Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων Και Μηχανικων Υπολογιστων
Τομεας Σηματων, Ελεγχου και Ρομποτικης

Εντοπισμός Επιθετικών Παραδειγμάτων σε Συνελικτικά Νευρωνικά Δίκτυα

Διπλωματικη Εργασια

του

ΣΤΕΦΑΝΟΥ ΠΕΡΤΙΓΚΙΟΖΟΓΛΟΥ

Επιβλέπων : Πέτρος Μαραγκός
Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2018

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Σημάτων, Ελέγχου και Ρομποτικής

Εντοπισμός Επιθετικών Παραδειγμάτων σε Συνελικτικά Νευρωνικά Δίκτυα

Διπλωματικη Εργασια

του

ΣΤΕΦΑΝΟΥ ΠΕΡΤΙΓΚΙΟΖΟΓΛΟΥ

Επιβλέπων : Πέτρος Μαραγκός
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 24η Ιουλίου 2018.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Πέτρος Μαραγκός	Γεράσιμος Ποταμιάνος	Χαράλαμπος Ψυλλάκης
Καθηγητής Ε.Μ.Π.	Αν. Καθηγητής Π.Θ.	Λέκτορας Ε.Μ.Π.

Αθήνα, Ιούλιος 2018

(Υπογραφή)

.....
Περτιγκιοζογλου Στεφανος
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον επιβλέπων καθηγητή μου κ.Πέτρο Μαραγκό για την ευκαιρία που μου έδωσε να εκπονήσω την διπλωματική μου εργασία στο εργαστήριο Όρασης Υπολογιστών, Επικοινωνίας Λόγου και Επεξεργασίας Σημάτων, καθώς και για την συνεχή καθοδήγησή του καθόλη την διάρκεια συγγραφής της διπλωματικής αυτής εργασίας. Επιπρόσθετα κατά την διάρκεια της φοίτησής μου στην σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, τα μαθήματα του κ.Μαραγκού με εισήγαγαν στο τομέα της Επεξεργασίας Σήματος και έπαιξαν καθοριστικό ρόλο στην απόφασή μου να συνεχίσω τις σπουδές μου στον τομέα αυτόν και ιδιαίτερα στον τομέα της Όρασης Υπολογιστών.

Ακόμα θα ήθελα να ευχαριστήσω τους γονείς μου και την αδελφή μου για την συμπαράστασή τους, την ηθική καθοδήγησή τους και την πίστη που έδειξαν σε εμένα. Τέλος θα ήθελα να ευχαριστήσω και τους στενούς μου φίλους που στάθηκαν δίπλα μου όλα αυτά τα χρόνια.

Περίληψη

Τα εντυπωσιακά αποτελέσματα που επιτυγχάνονται με την χρήση βαθιών νευρωνικών δικτύων, έχουν ως αποτέλεσμα την μεγάλη εξάπλωση της χρήσης τους σε πολλές εφαρμογές μηχανικής μάθησης. Τα συστήματα αυτά όμως είναι ευάλωτα σε ειδικά κατασκευασμένες εισόδους, τα επιθετικά παραδείγματα, τα οποία ενώ δεν γίνονται εύκολα αντιληπτά από ανθρώπινους παρατηρητές, οδηγούν τα νευρωνικά δίκτυα σε λανθασμένα συμπεράσματα.

Η εργασία αυτή εκτελεί ανάλυση των βασικών χαρακτηριστικών των επιθετικών παραδειγμάτων για συνελικτικά νευρωνικά δίκτυα τα οποία χρησιμοποιούνται σε εφαρμογές αναγνώρισης εικόνων, δίνοντας ιδιαίτερη βαρύτητα στα επιθετικά παραδείγματα που παράγονται από την Fast Gradient Sign Method, ενώ παράλληλα προτείνει τρεις νέες μεθόδους οι οποίες στοχεύουν στον εντοπισμό πιθανών επιθετικών παραδειγμάτων. Η πρώτη από τις προτεινόμενες μεθόδους βασίζεται στην ομαλοποίηση του διανύσματος χαρακτηριστικών που παράγει σαν έξοδο το νευρωνικό δίκτυο, προκειμένου να εντοπίσει πιθανά επιθετικά παραδείγματα. Η δεύτερη μέθοδος κάνει χρήση των ιστογραμμάτων των τιμών των εξόδων των ενδιάμεσων επιπέδων του νευρωνικού δικτύου, για να συμπεράνει αν η είσοδος του δικτύου αποτελεί επιθετικό παράδειγμα. Τα ιστογράμματα αυτά συγκροτούν ένα διάνυσμα χαρακτηριστικών το οποίο εισάγεται σαν είσοδος σε ένα SVM που ταξινομεί την αρχική είσοδο είτε ως επιθετικό παράδειγμα είτε ως πραγματική είσοδο. Τέλος για την τρίτη μέθοδο παρουσιάζουμε την έννοια τη υπολειπόμενης εικόνας, η οποία περιέχει πληροφορία σχετικά με τα μέρη του προτύπου εισόδου τα οποία αγνοούνται από το νευρωνικό δίκτυο. Η μέθοδος αυτή στοχεύει στον εντοπισμό πιθανών επιθετικών εικόνων, χρησιμοποιώντας την πληροφορία που παρέχει η υπολειπόμενη εικόνα και ενισχύοντας τα μέρη του προτύπου εισόδου που αγνοούνται από το νευρωνικό δίκτυο.

Για τις τρεις μεθόδους που προτείνονται, παρουσιάζονται τα αποτελέσματα εντοπισμού επιθετικών παραδειγμάτων σε ένα νευρωνικό δίκτυο εκπαιδευμένο στο MNIST σύνολο δεδομένων. Επίσης για την τρίτη μέθοδο παρουσιάζονται αποτελέσματα εντοπισμού επιθετικών παραδειγμάτων και σε ένα νευρωνικό δίκτυο εκπαιδευμένο στο CIFAR-10 σύνολο δεδομένων. Τέλος παρουσιάζεται η δυνατότητα συνδυασμού των μεθόδων για περαιτέρω ενίσχυση των αποτελεσμάτων εντοπισμού.

Λέξεις Κλειδιά

Μηχανική Μάθηση, Συνελικτικό Νευρωνικό Δίκτυο, Επιθετικά Παραδείγματα, Βαθιά Μάθηση, Όραση Υπολογιστών, Επεξεργασία Εικόνας

Abstract

The great success of deep neural networks, has caused a massive spread of the use of such systems in a large variety of machine learning applications. However these systems are vulnerable to certain inputs, the adversarial examples, which although are not easily perceived by humans can lead a neural network to produce faulty results.

This thesis analyzes the basic characteristics of adversarial examples for convolutional neural networks which are used in image recognition applications, emphasizing particularly in adversarial examples produced by the Fast Gradient Sign Method, while at the same time proposes three new methods for detecting possible adversarial examples. The first of the proposed methods is based on the regularization of the feature vector that the neural network produces as an output, in order to detect possible adversarial examples. The second method uses the histograms of the values from the outputs of the hidden layers of the neural network, in order to detect adversarial examples. These histograms create a feature vector which is the input of an SVM that classifies the original input either as an adversarial example or as a real input. Finally for the third method we introduce the concept of the residual image, which contains information about the parts of the input pattern that are ignored by the neural network. This method aims in the detection of possible adversarial examples, by using the residual image and reinforcing the parts of the input pattern that are ignored by the neural network.

For the three proposed methods we present the results of detecting adversarial examples in a convolutional neural networks trained in the MNIST dataset. Furthermore for the third method we present the results of detecting adversarial examples in a convolutional neural network trained in the CIFAR-10 dataset. Finally the possibility of combining these methods is presented as a way to further boost the results of detection.

Keywords

Machine Learning, Convolutional Neural Network, Adversarial Examples, Deep Learning, Computer Vision, Image Processing

Περιεχόμενα

Ευχαριστίες **5**

Περίληψη **7**

Abstract **9**

Περιεχόμενα **12**

Κατάλογος Σχημάτων **14**

Κατάλογος Πινάκων **15**

1 Εισαγωγή **17**
1.1 Επιθετικά Παραδείγματα 18
1.2 Στόχοι και Συνεισφορές της Εργασίας 18
1.3 Διάρθρωση της Εργασίας 19

2 Ανασκόπηση Σχετικών Αναφορών της Βιβλιογραφίας **21**
2.1 Συνελικτικά νευρωνικά δίκτυα 21
2.2 Το πρόβλημα παραγωγής επιθετικών παραδειγμάτων 25
2.3 Μέθοδοι κατασκευής επιθετικών εικόνων 27
2.4 Τεχνικές άμυνας ενάντια σε επιθετικές εικόνες που έχουν προταθεί στην βιβλιογραφία 30

3 Lipschitz συνέχεια και επιθετικά παραδείγματα **35**
3.1 Ο Scattering μετασχηματισμός και η επέκτασή του για πολυκαναλικά συνελικτικά δίκτυα 35
3.2 Lipschitz συνέχεια στον επεκτεταμένο Scattering μετασχηματισμό 36
3.3 Η εξέλιξη της σταθεράς Lipschitz κατά την εκπαίδευση με επιθετικές εικόνες 39

4 Ανάλυση επιθετικών εικόνων που παράγονται από την Ταχεία Τεχνική Προσημασμένης Παραγώγου (FGSM) **43**
4.1 Σύγκριση των αποτελεσμάτων της FGSM , και της επίλυσης του προβλήματος ελαχιστοποίησης 43
4.2 Προσέγγιση επιθετικών κατευθύνσεων με χρήση του συνόλου εκπαίδευσης . . 47
4.3 Γενίκευση επιθετικών εικόνων της FGSM σε διαφορετικά δίκτυα 49

5 Προτεινόμενες μέθοδοι άμυνας κατά επιθετικών εικόνων **51**
5.1 Τα σύνολα δεδομένων MNIST, CIFAR-10 51
5.2 Ομαλοποίηση του διανύσματος χαρακτηριστικών 52

5.3 Χρήση ιστογραμμάτων ενεργοποιήσεων του δικτύου για ανίχνευση επιθετικών εισόδων 55
5.4 Χρήση της υπολειπόμενης από τον ταξινομητή εικόνας για ανίχνευση επιθετικών εισόδων 61
5.4.1 Η έννοια της υπολειπόμενης εικόνας 61
5.4.2 Πειραματικές μέθοδοι για ανίχνευση επιθετικών εισόδων με χρήση της υπολειπόμενης εικόνας 65
5.4.3 Αποτελέσματα ανίχνευσης επιθετικών εισόδων σε MNIST , CIFAR-10 72
5.5 Σύγκριση και συνδυασμός των τριών προτεινόμενων μεθόδων άμυνας 75

6 Συμπεράσματα και Συμβολές της Εργασίας 79
6.1 Συμπεράσματα και Συμβολές 79
6.2 Κατευθύνσεις για μελλοντική έρευνα 80

Βιβλιογραφία 81

Κατάλογος Σχημάτων

2.1 Εικόνα των εξόδων των διαφορετικών επιπέδων ενός συνελικτικού νευρωνικού που δέχεται σαν είσοδο μια έγχρωμη εικόνα, [20] 22
2.2 Αρχιτεκτονική βαθιού συνελικτικού νευρωνικού δικτύου που αναπτύχθηκε στο [18] 24
2.3 Παράδειγμα της αρχιτεκτονικής του ResNet που αναπτύχθηκε στο [14] 25
2.4 Παράδειγμα εφαρμογή της μεθόδου που αναπτύσσεται στο [36], όπου ο αριστερός χρήστης αναγνωρίζεται από το σύστημα ως ο δεξιός χρήστης 26
2.5 Παράδειγμα παραγωγής επιθετικής εικόνας με χρήση της FGSM για νευρωνικό δίκτυο που έχει εκπαιδευτεί στο ImageNet,[11] 28

3.1 Εξέλιξη της σταθεράς L_{out} κατά την διαδικασία εκπαίδευσης με χρήση και χωρίς την χρήση επιθετικών εισόδων 42
3.2 Εξέλιξη του λάθους στο train,validation set κατά την εκπαίδευση με χρήση και χωρίς χρήση επιθετικών εισόδων 42

4.1 Αρχικό δείγμα ψηφίου /9/,επιθετικό παράδειγμα που ταξινομείται ως /4/ και η διαφορά τους 44
4.2 Αρχικό δείγμα ψηφίου /7/,επιθετικό παράδειγμα που ταξινομείται ως /4/ και η διαφορά τους 44
4.3 Αρχικό δείγμα ψηφίου /0/,επιθετικό παράδειγμα που ταξινομείται ως /2/ και η διαφορά τους 45
4.4 Βεβαιότητα δικτύου καθώς κινούμαστε στην διεύθυνση του προσήμου του gradient($grad$) και σε αυτήν που προκύπτει από την διαδικασία ελαχιστοποίησης της βεβαιότητα για την σωστή κατηγορία(adv) 46
4.5 Βεβαιότητα του δικτύου καθώς κινούμαστε σε μια τυχαία διεύθυνση($rand$) και σε αυτήν που προκύπτει από διαδικασία ελαχιστοποίησης της βεβαιότητας για την σωστή κατηγορία(adv) 46
4.6 Ιστόγραμμα των κατηγοριών των δειγμάτων που χρησιμοποιούνται για την προσέγγιση της επιθετικής κατεύθυνσης με την οποία το αρχικό δείγμα (κατηγορία /9/,label 10) ταξινομείται λανθασμένα ως /4/ (label 5) 48
4.7 Βεβαιότητα του δικτύου καθώς κινούμαστε στην τυχαία επιθετική κατεύθυνση(art), και στην διεύθυνση που προέκυψε από την διαδικασία ελαχιστοποίησης της βεβαιότητας για την σωστή κατηγορία(adv) 49

5.1 Παραδείγματα εικόνων του MNIST για τις 10 κατηγορίες ψηφίων 51
5.2 Παραδείγματα εικόνων του CIFAR-10 για τις 10 διαφορετικές κατηγορίες . . . 52
5.3 Παραδείγματα πραγματικής και επιθετικής εικόνας εισόδου 55
5.4 Παραδείγματα εξόδων του πρώτου συνελικτικού επιπέδου όταν οι είσοδοι είναι οι εικόνες του Σχήματος 5.3 56

5.5 Ποσοστό της ενέργειας των τιμών των εξόδων του πρώτου συνελικτικού επιπέδου οι οποίες βρίσκονται πάνω από ορισμένο κατώφλι για την πραγματική είσοδο του Σχήματος 5.3α και την επιθετική είσοδο του Σχήματος 5.3β 56
5.6 Ποσοστό της ενέργειας των τιμών των εξόδων του πρώτου συνελικτικού επιπέδου οι οποίες βρίσκονται πάνω από ορισμένο κατώφλι για την επιθετική είσοδο του Σχήματος 5.3β και είσοδο που προκύπτει ελαχιστοποιώντας την επιθετική κατηγορία 57
5.7 Ιστόγραμμα τιμών εξόδου του πρώτου συνελικτικού επιπέδου όταν εισάγουμε σαν είσοδο πραγματική εικόνα 58
5.8 Ιστόγραμμα τιμών εξόδου του πρώτου συνελικτικού επιπέδου όταν εισάγουμε σαν είσοδο επιθετικό παράδειγμα 59
5.9 Ποσοστά αναγνώρισης των δύο μεθόδων ιστογραμμάτων συναρτήσει της τυπικής απόκλισης του προστιθέμενου θορύβου 60
5.10 Παραδείγματα εικόνων εισόδου 61
5.11 Κατευθύνσεις που προκύπτουν από το backpropagation και ενισχύουν το μέτρο του διανύσματος χαρακτηριστικών 62
5.12 Μέση τιμή του λόγου του μέτρου της υπολοιπόμενης εικόνας $\boldsymbol{x}_{ign}$ προ το μέτρο της υπολειπόμενης εικόνας της πραγματικής εισόδου $\boldsymbol{x}_{0ign}$, καθώς το $\boldsymbol{x}$ κινείται από την πραγματική είσοδο $\boldsymbol{x}_0$ προς την επιθετική είσοδο 64
5.13 Παράδειγμα επιθετικής εισόδου που εντοπίζει η μέθοδος Α και ενδιάμεσων αποτελεσμάτων της μεθόδου 66
5.14 Παράδειγμα επιθετικής εισόδου που δεν εντοπίζει η μέθοδος Α και ενδιάμεσων αποτελεσμάτων της μεθόδου 66
5.15 Παράδειγμα πραγματικής εισόδου που ανιχνεύεται με επιτυχία ως πραγματική από την μέθοδο Α και ενδιάμεσων αποτελεσμάτων της μεθόδου 67
5.16 Παράδειγμα πραγματικής εισόδου που ανιχνεύεται ως επιθετική από την μέθοδο Α και ενδιάμεσων αποτελεσμάτων της μεθόδου 67
5.17 Παράδειγμα των εικόνων που παράγει η μέθοδος Β για διαφορετικό αριθμό επαναλήψεων της μεθόδου 69
5.18 Βεβαιότητα του ταξινομητή για την σωστή κατηγορία με χρήση της μεθόδου Β για πραγματική και επιθετική είσοδο 70
5.19 Βεβαιότητα του ταξινομητή για την σωστή κατηγορία με χρήση της μεθόδου Γ για πραγματική και επιθετική είσοδο 70
5.20 Σύγκριση των οπτικών αποτελεσμάτων μεταξύ των μεθόδων Β,Γ 72
5.21 Παραδείγματα επιθετικών εικόνων του CIFAR-10 και ενδιάμεσων αποτελεσμάτων που προκύπτουν από την εφαρμογή της μεθόδου Α 74
5.22 Κατανομή των κοινών λαθών ανάμεσα στις τρεις προτεινόμενες μεθόδους . . . 76

Κατάλογος Πινάκων

4.1 Μέσος όρος της L_2 νόρμας των διαφορών ανάμεσα στις επιθετικές εικόνες και τις πραγματικές εικόνες από τις οποίες παράγονται 48

5.1 Αποτελέσματα αναγνώρισης επιθετικών εικόνων μετά την ομαλοποίηση του διανύσματος χαρακτηριστικών όταν χρησιμοποιούμε διαφορετικές συναρτήσεις απόστασης 54
5.2 Αποτελέσματα των δύο μεθόδων ιστογραμμάτων για διαφορετική ποσότητα θορύβου 59
5.3 Αποτελέσματα των μεθόδων του κεφαλαίου 5.4.2 στο σύνολο δεδομένων MNIST και CIFAR-10 73
5.4 Αποτελέσματα από διαφορετικούς συνδυασμούς των αποτελεσμάτων των μεθόδων του Κεφαλαίου 5 77

Κεφάλαιο 1

Εισαγωγή

Η εξέλιξη του κλάδου της Όρασης Υπολογιστών είναι στενά συνδεδεμένη με την εξέλιξη των σύγχρονων τεχνικών Μηχανικής Μάθησης. Η συνεχόμενη ανάπτυξη των τεχνικών Μηχανικής Μάθησης που χρησιμοποιούνται σε εφαρμογές της Όρασης Υπολογιστών σε συνδυασμό με την δραματική αύξηση των δεδομένων που καταγράφουμε και αποθηκεύουμε, έχει ως αποτέλεσμα την επίτευξη εντυπωσιακών αποτελεσμάτων σε πολλές εφαρμογές που στοχεύουν στην εξαγωγή πληροφορίας από σήματα εικόνων ή σήματα βίντεο. Μια μεγάλη κατηγορία τέτοιων τεχνικών είναι οι τεχνικές επιβλεπόμενης μάθησης, όπου το σύστημα δέχεται ένα σύνολο από παραδείγματα εισόδων και εξόδων, και με βάση το σύνολο αυτό μαθαίνει να αντιστοιχεί άγνωστες εισόδους με τις επιθυμητές εξόδους.

Τα μοντέλα μηχανικής μάθησης στα οποία επικεντρώνεται η εργασία αυτή είναι τα νευρωνικά δίκτυα, και συγκεκριμένα τα βαθιά νευρωνικά δίκτυα. Παραδοσιακά ένα μοντέλο μηχανικής μάθησης απαιτούσε και ένα προσεκτικά κατασκευασμένο σύστημα προεπεξεργασίας των δεδομένων, προκειμένου το μοντέλο να δέχεται εισόδους οι οποίες αντιπροσωπεύουν τα κυριότερα χαρακτηριστικά των δεδομένων. Αντίθετα οι τεχνικές βαθιάς μάθησης και συγκεκριμένα τα βαθιά νευρωνικά δίκτυα, τα οποία αποτελούνται από πολλά διαδοχικά επίπεδα, έχουν την δυνατότητα να δέχονται τα αρχικά δεδομένα χωρίς να είναι απαραίτητη η προεπεξεργασία, καθώς αυτή εκτελείται στο εσωτερικό του μοντέλου. Έτσι η χρήση των βαθιών νευρωνικών δικτύων έφερε σημαντική βελτίωση σε πολλές εφαρμογές και μείωσε την αναγκαιότητα κατασκευής περίπλοκων μεθόδων εξαγωγής των βασικών χαρακτηριστικών των δεδομένων. Μια τέτοια βαθιά αρχιτεκτονική νευρωνικών δικτύων, που χρησιμοποιείται ιδιαίτερα σε εφαρμογές αναγνώρισης αντικειμένων σε σήματα εικόνων, είναι τα βαθιά συνελικτικά νευρωνικά δίκτυα. Το χαρακτηριστικό των συνελικτικών νευρωνικών δικτύων είναι ότι εκμεταλλεύονται την τοπικότητα που παρουσιάζουν τα μοτίβα τα οποία αναζητούμε σε μια εικόνα, χρησιμοποιώντας συνελικτικά επίπεδα. Επίσης η χρήση διαδοχικών τέτοιων συνελικτικών επιπέδων επιτρέπει τον εντοπισμό των τοπικών μοτίβων της εικόνας για διαφορετικές κλίμακες.

Αξίζει επίσης να αναφερθεί ότι η ανάπτυξη των βαθιών αρχιτεκτονικών των νευρωνικών δικτύων είναι στενά συνδεδεμένη με την ανάπτυξη υλικού εξειδικευμένου σε αποδοτική παράλληλη επεξεργασία (GPUs). Η συσχέτιση των δύο αυτών τεχνολογιών προκύπτει από το γεγονός, ότι ο μεγάλος αριθμός των παραμέτρων προς εκπαίδευσης σε ένα βαθύ νευρωνικό δίκτυο οδηγεί σε μια ιδιαίτερα υπολογιστικά κοστοβόρα διαδικασία εκπαίδευσης που απαιτεί μεγάλο αριθμό από δεδομένα εκπαίδευσης και από το ότι η διαδικασία εκπαίδευσης μπορεί σε μεγάλο βαθμό να παραλληλοποιηθεί. Ως συνέπεια η χρήση GPUs οδηγεί σε σημαντική μείωση του απαιτούμενου χρόνου για μια πλήρη εκπαίδευση.

1.1 Επιθετικά Παραδείγματα

Στην εργασία αυτή μελετάμε ορισμένα παραδείγματα ειδικά κατασκευασμένων εισόδων για βαθιά νευρωνικά δίκτυα, τα οποία ονομάζονται επιθετικά παραδείγματα (adversarial examples), και οδηγούν τα νευρωνικά δίκτυα σε λάθος συμπεράσματα. Συγκεκριμένα μελετάμε επιθετικά παραδείγματα σε βαθιά συνελικτικά νευρωνικά δίκτυα, τα οποία ταξινομούν μια εικόνα ή ένα μέρος αυτής στην κατηγορία στην οποία ανήκει το αντικείμενο που απεικονίζεται. Το χαρακτηριστικό των παραδειγμάτων αυτών είναι ότι ενώ ένας άνθρωπος δεν μπορεί να τα διακρίνει από τα τις υπόλοιπες εισόδους, όταν εισάγονται σε ένα νευρωνικό δίκτυο το οδηγούν σε λάθος συμπεράσματα ανεξάρτητα από την ακρίβεια που παρουσιάζει το δίκτυο για ένα γενικό σύνολο δεδομένων.

Μια εικόνα η οποία αποτελεί επιθετικό παράδειγμα για ένα βαθύ συνελικτικό νευρωνικό δίκτυο μπορεί να παραχθεί από μια πραγματική εικόνα του συνόλου δεδομένων. Η διαφορά μεταξύ των δύο αυτών εικόνων την οποία αντιλαμβάνεται ένας άνθρωπος είναι πολύ μικρή, με αποτέλεσμα να μην αλλάζει η πραγματική κατηγορία στην οποία θα πρέπει να ταξινομηθούν και οι δύο εικόνες σε ένα πρόβλημα αναγνώρισης. Αντίθετα όταν εισάγονται αυτές οι εικόνες σε ένα νευρωνικό δίκτυο, οι έξοδοι τους διαφέρουν σε μεγάλο βαθμό, με αποτέλεσμα να ταξινομούνται σε διαφορετικές κατηγορίες σε ένα πρόβλημα αναγνώρισης. Βλέπουμε λοιπόν ότι ο τρόπος με τον οποίο το νευρωνικό αντιλαμβάνεται τις εικόνες και επιλύει το πρόβλημα αναγνώρισης δεν προσεγγίζει στον βαθμό που αναμένουμε τον τρόπο με τον οποίο επιλύεται το πρόβλημα από έναν άνθρωπο. Η διαφορά ανάμεσα στην αναμενόμενη και την πραγματική συμπεριφορά ενός βαθιού νευρωνικού δικτύου αναδεικνύει την ανάγκη αναθεώρησης τόσο των αρχιτεκτονικών όσο και των μεθόδων εκπαίδευσης των νευρωνικών δικτύων.

Με την μελέτη των επιθετικών παραδειγμάτων μπορούμε να βελτιώσουμε το μοντέλο του συνελικτικού νευρωνικού δικτύου, αφού αν βελτιώσουμε την συμπεριφορά του νευρωνικού δικτύου ενάντια σε επιθετικές εικόνες, καταφέρνουμε να προσεγγίσουμε καλύτερα την ιδανική λύση του προβλήματος, όπου η αναγνώριση των εικόνων από ένα νευρωνικό δίκτυο ταυτίζεται με την αναγνώριση την οποία πραγματοποιεί ένας άνθρωπος. Έτσι μπορούμε να πετύχουμε καλύτερα αποτελέσματα όχι μόνο στα σύνολα επιθετικών παραδειγμάτων αλλά και σε τυχαία σύνολα άγνωστων εικόνων του προβλήματος. Επίσης η ύπαρξη εισόδων στις οποίες το νευρωνικό δίκτυο είναι ευάλωτο δημιουργεί προβλήματα ασφαλείας σε εφαρμογές όπου η σωστή αναγνώριση είναι κρίσιμη, αφού ένας κακόβουλος χρήστης εισάγοντας μια τέτοια είσοδο μπορεί να οδηγήσει το σύστημα στην εξαγωγή λανθασμένων συμπερασμάτων. Ένα παράδειγμα μιας τέτοιας εφαρμογής είναι η αναγνώριση σημάτων οδικής κυκλοφορίας με χρήση κάμερας η οποία είναι προσαρτημένη σε ένα όχημα, αφού ένας κακόβουλος χρήστης μπορεί επικαλύπτοντας μικρές περιοχές της πινακίδας οδοσήμανσης να μεταβάλει το σήμα το οποίο αναγνωρίζεται από την κάμερα. Σημαντική λοιπόν είναι και η ανάπτυξη μεθόδων οι οποίες, δεδομένου ότι ένα βαθύ νευρωνικό δίκτυο είναι ευάλωτο σε συγκεκριμένης επιθετικές εισόδους, έχουν την δυνατότητα να διακρίνουν τις εισόδους αυτές μέσα σε ένα σύνολο εισόδων το οποίο περιέχει μείγμα πραγματικών και επιθετικών εισόδων.

1.2 Στόχοι και Συνεισφορές της Εργασίας

Ο στόχος της εργασίας αυτής είναι η μελέτη των βασικών χαρακτηριστικών που παρουσιάζουν τα επιθετικά παραδείγματα τα οποία αποτελούν εισόδους βαθιών συνελικτικών νευρωνικών δικτύων, και η χρήση των χαρακτηριστικών αυτών για τον σχεδιασμό μεθόδων οι οποίες στοχεύουν στην διάκριση μεταξύ επιθετικών παραδειγμάτων και πραγματικών εισόδων. Τα επιθετικά παραδείγματα που μελετάμε αποτελούν σήματα εικόνων, οπότε αναφερόμαστε σε αυτά και ως επιθετικές εικόνες

Για την μελέτη των επιθετικών εικόνων θα επικεντρωθούμε κυρίως στις εικόνες οι οποίες παράγονται από την Fast Gradient Sign Method (FGSM), η οποία αποτελεί μια από τις βασικότερες μεθόδους κατασκευής επιθετικών εικόνων και παρουσιάζεται στο Κεφάλαιο 2. Μελετάμε λοιπόν τις ιδιότητες που έχουν οι εικόνες της FGSM και διερευνούμε τα αποτελέσματα μιας μεθόδου που προτείνουμε, η οποία παράγει μια προσέγγιση επιθετικών εικόνων χωρίς να κάνει χρήση ενός εκπαιδευμένου νευρωνικού δικτύου αλλά χρησιμοποιώντας μόνο το σύνολο εκπαίδευσης. Παράλληλα υλοποιούμε ορισμένες από τις μεθόδους άμυνας οι οποίες παρουσιάζονται στην βιβλιογραφία και εξετάζουμε το πώς οι μέθοδοι αυτές επηρεάζουν την δυνατότητα των επιθετικών εικόνων να επεκτείνονται σε περισσότερα από ένα νευρωνικά δίκτυα.

Επίσης προτείνουμε μια μέθοδο υπολογισμού μιας σταθεράς Lipschitz για ένα συνελικτικό νευρωνικό δίκτυο και παρουσιάζουμε το πώς μεταβάλλεται η σταθερά αυτή όταν το δίκτυο εκπαιδεύεται χρησιμοποιώντας μια μέθοδο άμυνας ενάντια σε επιθετικές εικόνες.

Τέλος η κυριότερη συνεισφορά της εργασίας αυτής είναι η πρόταση τριών διαφορετικών μεθόδων, οι οποίες στοχεύουν στον εντοπισμό πιθανών επιθετικών παραδειγμάτων τα οποία εισάγονται σαν είσοδοι σε ένα δεδομένο νευρωνικό δίκτυο. Οι μέθοδοι αυτές προτείνουν τρεις διαφορετικές κατευθύνσεις για την επίλυση του προβλήματος εντοπισμού επιθετικών εικόνων. Επίσης καθώς στο πρόβλημα εντοπισμού επιθετικών εικόνων το συνολικά καλύτερο αποτέλεσμα σε πολλές εφαρμογές προκύπτει από τον συνδυασμό διαφορετικών μεθόδων, κάτι που επιβεβαιώνεται και πειραματικά στην εργασία αυτή, οι μέθοδοι που προτείνουμε εκτός από τα μεμονωμένα αποτελέσματα που παρουσιάζονται μπορούν να συνδυαστούν τόσο μεταξύ τους όσο και με άλλες μεθόδους της βιβλιογραφίας για να παράξουν ακόμα καλύτερα αποτελέσματα.

1.3 Διάρθρωση της Εργασίας

Η εργασία αυτή είναι οργανωμένη σε 6 κεφάλαια

- Στο Κεφάλαιο 2 γίνεται ανασκόπηση των κυριότερων αναφορών της βιβλιογραφίας σχετικά με τα συνελικτικά νευρωνικά δίκτυα και τα επιθετικά παραδείγματα και παρουσιάζονται οι κυριότερες μέθοδοι κατασκευής επιθετικών παραδειγμάτων και άμυνας ενάντια σε αυτά

- Στο Κεφάλαιο 3 παρουσιάζεται ο Scattering μετασχηματισμός και μια επέκτασή του, ως εργαλεία για την μοντελοποίηση της λειτουργίας ενός συνελικτικού νευρωνικού, ενώ παράλληλα εξετάζεται η συσχέτιση της σταθεράς Lipschitz του δικτύου με την ικανότητά του να αντιμετωπίζει επιθετικές εικόνες.

- Στο Κεφάλαιο 4 αναλύονται βασικές ιδιότητες των επιθετικών εικόνων, όπου δίνεται ιδιαίτερη έμφαση στις εικόνες που προκύπτουν από την FGSM μέθοδο

- Στο Κεφάλαιο 5 παρουσιάζονται οι τρεις προτεινόμενες μέθοδοι για τον εντοπισμό επιθετικών εικόνων. Αρχικά για κάθε μία από τις μεθόδους γίνεται ανάλυση της λειτουργίας της και παρουσιάζονται τα πειραματικά αποτελέσματα που επιτυγχάνει. Στην συνέχεια του κεφαλαίου εκτελούμε σύγκριση των αποτελεσμάτων των μεθόδων και εξετάζουμε πιθανούς τρόπους συνδυασμού των τριών μεθόδων.

- Στο Κεφάλαιο 6 γίνεται σύνοψη των κυριότερων αποτελεσμάτων και συνεισφορών της εργασίας και προτείνονται πιθανές κατευθύνσεις για μελλοντική έρευνα οι οποίες προκύπτουν από την εργασία αυτή.

Κεφάλαιο 2

Ανασκόπηση Σχετικών Αναφορών της Βιβλιογραφίας

2.1 Συνελικτικά νευρωνικά δίκτυα

Στην ενότητα αυτή παρουσιάζουμε τα βασικά χαρακτηριστικά των νευρωνικών δικτύων. Στο [3], παρουσιάζεται η βασική δομή ενός απλού νευρωνικού δικτύου το οποίο αποτελείται από ένα σύνολο μονάδων (perceptron), οι οποίες λαμβάνουν το διάνυσμα εισόδου, υπολογίζουν έναν γραμμικό συνδυασμό των στοιχείων της εισόδου, τον περνάνε από μια μη-γραμμική συνάρτηση ενεργοποίησης και δίνουν το αποτέλεσμα σαν έξοδο. Η πιο συνηθισμένη αρχιτεκτονική ενός απλού νευρωνικού δικτύου αποτελείται από δύο επίπεδα perceptron, όπου το πρώτο λαμβάνει σαν είσοδο το διάνυσμα εισόδου και το δεύτερο λαμβάνει σαν είσοδο την έξοδο του πρώτου επιπέδου και παράγει την τελική έξοδο του δικτύου. Οπότε ένας ταξινομητής ο οποίος υλοποιείται με την παραπάνω αρχιτεκτονική, δέχεται σαν είσοδο τα βαθμωτά χαρακτηριστικά $x_1, x_2, ..., x_D$ και παράγει την έξοδο y_k ως εξής

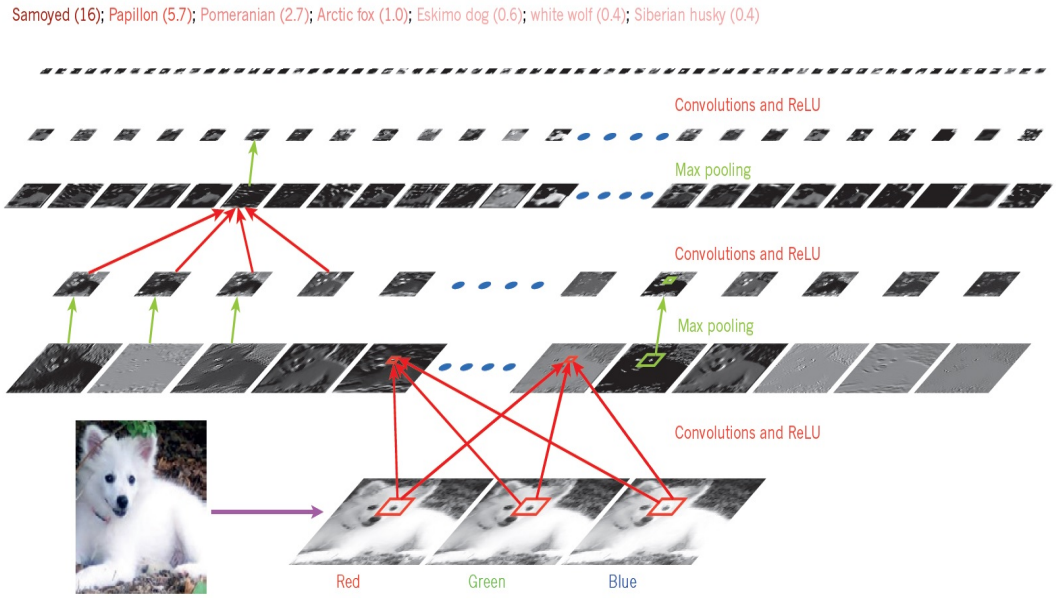
$$y_k = \sigma\left(\sum_{j=0}^{M} w_{kj}^{(2)} \sigma\left(\sum_{i=0}^{D} w_{ji}^{(1)} x_i\right)\right) \quad k = 1, 2, ..., K \tag{2.1}$$

όπου K είναι ο συνολικός αριθμός των εξόδων του δικτύου, M είναι ο αριθμός των perceptron του πρώτου επιπέδου του δικτύου, σ είναι η συνάρτηση ενεργοποίησης και $w_{ij}^{(p)}$ είναι τα βάρη που μαθαίνονται κατά την εκπαίδευση του δικτύου.

Το μοντέλο του δικτύου που παρουσιάστηκε προηγουμένως, αποτελεί έναν ρηχό ταξινομητή, καθώς αποτελείται από δύο επίπεδα perceptron . Τέτοιου είδους ταξινομητές απαιτούν σαν είσοδο ένα διάνυσμα χαρακτηριστικών το οποίο αποτελεί μια καλή αναπαράσταση των σημαντικών χαρακτηριστικών του προτύπου εισόδου, τα οποία το διαχωρίζουν από πρότυπα μιας διαφορετικής κατηγορίας. Για την αποφυγή του σχεδιασμού ενός πολύπλοκου συστήματος εξαγωγής χαρακτηριστικών επιθυμούμε ο ταξινομητής να έχει την δυνατότητα να μάθει τον καλύτερο τρόπο για την εξαγωγή χαρακτηριστικών κατά την εκπαίδευση.

Έτσι στο [20] παρουσιάζονται τα βαθιά νευρωνικά δίκτυα τα οποία ακολουθούν την ίδια μεθοδολογία με το δίκτυο που παρουσιάστηκε στην προηγούμενη παράγραφο, αλλά αποτελούνται από μεγαλύτερο αριθμό επιπέδων. Αυτές οι βαθιές αρχιτεκτονικές μπορούν να υλοποιήσουν περίπλοκες συναρτήσεις που εκτελούν ταξινόμηση όταν σαν είσοδο δέχονται το ίδιο το διάνυσμα του προτύπου που θέλουμε να αναγνωρίσουμε, όπως για παράδειγμα τις τιμές των pixel μιας εικόνας.

Μια κατηγορία βαθιών νευρωνικών δικτύων είναι τα βαθιά συνελικτικά νευρωνικά δίκτυα που παρουσιάζονται στο [20]. Η ιδιαιτερότητα των δικτύων αυτών είναι ότι δέχονται ως ε-



Σχήμα 2.1: Εικόνα των εξόδων των διαφορετικών επιπέδων ενός συνελικτικού νευρωνικού που δέχεται σαν είσοδο μια έγχρωμη εικόνα, [20]

ίσοδο πρότυπα των οποίων οι αναπαραστάσεις αποτελούν πολυδιάστατα διακριτά σήματα και εκμεταλεύονται την τοπικότητα των μοτίβων που αναζητούμε στα πρότυπα αυτά . Έτσι τα συνελικτικά δίκτυα βρίσκουν μεγάλη εφαρμογή στην αναγνώριση εικόνων, όπου έχουμε 2D σήματα στην περίπτωση ασπρόμαυρων εικόνων και 3D σήματα στην περίπτωση έγχρωμων εικόνων. Τα ενδιάμεσα επίπεδα του συνελικτικού δικτύου δίνουν σαν έξοδο ένα σύνολο από πολυδιάστατα διακριτά σήματα τα οποία αναφέρονται και ως feature maps. Ένα παράδειγμα feature map μπορεί να είναι ένα $N \times N$ διακριτό σήμα που μπορεί να ερμηνευτεί και ως ένα $N \times N$ σήμα εικόνας όπως φαίνεται και στο Σχήμα 2.1. Για απλοποίηση των συμβολισμών αναφερόμαστε σε ένα feature map και ώς το διάνυσμα $\boldsymbol{x}$ το οποίο προκύπτει αν λάβουμε τα στοιχεία του feature map σε μορφή διανύσματος. Τα στοιχεία του διανύσματος $\boldsymbol{x}$ διατηρούν τις σχέσεις γειτονίας που είχαν στο feature map από το οποίο προέκυψε το $\boldsymbol{x}$. Έτσι συμβολίζουμε με $(\psi * \boldsymbol{x})$ την διακριτή συνέλιξη του φίλτρου κρουστικής απόκρισης ψ με το feature map που αντιστοιχεί στο διάνυσμα $\boldsymbol{x}$, η οποία μπορεί να υλοποιηθεί και με τον πολλαπλασιασμό $\boldsymbol{A}_\psi \boldsymbol{x}$ όπου $\boldsymbol{A}_\psi$ ο πίνακας συνέλιξης που προκύπτει από τον πυρήνα ψ. Τον συμβολισμό αυτόν ακολουθούμε και για τα πρότυπα εισόδου, οπότε για παράδειγμα αναφερόμαστε σε μια εικόνα εισόδου ως εικόνα εισόδου $\boldsymbol{x}$.

Τα αρχικά επίπεδα ενός συνελικτικού δικτύου αποτελούνται κυρίως από 3 είδη επιπέδων

- Συνελικτικό επίπεδο: Το συνελικτικό επίπεδο για κάθε διάνυσμα χαρακτηριστικών (feature map) που δέχεται στην είσοδο, εκτελεί διακριτή συνέλιξη με ένα σύνολο από φίλτρα των οποίων οι παράμετροι μαθαίνονται κατά την διαδικασία εκπαίδευσης. Έτσι αν $\boldsymbol{x}_{ij}$ είναι το i-οστό διάνυσμα χαρακτηριστικών της εξόδου του j επιπέδου, το οποίο αποτελεί συνελικτικό επίπεδο, και ψ_{ki} είναι η κρουστική απόκριση του φίλτρου που συνδέει το k διάνυσμα του προηγούμενου επιπέδου με το παρόν i-οστό διάνυσμα, τότε το συνελικτικό επίπεδο παράγει την έξοδο

$$\boldsymbol{x}_{ij} = \sum_k (\psi_{ki} * \boldsymbol{x}_{k(j-1)})$$

- Pooling επίπεδο: Το επίπεδο αυτό στοχεύει στην συγχώνευση των χαρακτηριστικών των διανυσμάτων που προκύπτουν ως έξοδοι προηγούμενων επιπέδων του δίκτυο. Η χρήση συνελικτικών επιπέδων έχει ως αποτέλεσμα γειτονικά στοιχεία του διανύσματος χαρακτηριστικών εξόδου του συνελικτικού επιπέδου να αναφέρονται στην ανίχνευση μοτίβων σε γειτονικές περιοχές του διανύσματος εισόδου. Για τον λόγο αυτό η διαδικασία του Pooling εφαρμόζεται ξεχωριστά σε τοπικές γειτονιές του διανύσματος χαρακτηριστικών. Έτσι το επίπεδο αυτό χωρίζει το διάνυσμα σε τοπικές γειτονιές, οι οποίες μπορούν και να επικαλύπτονται, και για κάθε μια από τις γειτονιές αυτές δίνει στην έξοδο τον μέσο όρο των στοιχείων της γειτονιάς (Average Pooling) ή την μέγιστη τιμή των στοιχείων της γειτονιάς (Max Pooling).

- Επίπεδο μη-γραμμικότητας: Το επίπεδο αυτό εφαρμόζει σε κάθε ένα από τα στοιχεία των διανύσματων εισόδου μια μη-γραμμική συνάρτηση. Υπάρχει μια ποικιλία από συναρτήσεις όπως η υπερβολική εφαπτομένη $f(x) = \tanh(x)$, η λογιστική συνάρτηση $f(x) = 1/(1 + \exp(-x))$, με την ποιο διαδεδομένη συνάρτηση για τα συνελικτικά νευρωνικά δίκτυα να είναι η ReLU(Rectified Linear Unit) $f(x) = \max(0, x)$.

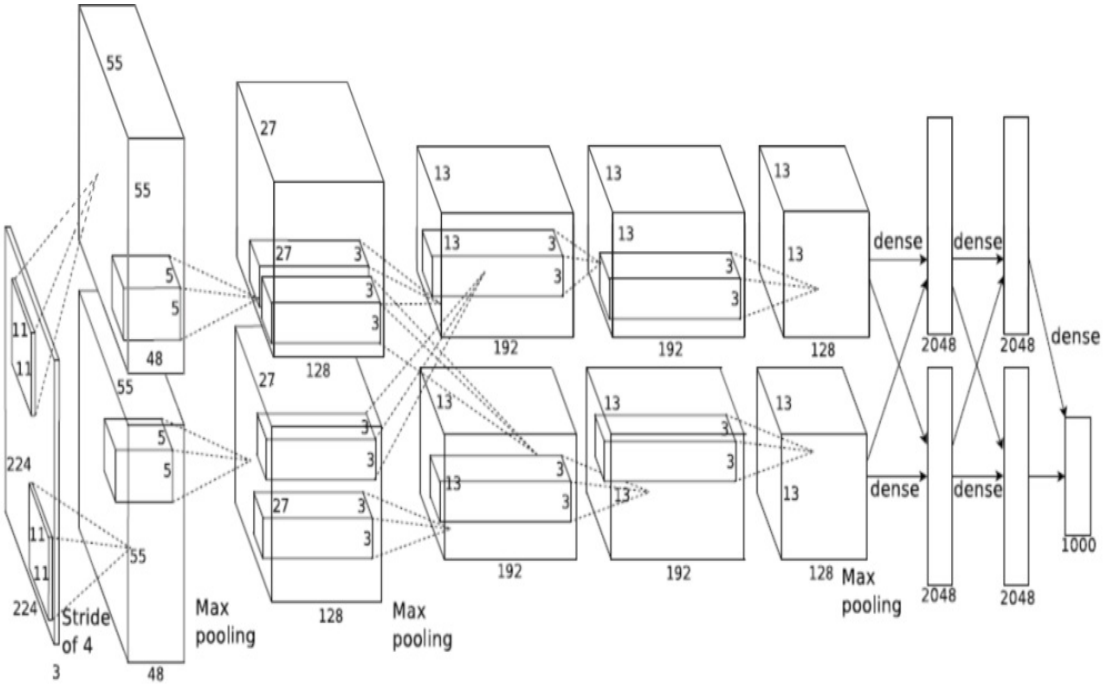
Μια συνηθισμένη αρχιτεκτονική που χρησιμοποιείται στα περισσότερα συνελικτικά δίκτυα είναι η διαδοχική χρήση μιας τριάδας επιπέδων η οποία αποτελείται από ένα συνελικτικό επίπεδο ένα Pooling επίπεδο και ένα επίπεδο μη-γραμμικότητας (ReLU). Επίσης σε πολλές αρχιτεκτονικές μετά από ορισμένο αριθμό επιπέδων παράγεται ένα διάνυσμα χαρακτηριστικών το οποίο αποτελεί είσοδο ενός πλήρους συνδεδεμένου νευρωνικού δικτύου, όπου ακολουθεί το μοντέλο του απλού νευρωνικού που περιγράφτηκε στην αρχή του κεφαλαίου, και το οποίο μας δίνει την τελική έξοδο του συνολικού δικτύου. Στο Σχήμα 2.1 παρουσιάζεται ένα παράδειγμα των εξόδων των διαφορετικών επιπέδων ενός συνελικτικού νευρωνικού.

Τα επίπεδα των συνελικτικών νευρωνικών δικτύων, όπως και τα επίπεδα των απλών νευρωνικών, επιτρέπουν την παραγωγή του gradient της εξόδου ως προς τις παραμέτρους του δικτύου με χρήση του backpropagation αλγορίθμου. Αυτό επιτρέπει την εύκολη εκπαίδευση των δικτύων αυτών με χρήση της μεθόδου SGD (stohastic gradient descent), με την οποία λαμβάνουμε τον μέσο όρο του gradient του λάθους ως προς κάθε παράμετρο, όταν εισάγουμε ένα σύνολο από εισόδους στο δίκτυο, και τον χρησιμοποιούμε για την επαναληπτική ανανέωση των παραμέτρων με σκοπό την ελαχιστοποίηση του λάθους.

Στο [18] παρουσιάζεται ένα παράδειγμα υλοποίησης ενός βαθιού συνελικτικού δικτύου το οποίο πέτυχε εντυπωσιακά αποτελέσματα στο σύνολο δεδομένων ImageNet, το οποίο αποτελείται από 15 εκατομμύρια έγχρωμες εικόνες υψηλής ποιότητας οι οποίες ανήκουν σε 22000 διαφορετικές κατηγορίες. Το δίκτυο αυτό αποτελείται από 5 συνελικτικά επίπεδα τα οποία ακολουθούνται από ReLU,Max Pooling επίπεδα και από 3 πλήρως συνδεδεμένα επίπεδα όπως φαίνεται στο Σχήμα 2.2. Παρά τον μεγάλο αριθμό εικόνων που είναι διαθέσιμες στο ImageNet, ο μεγάλος αριθμός παραμέτρων οδηγούσε σε overfitting, με αποτέλεσμα να μην επιτυγχάνεται ικανοποιητική γενίκευση. Για την καταπολέμηση του προβλήματος αυτού, στο [18] προτείνονται οι παρακάτω τεχνικές που εφαρμόζονται κατά την εκπαίδευση.

- Dropout : Με την τεχνική αυτή, κάθε έξοδος των ενδιάμεσων επιπέδων του δικτύου μηδενίζεται με πιθανότητα 0.5. Έτσι κατά την εκπαίδευση δειγματοληπτούμε από έναν μεγάλο αριθμό αρχιτεκτονικών οι οποίες μοιράζονται τα ίδια βάρη, με αποτέλεσμα να οδηγούμαστε σε καλύτερη γενίκευση. Η τεχνική αυτή παρουσιάζεται στο [15]

- Αύξηση του συνόλου δεδομένων εκπαίδευσης: Η τεχνική αυτή βασίζεται στην εφαρμογή μετασχηματισμών οι οποίοι διατηρούν την κατηγορία αναγνώρισης στις εικόνες του συνόλου εκπαίδευσης. Με αυτό τον τρόπο επιτυγχάνεται η αύξηση των συνολικών

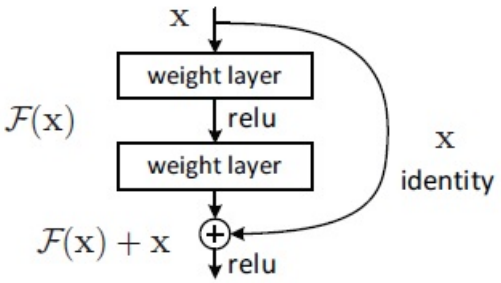


Σχήμα 2.2: Αρχιτεκτονική βαθιού συνελικτικού νευρωνικού δικτύου που αναπτύχθηκε στο [18]

διαφορετικών εικόνων που είναι διαθέσιμες κατά την εκπαίδευση με αποτέλεσμα να μειώνεται το overfitting. Στο [37] παρουσιάζονται τέτοιου είδους μετασχηματισμοί, όπως η εφαρμογή ενός αφινικού μετασχηματισμού της αρχικής εικόνας ο οποίος δεν μεταβάλει την κατηγορία στην οποία ανήκει η εικόνα.

Το πρόβλημα που προκύπτει στην περίπτωση των βαθιών νευρωνικών δικτύων είναι ότι παρόλο που η εισαγωγή επιπλέον επιπέδων μας επιτρέπει την επίλυση δυσκολότερων προβλημάτων, με την εισαγωγή ενός νέου επιπέδου στο δίκτυο, η εκπαίδευση του και η εύρεση των βέλτιστων παραμέτρων γίνεται δυσκολότερη. Στο [14] αναφέρεται ότι η εισαγωγή ενός επιπλέον επιπέδου δεν θα έπρεπε να οδηγήσει σε αύξηση του λάθους του δικτύου, καθώς το νέο επίπεδο μπορεί να λάβει την μορφή ενός ταυτοτικού επιπέδου το οποίο περνάει την είσοδό του στην έξοδο. Στην πράξη αυτό δεν συμβαίνει και η εισαγωγή επιπλέον επιπέδων μπορεί να οδηγήσει στην αύξηση του λάθους ακόμα και στο σύνολο εκπαίδευσης, γεγονός που δείχνει ότι το πρόβλημα δεν οφείλεται σε overfitting .

Για την αντιμετώπιση των προβλημάτων που προκύπτουν με την εισαγωγή επιπρόσθετων επιπέδων, στο [14] προτείνεται η αρχιτεκτονική του ResNet, με το οποίο βαθύτερες αρχιτεκτονικές δεν εμφανίζουν αύξηση του λάθους. Η αρχιτεκτονική αυτή βασίζεται στην απευθείας προώθηση της εισόδου ενός επιπέδου στις εισόδους ανώτερων επιπέδων. Έτσι αν έχουμε μια ομάδα από διαδοχικά επίπεδα που δέχονται είσοδο $\boldsymbol{x}$ και δίνουν έξοδο $F(\boldsymbol{x})$, τότε η τελική έξοδος των διαδοχικών επιπέδων θα είναι $F(\boldsymbol{x}) + \boldsymbol{x}$. Με αυτόν τον τρόπο αν η βέλτιστη λύση προκύπτει όταν το επίπεδο είναι ταυτοτικό επίπεδο, είναι ευκολότερο κατά την εκπαίδευση να οδηγηθούμε σε αυτήν με $F(\boldsymbol{x}) = 0$.



Σχήμα 2.3: Παράδειγμα της αρχιτεκτονικής του ResNet που αναπτύχθηκε στο [14]

2.2 Το πρόβλημα παραγωγής επιθετικών παραδειγμάτων

Σύμφωνα με τα [40],[11] πολλά από τα βαθιά νευρωνικά δίκτυα παρουσιάζουν σημεία στο χώρο εισόδου τους τα οποία ονομάζουμε επιθετικά παραδείγματα (adversarial examples) . Τα σημεία αυτά, όπως αναφέρουμε και στο Κεφάλαιο 1, ταξινομούνται λάθος από το δίκτυο και προκύπτουν μετά από πολύ μικρή μεταβολή των *pixel* της εικόνας εισόδου, η οποία πριν την μεταβολή ταξινομείται με επιτυχία από το νευρωνικό στην κατηγορία στην οποία ανήκει. Έτσι έχουμε δύο εικόνες των οποίων οι διαφορές είναι δύσκολα αναγνωρίσιμες από τον άνθρωπο, και το νευρωνικό τις ταξινομεί σε διαφορετικές κατηγορίες και μάλιστα σε πολλές περιπτώσεις με μεγάλη βεβαιότητα.

Γενικά οι μέθοδοι που παράγουν επιθετικά παραδείγματα έχουν την ελευθερία να αλλάζουν την τιμή σε κάθε pixel της εικόνας ξεχωριστά. Μια τέτοια μέθοδος είναι η Dense Adversary Generation που αναπτύσσεται στο [43]. Η μέθοδος αυτή εφαρμόζεται σε συνελικτικά νευρωνικά που έχουν την δυνατότητα εντοπισμού αντικειμένων μέσα σε μια εικόνα και ταξινόμηση κάθε ενός από τα αντικείμενα αυτά στις αντίστοιχες κατηγορίες στις οποίες ανήκουν. Στο [43] παρουσιάζονται παραδείγματα επιθετικών εικόνων που επιτυγχάνουν να οδηγήσουν ένα VGGnet [38] δίκτυο τόσο σε λάθος εντοπισμό αντικειμένων όσο και σε λάθος ταξινόμηση των αντικειμένων αυτών.

Αντίθετα με την παραπάνω μέθοδο, μια διαδικασία που επιδιώκει την υλοποίηση των επιθετικών εικόνων στο φυσικό κόσμο παρουσιάζεται στο [36]. Συγκεκριμένα έχει ως σκοπό να οδηγήσει ένα σύστημα αναγνώρισης προσώπων σε λάθος αναγνώριση, δημιουργώντας ένα πλαίσιο γυαλιών το οποίο μπορεί να φοράει ο χρήστη. Έτσι στοχεύουν στην μεταβολή των pixel της εικόνας στην περιοχή γύρω από τα μάτια του ατόμου που καλύπτει ένα πλαίσιο γυαλιών, ενώ παράλληλα ενισχύουν την δυνατότητα της εκτύπωσης ενός τέτοιου πλαισίου εισάγοντας μια ποινή για πλαίσια τα οποία δεν μπορούν να παραχθούν από έναν εκτυπωτή. Στο Σχήμα 2.4 βλέπουμε ένα τέτοιο πλαίσιο.

Αντίστοιχα με το [36], εφαρμογές που στοχεύουν στην φυσική υλοποίηση επιθετικών παραδειγματων παρουσιάζονται στο [19] καθώς και στο [9], το οποίο στοχεύει στην υλοποίηση επιθετικών παραδειγμάτων για εφαρμογή αναγνώρισης πινακίδων οδοσήμανσης, μια εφαρμογή όπου η σωστή αναγνώριση είναι κρίσιμη. Παράλληλα έξω από τον κλάδο της Όρασης Υπολογιστών, στα [13], [1] παρουσιάζονται επιθετικά παραδείγματα για συστήματα αυτόματης αναγνώρισης κακόβουλου λογισμικού, όπου η σωστή αναγνώριση είναι εξίσου κρίσιμη.

Δεδομένου λοιπόν ενός εκπαιδευμένου δικτύου που υλοποιεί την συνάρτηση f, και το οποίο ταξινομεί την είσοδο $\boldsymbol{x}$ στην κατηγορία $f(\boldsymbol{x}) = \ell \quad \ell \in \{1, \ldots, k\}$, μια απλή μοντελοποίηση του προβλήματος εύρεση ενός επιθετικού παραδείγματος είναι η επίλυση του παρακάτω προβλήματος

Σχήμα 2.4: Παράδειγμα εφαρμογή της μεθόδου που αναπτύσσεται στο [36], όπου ο αριστερός χρήστης αναγνωρίζεται από το σύστημα ως ο δεξιός χρήστης

Ελαχιστοποίηση του $\|\boldsymbol{r}\|_2$ έτσι ώστε:

- $f(\boldsymbol{x}+\boldsymbol{r}) = \ell_2 \neq \ell = f(\boldsymbol{x})$
- $\boldsymbol{x}+\boldsymbol{r} \in I$ όπου I ο χώρος εισόδου

Στην παραπάνω μοντελοποίηση θεωρούμε ότι δύο είσοδοι είναι όμοιες όταν η L_2 νόρμα της διαφορά τους είναι μικρή. Αυτό όμως αποτελεί μια απλοποίηση του προβλήματος καθώς στην πραγματικότητα μας ενδιαφέρει να μεγιστοποιήσουμε την ομοιότητα την οποία αντιλαμβάνεται ένας παρατηρητής όταν βλέπει τις δύο εικόνες. Έτσι το μέτρο το οποίο χρησιμοποιούμε για να μετρήσουμε την ομοιότητα δύο εικόνων είναι μια σημαντική παράμετρος του προβλήματος, η οποία επηρεάζει και τις προσεγγιστικές λύσεις που βρίσκουμε όταν προσπαθούμε να το λύσουμε.

Αν θέλουμε λοιπόν να γενικεύσουμε το πρόβλημα εύρεση επιθετικών παραδειγμάτων για απόσταση $d(\boldsymbol{x},\boldsymbol{y})$ διαφορετική από την L_2 νόρμα της διαφοράς των $\boldsymbol{x}$,$\boldsymbol{y}$, τότε το πρόβλημα ελαχιστοποίησης γίνεται

$$\arg\min_{\boldsymbol{y}} d(\boldsymbol{x},\boldsymbol{y}) : f(\boldsymbol{x}) \neq \ell \quad \text{και } d(\boldsymbol{x},\boldsymbol{y}) < \tau \tag{2.2}$$

όπου τ το κατώφλι για το οποίο ισχύει ότι αν δύο εικόνες $\boldsymbol{x},\boldsymbol{y}$ έχουν απόσταση $d(\boldsymbol{x},\boldsymbol{y}) < \tau$, τότε οι δύο εικόνες θεωρούνται ως όμοιες από έναν ανθρώπινο παρατηρητή.

Στο [34] προτείνεται το PASS, μια μέθοδος για εύρεση της οπτικής ομοιότητας ανάμεσα σε δύο εικόνες την οποία αντιλαμβάνεται ένας παρατηρητής. Η μέθοδος αποτελείται από δύο στάδια. Το πρώτο στάδιο σκοπεύει στην ευθυγράμμιση των δύο εικόνων, έτσι ώστε η απόσταση να μην εξαρτάται από παραμορφώσεις της εικόνας που οφείλονται σε περιστροφές, μετατοπίσεις ή την οπτική γωνία με την οποία παρατηρούμε την εικόνα.

Έστω λοιπόν ότι έχουμε δύο εικόνες $\boldsymbol{x}$,$\boldsymbol{y}$ και η $\psi(\boldsymbol{y},\boldsymbol{x})$ είναι ένας ομογραφικός μετασχηματισμό από την εικόνα $\boldsymbol{y}$ στην εικόνα $\boldsymbol{x}$ με πίνακα ομογραφίας $\boldsymbol{H}$. Τον πίνακα $\boldsymbol{H}$ τον βρίσκουμε ελαχιστοποιώντας την συνάρτηση

$$\arg\min_{\boldsymbol{H}} \left\| \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2} - \frac{\psi(\boldsymbol{y},\boldsymbol{x})}{\|\psi(\boldsymbol{y},\boldsymbol{x})\|_2} \right\|_2 \tag{2.3}$$

Στο δεύτερο στάδιο η μέθοδος χρησιμοποιεί τρεις διαφορετικές μετρικές ομοιότητας ανάμεσα σε δύο εικόνες $\boldsymbol{x}$,$\boldsymbol{y}$ οι οποίες αναπτύχθηκαν στο [41]. Οι μετρικές αυτές είναι οι $L(\boldsymbol{x},\boldsymbol{y})$, $C(\boldsymbol{x},\boldsymbol{y})$, $S(\boldsymbol{x},\boldsymbol{y})$ οι οποίες καθορίζουν την ομοιότητα σε φωτεινότητα, αντίθεση , δομική ομοιότητα.

$$L(\boldsymbol{x},\boldsymbol{y}) = \left[\frac{2\mu_{\boldsymbol{x}}\mu_{\boldsymbol{y}} + C_1}{\mu_{\boldsymbol{x}}^2 + \mu_{\boldsymbol{y}}^2 + C_1} \right] \tag{2.4}$$

$$C(\boldsymbol{x},\boldsymbol{y}) = \left[\frac{2\sigma_{\boldsymbol{x}}\sigma_{\boldsymbol{y}} + C_2}{\sigma_{\boldsymbol{x}}^2 + \sigma_{\boldsymbol{y}}^2 + C_2}\right] \tag{2.5}$$

$$S(\boldsymbol{x},\boldsymbol{y}) = \left[\frac{\sigma_{\boldsymbol{xy}} + C_3}{\sigma_{\boldsymbol{x}}\sigma_{\boldsymbol{y}} + C_3}\right] \tag{2.6}$$

όπου τα $\mu_{\boldsymbol{x}}$,$\sigma_{\boldsymbol{x}}$,$\sigma_{\boldsymbol{xy}}$ αποτελούν σταθμισμένο μέσον όρο, τυπική απόκλιση και συνδιακύμανση αντίστοιχα και οι C_1, C_2, C_3 είναι σταθερές που εισάγουμε προκειμένου να μην οδηγηθούμε σε διαίρεση με το μηδέν. Με τις μετρικές αυτές ορίζεται ο τοπικός δείκτης ομοιότητας RSSIM, όπου αν λάβουμε μια τοπική περιοχή $\boldsymbol{x}_r, \boldsymbol{y}_r$ των εικόνων $\boldsymbol{x}, \boldsymbol{y}$ αντίστοιχα, ορίζεται ως εξής

$$\text{RSSIM}(\boldsymbol{x}_r, \boldsymbol{y}_r) = L(\boldsymbol{x}_r, \boldsymbol{y}_r)C(\boldsymbol{x}_r, \boldsymbol{y}_r)S(\boldsymbol{x}_r, \boldsymbol{y}_r) \tag{2.7}$$

Τον οποίο αν αθροίσουμε για την συνολική εικόνα λαμβάνουμε το δείκτη ομοιότητας SSIM που αναπτύχθηκε στο [41]. Έτσι η μετρική ομοιότητας PASS προκύπτει αν συνδυάσουμε και τα δύο στάδια ως εξής.

$$\text{PASS}(\boldsymbol{y},\boldsymbol{x}) = \text{SSIM}(\psi(\boldsymbol{y},\boldsymbol{x}),\boldsymbol{x}) \tag{2.8}$$

Χρησιμοποιόντας τον παραπάνω δείκτη ομοιότητας λαμβάνουμε την απόσταση $d(\boldsymbol{y},\boldsymbol{x}) = 1 - \text{PASS}(\boldsymbol{y},\boldsymbol{x})$ που προσεγγίζει καλύτερα την ανθρώπινη αντίληψη, από ότι οι L_p νόρμες των διαφορών των δύο εικόνων, οι οποίες χρησιμοποιούνται συνήθως. Το μειονέκτημα χρήσης του PASS είναι ότι αυξάνει σημαντικά την πολυπλοκότητα του προβλήματος εύρεσης επιθετικών παραδειγμάτων.

2.3 Μέθοδοι κατασκευής επιθετικών εικόνων

Για την εύρεση των σημείων στο χώρο εισόδου του ταξινομητή στα οποία εμφανίζεται μεταβολή της κατηγορίας αναγνώριση με μικρή μεταβολή της εισόδου, έχουν προταθεί μια μεγάλη ποικιλία από αλγορίθμους.

Fast Gradient Sign Method (FGSM)

Η πρώτη μέθοδος που παρουσιάζεται στο [40] προτείνει την εύρεση μιας επιθετικής εικόνας $\boldsymbol{x}_{adv} = \boldsymbol{x} + \boldsymbol{r}$, δεδομένης μιας αρχικής εικόνας $\boldsymbol{x}$, με την επίλυση του προβλήματος ελαχιστοποίησης

$$\begin{aligned} &\min_{\boldsymbol{r}} \; c\|\boldsymbol{r}\|_2 - J(f(\boldsymbol{x}+\boldsymbol{r}),\ell) \\ &\text{s.t.}: \quad \boldsymbol{x}+\boldsymbol{r} \in I \end{aligned} \tag{2.9}$$

όπου ℓ είναι η σωστή κατηγορία ταξινόμησης της εισόδου $\boldsymbol{x}$, $J(f(\boldsymbol{x}),\ell)$ είναι το λάθος της εξόδου του ταξινομητή για είσοδο $\boldsymbol{x}$, δεδομένης επιθυμητής εξόδου ℓ και I ο χώρος εισόδου.

Για την επίλυση του προβλήματος ελαχιστοποίησης χρησιμοποιείται η μέθοδος L-BFGS. Η προσεγγιστική αυτή μέθοδος επιλύει ένα δύσκολο πρόβλημα ελαχιστοποίησης με αποτέλεσμα να είναι χρονοβόρα η εύρεση επιθετικών σημείων στον χώρο εισόδου. Σαν λύση του προβλήματος αυτού, στο [11] προτάθηκε η Fast Gradient Sign Method με την οποία επιτυγχάνεται εύρεση μια προσεγγιστικής λύσης του προβλήματος με σημαντικά μικρότερο υπολογιστικό κόστος. Συγκεκριμένα η προσεγγιστική λύση για την εύρεση επιθετικών εικόνων $\boldsymbol{x}_{adv}$ χρησιμοποιώντας ως αρχική εικόνα την $\boldsymbol{x}$, προκύπτει ως εξής:

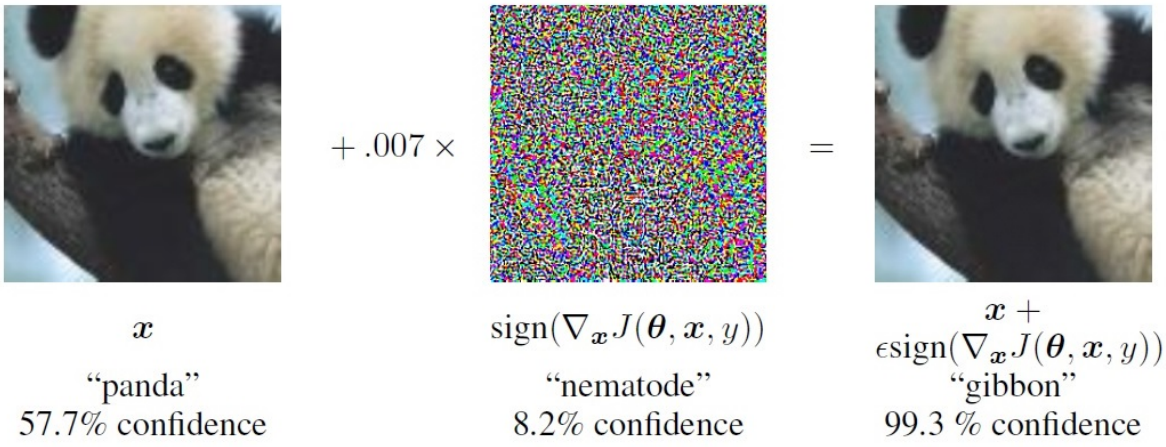
$$\boldsymbol{x}_{adv} = \boldsymbol{x} + \epsilon * \text{sign}(\nabla_x J(f(\boldsymbol{x}),\ell)) \tag{2.10}$$

Έτσι μπορούμε να βρούμε εύκολα την κατεύθυνση στην οποία πρέπει να κινηθούμε για να βρούμε το επιθετικό σημείο, υπολογίζοντας το $\text{sign}(\nabla_x J(f(\boldsymbol{x}), \ell))$ με χρήση *backpropagation*. Για να βρούμε το τελικό σημείο αρκεί να εκτελέσουμε μια αναζήτηση γραμμής ώστε να υπολογίσουμε το ϵ της Εξίσωσης (2.10) για το οποίο ισχύει ότι $f(\boldsymbol{x}_{adv}) \neq \ell$. Στο Κεφάλαιο 4 γίνεται σύγκριση των αποτελεσμάτων που μας δίνουν αυτές οι δύο μέθοδοι.

Επίσης στο [34] προτείνεται η Fast Gradient Value μέθοδος σαν παραλλαγή της FGSM. Στην μέθοδο αυτή δεν λαμβάνουμε το πρόσημο του gradient του δικτύου αλλά το ίδιο το gradient. Έτσι έχουμε ότι

$$\boldsymbol{x}_{adv} = \boldsymbol{x} + \epsilon * (\nabla_x J(f(\boldsymbol{x}), \ell)) \tag{2.11}$$

Η αφαίρεση της συνάρτησης πρόσημου έχει ως αποτέλεσμα να μεταβάλλονται περισσότερο pixel όπου το μέτρο του gradient είναι μεγαλύτερο, με αποτέλεσμα να είναι δυνατόν να πετύχουμε την ίδια μεταβολή της εξόδου με μικρότερη μεταβολή της εισόδου. Όπου την μεταβολή της εισόδου την μετράμε ώς $\|\boldsymbol{x}_{adv} - \boldsymbol{x}\|_1$



Σχήμα 2.5: Παράδειγμα παραγωγής επιθετικής εικόνας με χρήση της FGSM για νευρωνικό δίκτυο που έχει εκπαιδευτεί στο ImageNet,[11]

DeepFool

Μια ακόμα επιτυχημένη μέθοδος που είναι ταυτόχρονα και αρκετά απλή για την εύρεση επιθετικών σημείων είναι το *DeepFool* που παρουσιάζεται στο [30]. Η μέθοδος αυτή εκτελείται επαναληπτικά γραμμικοποιώντας τον ταξινομητή κοντά στο σημείο εισόδου και εκτελώντας βήματα σύμφωνα με το γραμμικό μοντέλο που θα οδηγήσουν την είσοδο στην λανθασμένη κατηγορία. Στην απλή περίπτωση ενός δυαδικού ταξινομητή όπου

- Αν $f(\boldsymbol{x}) > 0$ τότε το $\boldsymbol{x}$ ταξινομείται στην κατηγορία c_1
- Αν $f(\boldsymbol{x}) < 0$ τότε το $\boldsymbol{x}$ ταξινομείται στην κατηγορία c_2

Έχουμε ότι αν $f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + \boldsymbol{b}$ τότε το ελάχιστο διάνυσμα που αρκεί να προστεθεί στο $\boldsymbol{x}$ για να μεταβληθεί η κατηγορία του είναι

$$r(\boldsymbol{x}) = -\frac{f(\boldsymbol{x})}{\|\boldsymbol{w}\|_2^2}\boldsymbol{w} \tag{2.12}$$

Οπότε αν θεωρήσουμε ότι ο ταξινομητής δεν είναι αφινικός αλλά εκτελούμε γραμμικοποίηση, τότε σύμφωνα με την μέθοδο *DeepFool* εκτελούμε επαναληπτικά βήματα ανανεώνοντας την είσοδο

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \frac{f(\boldsymbol{x}_k)}{\|\nabla f(\boldsymbol{x}_k)\|_2^2} \nabla f(\boldsymbol{x}_k) \tag{2.13}$$

Μέχρι να βρεθεί $\boldsymbol{x}_k$ που ταξινομείται στην λάθος κατηγορία, όπου με $\boldsymbol{x}_k$ συμβολίζουμε την ανανεωμένη είσοδο $\boldsymbol{x}$ στην k επανάληψη της μεθόδου.

Έστω τώρα ότι έχουμε ένα γενικό συνελικτικό νευρωνικό δίκτυο με περισσότερες από δύο κατηγορίες, οπότε συμβολίζουμε την έξοδο για την c κατηγορία ως $f_c(\boldsymbol{x})$. Αν η αρχική εικόνα $\boldsymbol{x}_0$ ταξινομείται στην κατηγορία c_0 τότε η μέθοδος γενικεύεται ως εξής

$$\ell = \arg\min_{c \neq c_0} \frac{|f_c(\boldsymbol{x}_k) - f_{c_0}(\boldsymbol{x}_k)|}{\|\nabla f_c(\boldsymbol{x}_k) - \nabla f_{c_0}(\boldsymbol{x}_k)\|_2} \tag{2.14}$$

$$\boldsymbol{w}_k = \nabla f_\ell(\boldsymbol{x}_k) - \nabla f_{c_0}(x_k) \tag{2.15}$$

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \frac{|f_\ell(\boldsymbol{x}_k) - f_{c_0}(\boldsymbol{x}_k)|}{\|\boldsymbol{w}_k\|_2^2} \boldsymbol{w}_k \tag{2.16}$$

Όπου ομοίως με πριν, επαναλαμβάνεται η διαδικασία μέχρι να βρεθεί $\boldsymbol{x}_k$ που ταξινομείται σε λάθος κατηγορία.

C&W's Attack

Στο [4] παρουσιάζεται μια μέθοδος η οποία στοχεύει να δημιουργήσει επιθετικές εισόδους σε δίκτυα στα οποία έχει εφαρμοστεί η μέθοδος άμυνας του Defense Distillation που αναφέρεται στο Κεφάλαιο 2.4. Στην μέθοδο αυτή προτείνεται το παρακάτω συναρτησιακό, η ελαχιστοποίηση του οποίου μας δίνει την εικόνα $\boldsymbol{\eta}$ την οποία αν προσθέσουμε στο αρχικό πρότυπο $\boldsymbol{x}$ λαμβάνουμε την επιθετική εικόνα $\boldsymbol{x} + \boldsymbol{\eta}$.

$$\min_{\boldsymbol{\eta}} (\|\boldsymbol{\eta}\|_p + c * g(\boldsymbol{x} + \boldsymbol{\eta})) \tag{2.17}$$

Αν η αρχική κατηγορία αναγνώρισης είναι η $f(\boldsymbol{x}) = \ell$, η συνάρτηση $g(\boldsymbol{x})$ ορίζεται ως εξής

$$g(\boldsymbol{x}) = \max(\max_{i \neq \ell}(Z(\boldsymbol{x})_i) - Z(\boldsymbol{x})_\ell - k) \tag{2.18}$$

Όπου το $Z(\boldsymbol{x})_i$ είναι η έξοδος του $softmax$ επιπέδου του δικτύου για την i κατηγορία.

Επιπρόσθετα για την αποφυγή της χρήσης της μεθόδου L-BFGS όπου ορίζουμε περιορισμούς κουτιού στην μεταβλητή $\boldsymbol{\eta}$, προτείνεται η μεταβλητή $\boldsymbol{w}$ η οποία ικανοποιεί την σχέση $\boldsymbol{\eta} = \frac{1}{2}(\tanh(\boldsymbol{w}) + 1) - \boldsymbol{x}$. Έχοντας ορίσει την μεταβλητή $\boldsymbol{w}$, το πρόβλημα ελαχιστοποίησης που προτείνεται είναι το

$$\min_{\boldsymbol{w}} \left\| \frac{1}{2}(\tanh(\boldsymbol{w}) + 1) \right\|_2 + c * g\left(\frac{1}{2}\tanh(\boldsymbol{w}) + 1\right) \tag{2.19}$$

Παράλληλα με τον ορισμό του γενικού προβλήματος ελαχιστοποίησης έχουν προταθεί και διάφορες παραλλαγές οι οποίες βελτιώνουν τις επιδόσεις της μεθόδου. Μια τέτοια παραλλαγή είναι σε κάθε επανάληψη να εντοπίζονται τα pixel τα οποία δεν συνεισφέρουν στην μεταβολή της κατηγορίας και να αφαιρούνται από το σύνολο των pixel που ανανεώνονται. Σύμφωνα με το [4] η μέθοδος του Defense Distillation δεν μπορεί να προστατεύσει το δίκτυο από επιθετικές εικόνες οι οποίες έχουν παραχθεί από την C&W's Attack.

Universal Petrubation

Στα [29], [28] παρουσιάζονται μέθοδοι οι οποίες επιδιώκουν να βρουν ένα καθολικό διάνυσμα $\boldsymbol{\eta}$ για ένα σύνολο δεδομένων, με το οποίο μπορούμε να δημιουργήσουμε επιθετικές εικόνες. Συγκεκριμένα δεδομένης μιας σταθεράς $\epsilon > 0$ μιας σταθεράς $\delta \in [0, 1)$ και ενός νευρωνικού δικτύου οι μέθοδοι αυτές επιδιώκουν να βρουν ένα διάνυσμα $\|\boldsymbol{\eta}\|_2 \leq \epsilon$, το οποίο αν προστεθεί σε ένα σύνολο εικόνων, οι οποίες αρχικά ταξινομούνται σωστά από το δίκτυο, τότε το ποσοστό των εικόνων που οδηγούνται σε λάθος ταξινόμηση είναι μεγαλύτερο από δ.

Στο [29] αναφέρεται ότι χρησιμοποιώντας την μέθοδο που παρουσιάζεται για την εύρεση ενός καθολικού διανύσματος, καταφέρνουν να βρουν ένα διάνυσμα το οποίο οδηγεί σε ποσοστό λάθος αναγνώρισης που ξεπερνάει το 80% σε μια μεγάλη ποικιλία από βαθιά νευρωνικά δίκτυα.

Black Box Attack

Οι παραπάνω μέθοδοι έχουν ως κοινό χαρακτηριστικό ότι για την κατασκευή της επιθετικής εισόδου είναι απαραίτητη η γνώση των παραμέτρων του δικτύου που επιθυμούμε να ¨επιτεθούμε¨. Στα [31], [5], [32], αναπτύσσονται μεθοδολογίες οι οποίες δίνουν την δυνατότητα κατασκευής επιθετικών εισόδων χωρίς την γνώση των παραμέτρων του δικτύου. Συγκεκριμένα θεωρούμε ότι η μόνη ενέργεια που μπορούμε να εκτελέσουμε με τον ταξινομητή για τον οποίο ενδιαφερόμαστε να κατασκευάσουμε επιθετικές εισόδους, είναι η εισαγωγή ενός πεπερασμένου αριθμού εικόνων εισόδου και η καταγραφή των εξόδων του ταξινομητή.

Υπό αυτόν τον περιορισμό η μέθοδος Black Box Attack που αναπτύσσεται στο [32] χρησιμοποιεί έναν ευριστικό τρόπο για την κατασκευή συνθετικών εισόδων τις οποίες εισάγει στον ταξινομητή και καταγράφει τις εξόδους του. Οι συνθετικές αυτές εικόνες επικεντρώνονται σε κατευθύνσεις στον χώρο εισόδου όπου παρατηρείται η μεγαλύτερη διακύμανση των εξόδων του αρχικού ταξινομητή, και επιτρέπουν να περιορίσουν τον αριθμό των αναγκαίων εισόδων που πρέπει να εισάγουμε στον αρχικό ταξινομητή. Στην συνέχεια το συνθετικό σύνολο από ζεύγη εισόδου-εξόδου χρησιμοποιείται για την εκπαίδευση ενός δεύτερου ταξινομητή. Τέλος χρησιμοποιώντας μια από τις γνωστές μεθόδους που αναφέρθηκαν και παραπάνω, κατασκευάζονται επιθετικές είσοδοι για τον δεύτερο ταξινομητή, για τον οποίο γνωρίζουμε τις παραμέτρους. Οι επιθετικές είσοδοι του δεύτερου ταξινομητή επεκτείνονται στον πρώτο με ποσοστό επιτυχίας που σύμφωνα με το [32] φτάνει το 85% σε νευρωνικά δίκτυα που έχουν εκπαιδευτεί στο MNIST. Αξίζει επίσης να σημειωθεί ότι η μέθοδος αυτή φαίνεται να καταπολεμά μεθόδους άμυνας όπως το Defense Distillation το οποίο αναφέρεται στο Κεφάλαιο 2.4.

2.4 Τεχνικές άμυνας ενάντια σε επιθετικές εικόνες που έχουν προταθεί στην βιβλιογραφία

Παράλληλα με τεχνικές κατασκευής επιθετικών εικόνων, έχουν προταθεί και διαφορετικές τεχνικές άμυνας ενάντια σε επιθετικές εικόνες. Οι τεχνικές αυτές μπορούν να διαχωριστούν σε δύο μεγάλες κατηγορίες.

- Αποφυγή επιθετικών εικόνων με τροποποίηση του ταξινομητή: Η κατηγορία αυτή περιλαμβάνει τεχνικές οι οποίες στοχεύουν στην τροποποίηση των παραμέτρων του ταξινομητή προκειμένου να τον κάνουν πιο εύρωστο ενάντια στις επιθετικές εικόνες.

- Ανίχνευση επιθετικών εικόνων: Η κατηγορία αυτή θεωρεί έναν δεδομένο ταξινομητή, στον οποίο μπορούν να παραχθούν επιθετικές εικόνες, και στοχεύει στην ανίχνευση των

εικόνων αυτών ανάμεσα στο σύνολο των εικόνων που εισάγονται στον ταξινομητή .

Παρακάτω παρουσιάζουμε ορισμένες αντιπροσωπευτικές μεθόδους και από τις δύο κατηγορίες. Πέρα των μεθόδων άμυνας που παρουσιάζονται αναλυτικότερα παρακάτω επιπρόσθετες μέθοδοι εντοπισμού επιθετικών παραδειγμάτων προτείνονται στα [22], [27], [26]. Επίσης μια διαφορετική προσέγγιση άμυνας ενάντια σε επιθετικές εικόνες παρουσιάζεται στα [16], [12], στα οποία παρουσιάζονται μέθοδοι οι οποίες επιδιώκουν να εντοπίσουν περιοχές στον χώρο εισόδου του νευρωνικού δικτύου στις οποίες μπορούν να εγγυηθούν την σωστή λειτουργία του δικτύου.
Αρχικά οι κυριότερες τεχνικές που στοχεύουν στην τροποποίηση των παραμέτρων του ταξινομητή είναι οι παρακάτω:

Defense Distillation

Η διαδικασία αυτή περιγράφεται στο [33] και βασίζεται στην τεχνική του Distillation για νευρωνικά δίκτυα η οποίο προτάθηκε αρχικά στο [2] ως μια μέθοδος μείωσης του μεγέθους των νευρωνικών δικτύων. Αρχικά εκτελείται μια τυπική εκπαίδευση ενός νευρωνικού δικτύου F_1 το οποίο έχει ώς έξοδο ένα Softmax επίπεδο. Στην εκπαίδευση αυτή για την είσοδο $\boldsymbol{x}$, η αναμενόμενη έξοδος με βάση την οποία γίνεται η εκπαίδευση είναι ένα διάνυσμα πιθανοτήτων $Y(\boldsymbol{x})$, το οποίο δίνει πιθανότητα 1 στην κατηγορία που αντιστοιχεί η είσοδος και 0 στις υπόλοιπες κατηγορίες. Η απαίτηση να γίνεται αλλαγή από 0 σε 1 χωρίς ενδιάμεσες τιμές δημιουργεί περιοχές του χώρου εισόδου στις οποίες έχουμε πολύ μεγάλες μεταβολές των εξόδων με μικρές μεταβολές της εισόδου. Το πρόβλημα αυτό η μέθοδος του Distillation το αντιμετωπίζει εκτελώντας εκπαίδευση ενός δεύτερου νευρωνικού το οποίο έχει την ίδια αρχιτεκτονική με το πρώτο, αλλά η αναμενόμενη έξοδος για την είσοδο $\boldsymbol{x}$ δέν είναι το $Y(\boldsymbol{x})$ αλλά η έξοδος του πρώτου νευρωνικού $F_1(\boldsymbol{x})$ που είναι ένα διάνυσμα πιθανοτήτων που δίνει την μεγαλύτερη πιθανότητα στην σωστή κατηγορία (δεδομένου ότι η εκπαίδευση έχει ολοκληρωθεί με επιτυχία) αλλά η πιθανότητα αυτή δεν είναι απαραίτητα 1 .

Με την διαδικασία αυτή επιτυγχάνεται το δεύτερο δίκτυο να έχει πιο λεία κλίση στις περιοχές γύρω από τα σημεία εκπαίδευσης με αποτέλεσμα να γίνεται πιο δύσκολη η εύρεση επιθετικών σημείων και τα σημεία αυτά τείνουν να απαιτούν μεγαλύτερες μεταβολές της αρχικής εισόδου. Η τεχνική του *distillation* αποτυγχάνει όταν χρησιμοποιείται η διαδικασία Black Box Attack που αναφέρθηκε στο Κεφάλαιο 2.3, η οποία καταφέρνει να βρίσκει επιθετικά σημεία στο νευρωνικό μετά το *distillation* το ίδιο εύκολα με το αρχικό νευρωνικό. Αυτό επιτυγχάνεται γιατί η εξομάλυνση της κλίσης γίνεται μόνο τοπικά σε περιοχές γύρω από το σύνολο εκπαίδευσης, με αποτέλεσμα να δυσκολεύει την εύρεση των επιθετικών σημείων μόνο όταν χρησιμοποιείται το διάνυσμα κλίσης του νευρωνικού στις περιοχές αυτές.

Adversarial Training

Μια άλλη μέθοδος άμυνας που έχει προταθεί στο [11], είναι η εύρεση των επιθετικών σημείων χρησιμοποιώντας μια από τις μεθόδους που παρουσιάστηκαν στο Κεφάλαιο 2.3 (κυρίως μια από τις FGSM,DeepFool), η ενσωμάτωση τους στο σύνολο εκπαίδευσης μετά από μια δεδομένη εποχή εκπαίδευσης και η συνέχιση της εκπαίδευσης με το ενισχυμένο με τις επιθετικές εικόνες σύνολο εκπαίδευσης .

Αυτή η μέθοδος παράγει πολύ καλά αποτελέσματα και δυσκολεύει την εύρεση επιθετικών σημείων ακόμα και με την διαδικασία black box attack. Τα προβλήματα της μεθόδου αυτής είναι ότι η επιτυχία της εξαρτάται σημαντικά από το ποσοστό των επιθετικών σημείων που εισάγουμε στο σύνολο εκπαίδευσης καθώς και από το σημείο κατά την διαδικασία εκπαίδευσης στο οποίο αρχίζουμε να μεταβάλουμε το σύνολο εκπαίδευσης. Τέλος η μέθοδος αυτή μειώνει

το ποσοστό επιτυχίας με το οποίο μπορεί να παραχθεί μια επιθετική είσοδος, η οποία δεν διαφέρει οπτικά από μια πραγματική είσοδο, όμως οι επιθετικές είσοδοι τις οποίες καταφέρνουμε τελικά να βρούμε με επιτυχία, συνεχίζουν να ταξινομούνται λάθος από το νευρωνικό δίκτυο με μεγάλη βεβαιότητα.

Mixup Training

Η μέθοδος αυτή παρουσιάζεται στο [44] και συνδυάζει ιδέες από τις προηγούμενες μεθόδους που παρουσιάστηκαν. Σύμφωνα με το [44], κατά την εκπαίδευσή του, το νευρωνικό δίκτυο μαθαίνει να αναγνωρίζει με μεγάλη επιτυχία δείγματα που ανήκουν στην κατανομή των δεδομένων εκπαίδευσης, αλλά εμφανίζει μεγάλες μεταβολές στις εξόδους του για δείγματα που βρίσκονται σε μικρή απόσταση, αλλά έξω από την κατανομή των δεδομένων εκπαίδευσης. Για την αντιμετώπιση του προβλήματος αυτού προτείνεται η εισαγωγή στο σύνολο εκπαίδευσης προτύπων που αποτελούν γραμμικό συνδυασμό των προτύπων του αρχικού συνόλου εκπαίδευσης με επιθυμητή έξοδο τον αντίστοιχο γραμμικό συνδυασμό των επιθυμητών εξόδων των αρχικών προτύπων. Έτσι για $\lambda \in [0,1]$ και $\boldsymbol{x}_i, \boldsymbol{x}_j$ δύο εικόνες του συνόλου εκπαίδευσης, εισάγουμε στο σύνολο εκπαίδευσης την εικόνα $\boldsymbol{x}_{\text{new}}$ για την οποία ισχύει

$$\boldsymbol{x}_{\text{new}} = \lambda \boldsymbol{x}_i + (1-\lambda)\boldsymbol{x}_j$$

$$f(\boldsymbol{x}_{\text{new}}) = \lambda f(\boldsymbol{x}_i) + (1-\lambda) f(\boldsymbol{x}_j)$$

Έτσι το σύνολο εκπαίδευσης αποκτά μια πιο γενικευμένη κατανομή, με αποτέλεσμα το δίκτυο να γενικεύει καλύτερα και να μην έχει μεγάλες μεταβολές σε σημεία του χώρου εισόδου έξω από την κατανομή του αρχικού συνόλου εκπαίδευσης. Η μέθοδος επιτυγχάνει να βελτιώνει την επίδοση του δικτύου και σε σύνολα δεδομένων όπως το CIFAR και το Imagenet στα οποία αναμένουμε η συνάρτηση ταξινόμησης να εμφανίζει σημαντικές μη-γραμμικότητες.

Παράλληλα με τις παραπάνω μεθόδους έχουν προταθεί οι παρακάτω μέθοδοι οι οποίες στοχεύουν στον εντοπισμό ενός επιθετικού παραδείγματος κατά την λειτουργία του εκπαιδευμένου δικτύου

Artifact Detection

Στο [10] προτείνονται δύο διαφορετικοί τρόποι εντοπισμού επιθετικών εικόνων. Η πρώτη μέθοδος βασίζεται στην παρατήρηση ότι οι επιθετικές εικόνες δεν ανήκουν στο manifold στο οποίο ανήκουν τα πραγματικά δεδομένα εισόδου, ενώ παράλληλα οι επιθετικές εικόνες βρίσκονται πιο κοντά στο submanifold στο οποίο ανήκουν οι εικόνες της πραγματικής τους κατηγορίας. Θεωρούμε ότι οι έξοδοι των τελευταίων επιπέδων του νευρωνικού δικτύου αποτελούν διανύσματα χαρακτηριστικών για τις εικόνες που εισάγονται σαν είσοδοι στο δίκτυο. Με βάση την παρατήρηση αυτή υπολογίζεται η κατανομή στην οποία ανήκουν τα διανύσματα χαρακτηριστικών που προκύπτουν ως έξοδοι του νευρωνικού δικτύου όταν σαν είσοδο έχουμε πραγματικά δεδομένα. Συγκεκριμένα αν Y_c το σύνολο των διανυσμάτων χαρακτηριστικών των δεδομένων εκπαίδευσης που ανήκουν σε μια κατηγορία c, $\boldsymbol{y}$ το διάνυσμα χαρακτηριστικών της εικόνας εισόδου, υπολογίζουμε την $\hat{f}_c(\boldsymbol{y})$ η οποία αποτελεί την εκτίμηση της πυκνότητας της κατανομής των πραγματικών χαρακτηριστικών της κατηγορίας c στο σημείο $\boldsymbol{y}$ ως εξής

$$\hat{f}_c(\boldsymbol{y}) = \frac{1}{|Y_c|} \sum_{\boldsymbol{y}_i \in Y_c} \exp\left(\frac{-\|\boldsymbol{y} - \boldsymbol{y}_i\|_2^2}{\sigma^2}\right) \tag{2.20}$$

όπου με $|Y_c|$ συμβολίζουμε τον αριθμό των στοιχείων του συνόλου Y_c.
Σύμφωνα με την μέθοδο αυτή στην περίπτωση μιας επιθετικής εισόδου με πραγματική κατηγορία την c_1, η οποία αναγνωρίζεται ως η κατηγορία c_2 θα ισχύει ότι $\hat{f}_{c_1}(\boldsymbol{x}) > \hat{f}_{c_2}(\boldsymbol{x})$.

Η δεύτερη μέθοδος βασίζεται στην εξαγωγή Bayesian αβεβαιότητας από το δίκτυο. Συγκεκριμένα προτείνεται ένας τρόπος εξαγωγής Bayesian αβεβαιότητας από ένα νευρωνικό δίκτυο το οποίο έχει εκπαιδευτεί χρησιμοποιώντας την dropout μέθοδο.

Έστω μια είσοδος $\boldsymbol{x}$ που επιθυμούμε να διαπιστώσουμε αν είναι επιθετική ή πραγματική. Για T διαφορετικά σύνολα παραμέτρων του δικτύου από την dropout μέθοδο λαμβάνουμε τις εξόδους $\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_T$ για είσοδο $\boldsymbol{x}$. Η αβεβαιότητα $U(\boldsymbol{x})$ του δικτύου στο σημείο $\boldsymbol{x}$ υπολογίζεται από την σχέση

$$U(\boldsymbol{x}) = \frac{1}{T}\sum_{i=1}^{T} \boldsymbol{y}_i^T \boldsymbol{y}_i - (\frac{1}{T}\sum_{i=1}^{T} \boldsymbol{y}_i)^\intercal (\frac{1}{T}\sum_{i=1}^{T} \boldsymbol{y}_i) \tag{2.21}$$

Δεδομένου της παραδοχής ότι οι επιθετικές εικόνες εμφανίζονται σε περιοχές του δικτύου όπου υπάρχει μεγάλη αβεβαιότητα, το $U(\boldsymbol{x})$ αποτελεί χρήσιμη μετρική για τον καθορισμό αν μια εικόνα $\boldsymbol{x}$ είναι επιθετική. Έτσι το [10] προτείνει τον συνδυασμό των δύο παραπάνω μετρικών για τον εντοπισμό πιθανών επιθετικών εικόνων.

PixelDefend

Στο [35] παρουσιάζεται το μοντέλο PixelCnn το οποίο κατά την εκπαίδευση προσεγγίζει το μοντέλο το οποίο παράγει τις εικόνες του συνόλου εκπαίδευσης και δεδομένης μιας καινούριας εικόνας υπολογίζει την πιθανότητα η εικόνα αυτή να έχει παραχθεί από το μοντέλο αυτό. Συγκεκριμένα για μια εικόνα εισόδου $\boldsymbol{x}$, διαβάζει ακολουθιακά τα pixel x_i της εικόνας, και παράγει την συνολική πιθανότητα της εικόνας $p(\boldsymbol{x})$ δεδομένου του μοντέλου που έχει παραχθεί από την εκπαίδευση. Η υλοποίηση του PixelCnn γίνεται με διαδοχικά συνελικτικά και ReLU επίπεδα.

Στο [39], παρατηρείται ότι στην περίπτωση των επιθετικών εικόνων η πιθανότητα $p(\boldsymbol{x})$ που προκύπτει από το PixelCnn είναι συστηματικά μικρότερη από την πιθανότητα των πραγματικών εικόνων. Έτσι προτείνουν την χρήση της τιμής $p(\boldsymbol{x})$ για την ανίχνευση επιθετικών εικόνων. Παράλληλα προτείνεται η μέθοδος PixelDefend η οποία σκοπεύει στην ανακατασκευή της αρχικής εικόνας από την οποία παράχθηκε η επιθετική εικόνα, βρίσκοντας μια εικόνα $\boldsymbol{x}^*$ για την οποία ισχύει ότι

$$\begin{aligned} &\max_{x^*} p(\boldsymbol{x}^*) \\ &\text{s.t.:} \quad \|\boldsymbol{x} - \boldsymbol{x}^*\|_2 \leq \epsilon_{\text{defend}} \end{aligned}$$

Κεφάλαιο 3

Lipschitz συνέχεια και επιθετικά παραδείγματα

3.1 Ο Scattering μετασχηματισμός και η επέκτασή του για πολυκαναλικά συνελικτικά δίκτυα

Στο [24] (Group Invariant Scattering) ορίζεται ο *scattering* μετασχηματισμός. Για τον ορισμό του μετασχηματισμού αυτού γίνεται χρήση ενός wavelet μετασχηματισμού όπου $\psi \in \boldsymbol{L}^2(\mathbb{R}^d)$ η μητρική συνάρτηση wavelet και ϕ η αντίστοιχη συνάρτηση κλίμακας (scaling function). Η μητρική συνάρτηση ψ διαστέλλεται κατα 2^{-j} και περιστρέφεται κατά $r \in G$, όπου G μια πεπερασμένη ομάδα περιστροφών, για να παράξει την συνάρτηση $\psi_{2^j r}$ όπου:

$$\psi_{2^j r}(x) = 2^j \psi(2^j r^{-1} x) \tag{3.1}$$

Για την απλοποίηση των συμβολισμών γράφουμε $\lambda = 2^j r \in 2^{\mathbb{Z}} \times G = \Lambda_\infty$ και ορίζουμε $\Lambda_J = \{\lambda = 2^j r : r \in G, 2^j > 2^{-J}\}$. Οι συναρτήσεις ψ_λ με $\lambda \in \Lambda_J$ καλύπτουν τις συχνότητες $2^j > 2^{-J}$, οπότε τις χαμηλότερες συχνότητες τις καλύπτει η

$$\phi_{2^J} = 2^{-J} \phi(2^{-J} x) \tag{3.2}$$

Με τα προηγούμενα λοιπόν ορίζεται ο scattering propagator για μια συνάρτηση $g \in \boldsymbol{L}^2(\mathbb{R}^d)$ ως εξής:

$$U[p] = U[\lambda_m] \dots U[\lambda_2] U[\lambda_1] \tag{3.3}$$

όπου $U[\lambda] g = |g * \psi_\lambda|$ ένας τελεστής, με ψ_λ *wavelet* κλίμακας λ, στον οποίο με $g * \psi_\lambda$ συμβολίζουμε την συνέλιξη μεταξύ των g, ψ_λ και $p = (\lambda_1, \lambda_2, \dots \lambda_m) \quad \lambda_i \in \Lambda_\infty$ ένα μονοπάτι.

Με χρήση του scattering propagator ορίζεται ο *windowed scattering transform* στο μονοπάτι $p = (\lambda_1, \lambda_2, \dots \lambda_m) \quad \lambda_i \in \Lambda_J$ ως εξής:

$$S[p] g = (U[p] g) * \phi_{2^J} \tag{3.4}$$

Για τον παραπάνω μετασχηματισμός στο [24] αποδεικνύεται ότι είναι αμετάβλητος κατά την μετατόπιση της εισόδου. Επίσης αποδεικνύεται ότι είναι Lipschitz συνεχής όταν στην συνάρτηση εισόδου g, δρα ο τελεστής $L_\tau g(x) = g(x - \tau(x))$ όπου $\|\nabla \tau\|_\infty < 1$. Ο περιορισμός $\|\nabla \tau\|_\infty < 1$, σημαίνει ότι ο τελεστής L_τ είναι αρκετά κοντά σε έναν τελεστή μετατόπισης.

Ο μετασχηματισμός αυτός, όπως παρουσιάζεται και στο [25], μπορεί να προσεγγίζει την λειτουργία μιας απλής αρχιτεκτονικής συνελικτικού νευρωνικού όπου δεν υπάρχει αλληλεπίδραση μεταξύ των καναλιών. Η προσέγγιση αυτή μας επιτρέπει να μελετήσουμε τον τρόπο

με τον οποίο τα νευρωνικά δίκτυα πετυχαίνουν την επίλυση δύσκολων προβλημάτων, τα οποία απαιτούν την πολυκλιμακωτή εξαγωγή χαρακτηριστικών. Παράλληλα οι ιδιότητες που αναφέρθηκαν παραπάνω εξηγούν το τρόπο με τον οποίο ένα τέτοιο απλοποιημένο νευρωνικό μπορεί να πετύχει αμεταβλητότητα στην μετατόπιση της εισόδου καθώς και σε μικρές παραμορφώσεις της εισόδου όπως η ελαστική παραμόρφωση που παρουσιάζεται στο [37]. Είναι σημαντικό να διευκρινιστεί ότι στο κεφάλαιο αυτό καθώς και στο Κεφάλαιο 3.2, όταν προσεγγίζουμε το νευρωνικό δίκτυο με χρήση του Scattering μετασχηματισμού και της επέκτασής του θεωρούμε ότι η είσοδος έχει την μορφή μιας πραγματικής συνάρτησης που ανήκει στο $\boldsymbol{L}^2(\mathbb{R}^d)$.

Για να μπορέσει το νευρωνικό δίκτυο να επιλύσει πιο περίπλοκα προβλήματα όπου είναι απαραίτητη η αμεταβλητότητα όχι μόνο σε μετατόπιση της εισόδου αλλά και στην δράση άλλων τελεστών, πρέπει να επιτρέψουμε για την ανάλυση να υπάρχει αλληλεπίδραση μεταξύ των διαφορετικών καναλιών σε κάθε επίπεδο. Δηλαδή η έξοδος $h_i^{(k)}$ του καναλιού i στο επίπεδο k να προκύπτει ως εξής:

$$h_i^{(k)}(u) = \left| \sum_j (h_j^{(k-1)} * F_{j,i})(u) \right| \tag{3.5}$$

όπου $h_i^{(k)} \in \boldsymbol{L}^2(\mathbb{R}^d)$ πραγματική συνάρτηση και $F_{j,i}$ η κρουστική απόκριση του φίλτρου που ενώνει το j κανάλι της εισόδου του επιπέδου με το i κανάλι της εξόδου του επιπέδου.
Έτσι χρειάζεται να επεκτείνουμε τον *scattering* μετασχηματισμό προκειμένου να ενσωματώσουμε την αλληλεπίδραση μεταξύ των καναλιών. Οπότε δεδομένου ότι η είσοδος του j καναλιού είναι είναι μια πραγματική συνάρτηση $h_j \in \boldsymbol{L}^2(\mathbb{R}^d)$, έχουμε ότι η συνολική είσοδος του επιπέδου, το οποίο αποτελείται από N κανάλια, είναι $h = \{h_1, h_2, ..., h_N\}$ και συμβολίζουμε ως $\|h_j\|$ την L_2 νόρμα της h_j και επίσης:

$$\|h\| = \left(\sum_{j=1}^{N} \|h_j\|^2 \right)^{1/2}$$

Έτσι έχουμε τους παρακάτω τελεστές

$$(Uh)_i(x) = |\sum_{j=1}^{N} \int h_j(u)\psi_{(j,i)}(x-u)du|$$

$$U[m]h = U(U[m-1]h) \quad U[1] = Uh$$

$$(Ah)_i(x) = \int h_i(u)\phi(x-u)du$$

$$S[m]h = A(U[m]h)$$

Στην περίπτωση αυτού του μετασχηματισμού οι συναρτήσεις $\psi_{(j,i)}$ θεωρούμε ότι μπορεί να είναι διαφορετικά ζωνοπερατά φίλτρα, ενώ η συνάρτηση ϕ θεωρούμε ότι αποτελεί ένα βαθυπερατό φίλτρο.

3.2 Lipschitz συνέχεια στον επεκτεταμένο Scattering μετασχηματισμό

Μια ιδιότητα που μας ενδιαφέρει να διερευνήσουμε στον επεκτεταμένο μετασχηματισμό είναι η *Lipschitz* συνέχεια. Συγκεκριμένα αν ισχύει ότι $\|S[m]g - S[m]h\| \leq \|g - h\|$. Ομοίως με πριν θεωρώντας ότι έχουμε N κανάλια συμβολίζουμε :

$$(Wh)_i(x) = \sum_{j=1}^{N} \int h_j(u)\psi_{(j,i)}(x-u)du$$

$$\hat{h}(\omega) = \int h(x)e^{-jx\omega}dx$$

Οπότε αρχικά θέλουμε να δείξουμε ότι $\|Wh\| \leq \|h\|$, δεδομένου ότι για κάθε ω ισχύει ότι $\sum_{i=1}^{N} |\hat{\psi}_{(j,i)}(\omega)|^2 \leq \frac{1}{N}$ $\forall j \in 1, 2, ..., N$. Έχουμε λοιπόν ότι:

$$\|Wh\|^2 = \sum_{i=1}^{N} \| \sum_{j=1}^{N} \int h_j(u)\psi_{(j,i)}(x-u)du\|^2 \tag{3.6}$$

$$\begin{aligned} \| \sum_{j=1}^{N} \int h_j(u)\psi_{(j,i)}(x-u)du\|^2 &\leq N \sum_{j=1}^{N} \| \int h_j(u)\psi_{(j,i)}(x-u)du\|^2 \\ &= N \sum_{j=1}^{N} \frac{1}{2\pi} \|\hat{h}_j(\omega)\hat{\psi}_{(j,i)}(\omega)\|^2 \end{aligned} \tag{3.7}$$

Από τις (3.6),(3.7) έχουμε:

$$\begin{aligned} \|Wh\|^2 &= \sum_{i=1}^{N} \| \sum_{j=1}^{N} \int h_j(u)\psi_{(j,i)}(x-u)du\|^2 \\ &\leq \sum_{i=1}^{N} \frac{N}{2\pi} \sum_{j=1}^{N} \|\hat{h}_j(\omega)\hat{\psi}_{(j,i)}(\omega)\|^2 \\ &= \sum_{i=1}^{N} \frac{N}{2\pi} \sum_{j=1}^{N} \int |\hat{h}_j(\omega)\hat{\psi}_{(j,i)}(\omega)|^2 d\omega \\ &= \frac{N}{2\pi} \sum_{j=1}^{N} \int |\hat{h}_j(\omega)|^2 \sum_{i=1}^{N} |\hat{\psi}_{(j,i)}(\omega)|^2 d\omega \\ &\leq \sum_{j=1}^{N} \frac{1}{2\pi} \int |\hat{h}_j(\omega)|^2 d\omega = \|h\|^2 \\ &\implies \|Wh\| \leq \|h\| \end{aligned} \tag{3.8}$$

Οπότε

$$\begin{aligned} \|Uh - Ug\| = \||Wh| - |Wg|\| &\leq \|Wh - Wg\| \\ &= \| \sum_{j=1}^{N} \int h_j(u)\psi_{(j,i)}(x-u)du - \sum_{j=1}^{N} \int g_j(u)\psi_{(j,i)}(x-u)du\| \\ &= \| \sum_{j=1}^{N} \int (h_j(u) - g_j(u))\psi_{(j,i)}(x-u)du\| \\ &= \|W(h-g)\| \leq \|h-g\| \end{aligned} \tag{3.9}$$

Έτσι από την (3.9) έχουμε

$$\begin{aligned} &\|U[m]h - U[m]g\| = \|U(U[m-1]h) - U(U[m-1]g)\| \leq \|U[m-1]h - U[m-1]g\| \\ &\implies \|U[m]h - U[m]g\| \leq \|h-g\| \end{aligned} \tag{3.10}$$

Ομοίως με την (3.8) μπορούμε να δείξουμε ότι $\|Ah\| \leq \|h\|$, δεδομένου ότι $|\hat{\phi}(\omega)|^2 \leq 1$.

$$\begin{aligned}\|Ah\|^2 = \sum_{i=1}^{N} \| \int h_i(u)\phi(x-u)du\|^2 &= \frac{1}{2\pi}\sum_{i=1}^{N} \|\hat{h}_i(\omega)\hat{\phi}(\omega)\|^2 \\ &= \frac{1}{2\pi}\sum_{i=1}^{N}\int |\hat{h}_i(\omega)|^2|\hat{\phi}(\omega)|^2 d\omega \\ &\leq \sum_{i=1}^{N}\frac{1}{2\pi}\int |\hat{h}_i(\omega)|^2 d\omega = \|h\|^2 \\ \implies \|Ah\| \leq \|h\| \qquad (3.11)\end{aligned}$$

Έτσι τελικά έχουμε ότι

$$\begin{aligned}\|S[m]h - S[m]g\| = \|A(U[m]h) - A(U[m]g)\| = \|A(U[m]h - U[m]g)\| \leq \|U[m]h - U[m]g\| \\ \leq \|h-g\| \quad (3.12)\end{aligned}$$

Χρησιμοποιώντας την σχέση (3.12) μπορούμε να μελετήσουμε την συμπεριφορά του μετασχηματισμού όταν σαν είσοδο εισάγουμε επιθετικές εικόνες. Συγκεκριμένα όπως περιγράφεται στο Κεφάλαιο 2, οι επιθετικές εικόνες παράγονται όταν στην είσοδο h προσθέσουμε μια πολύ μικρή μεταβολή p οπότε λαμβάνουμε την νέα είσοδο $h+p$, η οποία με κατάλληλα επιλεγμένη συνάρτηση εισόδου p ταξινομείται διαφορετικά από την αρχική είσοδο. Από την (3.12) έχουμε ότι:

$$\|S[m](h+p) - S[m](h)\| \leq \|p\| \qquad (3.13)$$

Έτσι προκύπτει ότι η έξοδος για μια επιθετική εικόνα δεν διαφέρει κατά μέτρο από την αρχική είσοδο περισσότερο από $\|p\|$. Οπότε αν ο μετασχηματισμός ακολουθεί τους περιορισμούς $\sum_{i=1}^{N}|\hat{\psi}_{(j,i)}(\omega)|^2 \leq \frac{1}{N}$,$|\hat{\phi}(\omega)|^2 \leq 1$ για κάθε ω, τότε παρουσιάζει σχετική ομαλότητα στην έξοδο, καθώς μικρές μεταβολές της εισόδου δεν προκαλούν μεγάλες μεταβολές στην έξοδο. Προφανώς κάθε μετασχηματισμός θα ικανοποιεί για κάθε ω τις συνθήκες $\sum_{i=1}^{N}|\hat{\psi}_{(j,i)}(\omega)|^2 \leq \frac{C^2}{N}$, $|\hat{\phi}(\omega)|^2 \leq C^2$ για κάποιο C, οπότε ακολουθώντας την ίδια λογική απόδειξης έχουμε ότι

$$\|Uh - Ug\| \leq C\|h-g\| \qquad (3.14)$$

και

$$\begin{aligned}\|U[m]h - U[m]g\| = \|U(U[m-1]h) - U(U[m-1]g)\| \leq C\|U[m-1]h - U[m-1]g\| \\ \implies \|U[m]h - U[m]g\| \leq C^m\|h-g\| \qquad (3.15)\end{aligned}$$

οπότε τελικά αντί της εξίσωσης (3.13) θα έχουμε την εξίσωση

$$\|S[m](h+p) - S[m](h)\| \leq C^{m+1}\|p\| \qquad (3.16)$$

Σε αυτή την περίπτωση η σταθερά C αποτελεί καθοριστικό παράγοντα για το πόσο ευάλωτος είναι ο μετασχηματισμός σε μεταβολές της εισόδου κατα p.

3.3 Η εξέλιξη της σταθεράς Lipschitz κατά την εκπαίδευση με επιθετικές εικόνες

Καθώς η σταθερά Lipschitz καθορίζει την ικανότητα του ταξινομητή να αντιμετωπίζει επιθετικές εικόνες , επιθυμούμε να ανιχνεύσουμε πως εξελίσσεται η σταθερά κατά την εκπαίδευση ενός συνελικτικού νευρωνικού δικτύου όταν χρησιμοποιούμε τις μεθόδους άμυνας που αναπτύχθηκαν στο Κεφάλαιο 2.4.

Στην ενότητα αυτή θεωρώ ότι η είσοδος του συνελικτικού νευρωνικού δικτύου έχει την μορφή διανύσματος. Οπότε συμβολίζω ως $f(\boldsymbol{x}_{in}, c)$ την έξοδο του δικτύου για την κατηγορία c με είσοδο το διάνυσμα εικόνας $\boldsymbol{x}_{in}$. Έστω ότι έχουμε δύο διαφορετικά διανύσματα εισόδου $\boldsymbol{y}_{in}$,$\boldsymbol{h}_{in}$ και τις αντίστοιχες εξόδους $f(\boldsymbol{y}_{in}, c)$,$f(\boldsymbol{h}_{in}, c)$, τότε συμβολίζω ως $\boldsymbol{y}_{ik}$,$\boldsymbol{h}_{ik}$ τις εξόδους του k-οστού επιπέδου στο κανάλι i για κάθε μια από τις δύο εισόδους.

Το συνελικτικό νευρωνικό δίκτυο που μελετάμε αποτελείται από 3 είδη επιπέδων

1. Συνελικτικά επίπεδα
2. *Pooling* επίπεδα
3. *ReLU* επίπεδα

Οπότε για κάθε ένα από τα 3 είδη επιπέδων έχουμε ότι
1) Έστω το επίπεδο k αποτελεί συνελικτικό επίπεδο. Καθώς εκφράζουμε τις εικόνες ως μονοδιάστατα διανύσματα, η συνέλιξη με έναν δισδιάστατο πυρήνα ψ_{ijk}, που συνδέει το i κανάλι της εξόδου με το j κανάλι της εισόδου, υλοποιείται με πολλαπλασιασμό του διανύσματος εισόδου με πίνακα $\boldsymbol{A}_{ijk}$ που παράγεται από τον αρχικό πυρήνα. Έτσι έχουμε ότι:

$$\boldsymbol{x}_{ik} = \sum_{j=1}^{N_k} \boldsymbol{A}_{ijk}\boldsymbol{x}_{j(k-1)} \quad i = 1, 2, ..., M_k \tag{3.17}$$

όπου N_k είναι ο αριθμός των καναλιών της εισόδου και M_k ο αριθμός των καναλιών της εξόδου του συνελικτικού επιπέδου k. Άρα

$$\begin{aligned}\|\boldsymbol{y}_{ik} - \boldsymbol{h}_{ik}\|_2 = \|\sum_{j=1}^{N_k} \boldsymbol{A}_{ijk}\boldsymbol{y}_{j(k-1)} - \sum_{j=1}^{N_k} \boldsymbol{A}_{ijk}\boldsymbol{h}_{j(k-1)}\|_2 = \|\sum_{j=1}^{N_k} \boldsymbol{A}_{ijk}(\boldsymbol{y}_{j(k-1)} - \boldsymbol{h}_{j(k-1)})\|_2 \\ \leq \sum_{j=1}^{N_k} \left\|\boldsymbol{A}_{ijk}(\boldsymbol{y}_{j(k-1)} - \boldsymbol{h}_{j(k-1)})\right\|_2 \leq \sum_{j=1}^{N_k} \|\boldsymbol{A}_{ijk}\|_2 \left\|\boldsymbol{y}_{j(k-1)} - \boldsymbol{h}_{j(k-1)}\right\|_2\end{aligned} \tag{3.18}$$

$$\implies \|\boldsymbol{y}_{ik} - \boldsymbol{h}_{ik}\|_2 \leq \sum_{j=1}^{N_k} \|\boldsymbol{A}_{ijk}\|_2 \left\|\boldsymbol{y}_{j(k-1)} - \boldsymbol{h}_{j(k-1)}\right\|_2 \tag{3.19}$$

2) Έστω ότι το επίπεδο k αποτελεί Pooling επίπεδο. Αν θεωρήσουμε ότι κατά το pooling δεν έχουμε επικάλυψη των περιοχών, τότε στο [42] αποδεικνύεται ότι

$$\|\boldsymbol{y}_{ik} - \boldsymbol{h}_{ik}\|_2 \leq \left\|\boldsymbol{y}_{i(k-1)} - \boldsymbol{h}_{i(k-1)}\right\|_2 \tag{3.20}$$

3) Έστω ότι το επίπεδο k αποτελεί ReLU επίπεδο. Τότε ισχύει ότι αν το διάνυσμα εξόδου του επιπέδου έχει την μορφή

$$\boldsymbol{x}_{ik} = \begin{bmatrix} x_{ik}(1) \\ x_{ik}(2) \\ \vdots \\ x_{ik}(m) \end{bmatrix}$$

Η έξοδος $x_{ik}(t)$ προκύπτει ως εξής

$$x_{ik}(t) = \max(0, x_{i(k-1)}(t))$$

Επομένως ισχύει ότι

$$\begin{aligned} \|\boldsymbol{y}_{ik} - \boldsymbol{h}_{ik}\|_2^2 &= \sum_{t=1}^{m} |\max(0, y_{i(k-1)}(t)) - \max(0, h_{i(k-1)}(t))|^2 \\ &\leq \sum_{t=1}^{m} |y_{i(k-1)}(t) - h_{i(k-1)}(t)|^2 = \left\|\boldsymbol{y}_{i(k-1)} - \boldsymbol{h}_{i(k-1)}\right\|_2^2 \\ &\implies \|\boldsymbol{y}_{ik} - \boldsymbol{h}_{ik}\|_2 \leq \left\|\boldsymbol{y}_{i(k-1)} - \boldsymbol{h}_{i(k-1)}\right\|_2 \end{aligned} \tag{3.21}$$

όπου χρησιμοποιήθηκε το ότι $|\max(0, a) - \max(0, b)| \leq |a - b|$

Χρησιμοποιώντας τις εξισώσεις (3.19),(3.20),(3.21) μπορούμε να βρούμε μια σταθερά L_{ik} για την οποία ισχύει ότι

$$\|\boldsymbol{y}_{ik} - \boldsymbol{h}_{ik}\|_2 \leq L_{ik} \|\boldsymbol{y}_{10} - \boldsymbol{h}_{10}\|_2 \tag{3.22}$$

Την σταθερά την ορίζουμε αναδρομικά ως εξής, $L_{10} = 1$ και

1. Συνελικτικό επίπεδο. Από τον τύπο (3.19) ισχύει ότι

$$\|\boldsymbol{y}_{ik} - \boldsymbol{h}_{ik}\|_2 \leq \sum_{j=1}^{N_k} \|\boldsymbol{A}_{ijk}\|_2 \left\|\boldsymbol{y}_{j(k-1)} - \boldsymbol{h}_{j(k-1)}\right\|_2 \leq \sum_{j=1}^{N_k} \|\boldsymbol{A}_{ijk}\|_2 L_{j(k-1)} \|\boldsymbol{y}_{10} - \boldsymbol{h}_{10}\|_2$$

$$\text{Άρα } L_{ik} = \sum_{j=1}^{N_k} \|\boldsymbol{A}_{ijk}\|_2 L_{j(k-1)} \tag{3.23}$$

2. Pooling επίπεδο. Από τον τύπο (3.20) ισχύει ότι

$$\|\boldsymbol{y}_{ik} - \boldsymbol{h}_{ik}\|_2 \leq \left\|\boldsymbol{y}_{i(k-1)} - \boldsymbol{h}_{i(k-1)}\right\|_2 \leq L_{i(k-1)} \|\boldsymbol{y}_{10} - \boldsymbol{h}_{10}\|_2$$

$$\text{Άρα } L_{ik} = L_{i(k-1)} \tag{3.24}$$

3. ReLU επίπεδο. Από τον τύπο (3.21) ισχύει ότι

$$\|\boldsymbol{y}_{ik} - \boldsymbol{h}_{ik}\|_2 \leq \left\|\boldsymbol{y}_{i(k-1)} - \boldsymbol{h}_{i(k-1)}\right\|_2 \leq L_{i(k-1)} \|\boldsymbol{y}_{10} - \boldsymbol{h}_{10}\|_2$$

$$\text{Άρα } L_{ik} = L_{i(k-1)} \tag{3.25}$$

Επομένως αν το δίκτυο έχει p επίπεδα μπορούμε να βρούμε την σταθερά Lipschitz που ικανοποιεί την σχέση

$$\|f(\boldsymbol{y}_{in}, c) - f(\boldsymbol{h}_{in}, c)\|_2 \leq L_{cp} \|\boldsymbol{y}_{in} - \boldsymbol{h}_{in}\|_2$$

χρησιμοποιώντας τους τύπους (3.23),(3.24),(3.25) αναδρομικά για τα διαφορετικά είδη επιπέδων.

Έχοντας αναπτύξει την μέθοδο για εύρεση μιας σταθερά Lipschitz για το δίκτυο, μελετάμε πως αυτή εξελίσσεται κατά την εκπαίδευση ενός συνελικτικού νευρωνικού δικτύου. Στο συγκεκριμένο πείραμα καταγράφουμε την εξέλιξη της σταθερά κατά την εκπαίδευση ενός συνελικτικού δικτύου που εκπαιδεύεται στο MNIST [21] σύνολο δεδομένων, που παρουσιάζεται στο Κεφάλαιο 5.1. Το δίκτυο αποτελείται από 8 επίπεδα , και στο τελευταίο επίπεδο έχουμε 10 διαφορετικές εξόδους, μία για κάθε ψηφίο. Έτσι καταγράφουμε την μέση τιμή των σταθερών L_{i8} και την συμβολίζουμε ως L_{out}

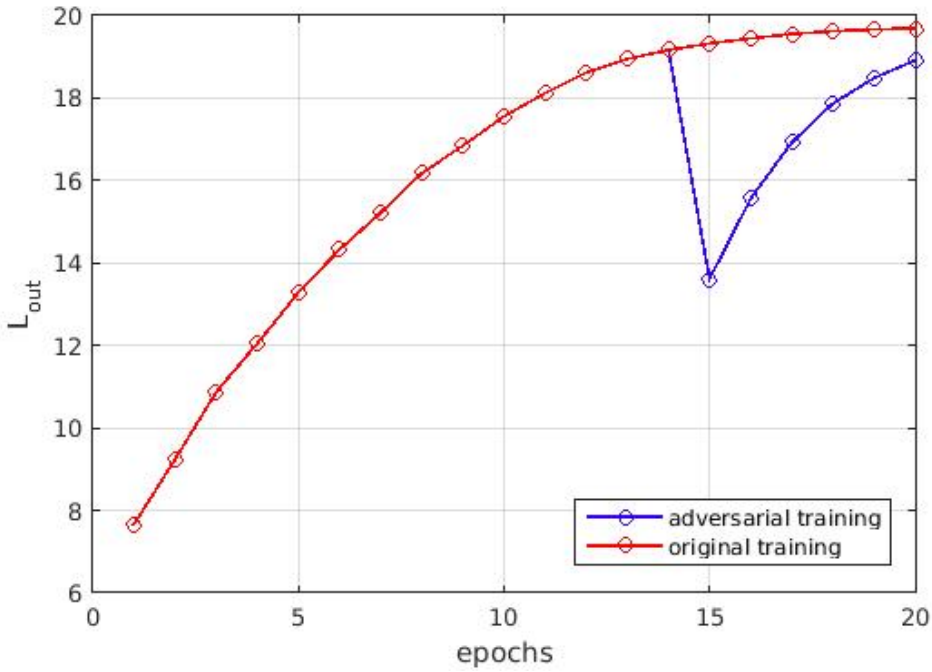
$$L_{out} = \frac{1}{10} \sum_{i=1}^{10} L_{i8} \tag{3.26}$$

Χρησιμοποιούμε 2 διαφορετικές μεθόδους εκπαίδευσης

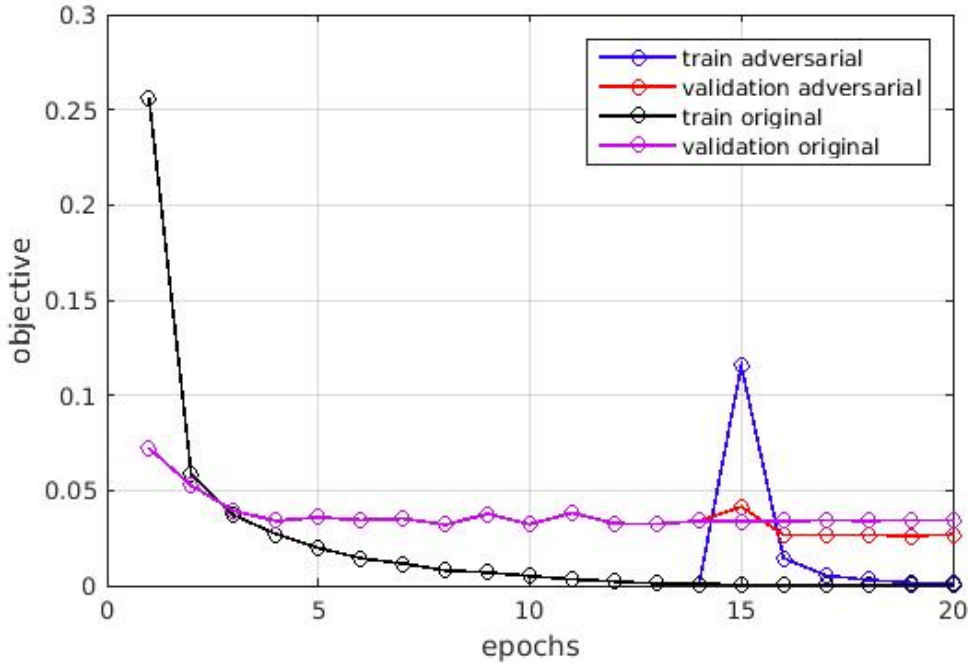
- Εκπαίδευση του δικτύου χρησιμοποιώντας ένα μείγμα πραγματικών και επιθετικών εισόδων, όπως περιγράφηκε στο Κεφάλαιο 2.4 στην μέθοδο του adversarial training
- Απλή εκπαίδευση του δικτύου

Στο τέλος κάθε εποχής εκπαίδευσης καταγράφουμε την σταθερά L_{out} για τις 2 μεθόδους. Τα αποτελέσματα παρουσιάζονται στο διάγραμμα του Σχήματος 3.1. Επίσης στο διάγραμμα του Σχήματος 3.2 παρουσιάζεται και η εξέλιξη του λάθους εκπαίδευσης για τις δύο μεθόδους τόσο στο training set όσο και στο validation set .

Είναι φανερό ότι η εκπαίδευση με επιθετικές εικόνες, η οποία στα αποτελέσματα που παρουσιάζονται ξεκίνησε στην 15η εποχή, οδηγεί σε ένα νευρωνικό δίκτυο με μικρότερη σταθερά L_{out}. Αυτό συμβαδίζει με την παραδοχή ότι δίκτυα με μικρότερη σταθερά Lipschitz θα παρουσιάζουν καλύτερη συμπεριφορά έναντι σε επιθετικές εικόνες. Βέβαια αξίζει να σημειωθεί ότι η σταθερά που υπολογίζουμε δεν αποτελεί κάτω φράγμα της σταθεράς Lipschitz του δικτύου, αλλά συνεχίζει να παρουσιάζει την συμπεριφορά που αναμένουμε. Επίσης αν παρατηρήσουμε την εξέλιξη του λάθους στο validation set, βλέπουμε ότι το δίκτυο που έχει εκπαιδευτεί με επιθετικές εικόνες και παρουσιάζει μικρότερη σταθερά L_{out} πετυχαίνει μικρότερο λάθος και γενικεύει καλύτερα.



Σχήμα 3.1: Εξέλιξη της σταθεράς L_{out} κατά την διαδικασία εκπαίδευσης με χρήση και χωρίς την χρήση επιθετικών εισόδων. Οι επιθετικές είσοδοι εισάγονται στο σύνολο εκπαίδευσης στην εποχή 15 της εκπαίδευσης



Σχήμα 3.2: Εξέλιξη του λάθους στο train,validation set κατά την εκπαίδευση με χρήση και χωρίς χρήση επιθετικών εισόδων. Οι επιθετικές είσοδοι εισάγονται στο σύνολο εκπαίδευσης στην εποχή 15 της εκπαίδευσης

Κεφάλαιο 4

Ανάλυση επιθετικών εικόνων που παράγονται από την Ταχεία Τεχνική Προσημασμένης Παραγώγου (FGSM)

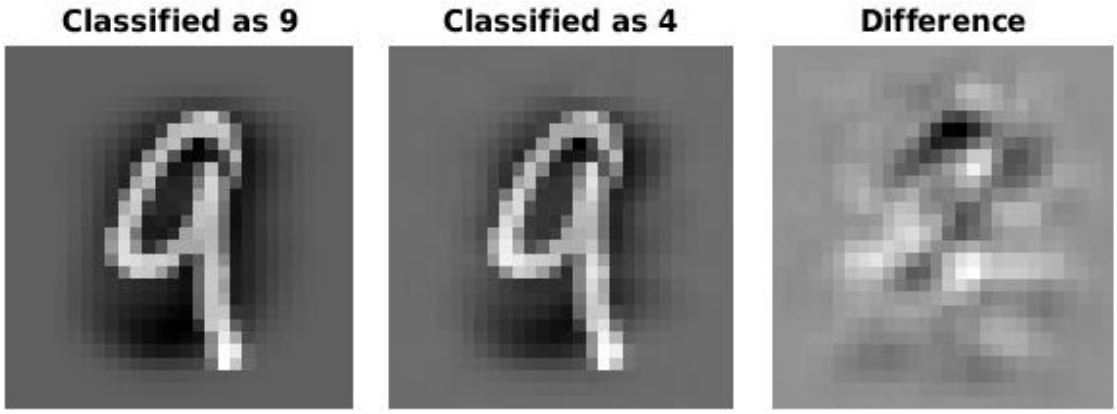
4.1 Σύγκριση των αποτελεσμάτων της FGSM , και της επίλυσης του προβλήματος ελαχιστοποίησης

Στο Κεφάλαιο 2 παρουσιάστηκε η μέθοδος Fast Gradient Sign Method (FGSM) σαν μια αποδοτικότερη μέθοδος για την εύρεση επιθετικών παραδειγμάτων από ότι η επίλυση του προβλήματος ελαχιστοποίησης που παρουσιάζεται στην (2.9), το οποίο μπορεί να επιλυθεί χρησιμοποιώντας την μέθοδο L-BFGS. Στο κεφάλαιο αυτό συγκρίνουμε τα αποτελέσματα που μας δίνουν οι δύο αυτές μέθοδοι σε ένα συνελικτικό νευρωνικό δίκτυο με 4 επίπεδα, το οποίο έχει εκπαιδευτεί στο σύνολο δεδομένων MNIST [21], που παρουσιάζεται στο Κεφάλαιο 5.1. Τα πρώτα 3 επίπεδα του δικτύου αποτελούνται από 3 ομάδες συνελικτικών, ReLU,MaxPooling επιπέδων ενώ το τελευταίο επίπεδο είναι ένα πλήρες συνδεδεμένο επίπεδο

Αρχικά επιλύουμε το πρόβλημα ελαχιστοποίησης (2.9). Οι επιθετικές εικόνες που βρίσκουμε παρουσιάζονται στα Σχήματα 4.1, 4.2, 4.3 μαζί με την αρχική εικόνα $\boldsymbol{x}_0$ και την εικόνα $\boldsymbol{r}$ που προστέθηκε. Στα σχήματα αυτά παρατηρούμε ότι οι δύο εικόνες έχουν πολύ μικρές διαφορές, με την επιθετική εικόνα να εμφανίζεται πιο θορυβώδης από την πραγματική εικόνα από την οποία προέκυψε. Επίσης όταν εκτελούμε την διαδικασία για 2000 διαφορετικές εικόνες και υπολογίζουμε τον μέσο όρο του λόγου του μέτρου της διαφοράς της επιθετικής εικόνας με την πραγματική προς το μέτρο της πραγματικής εικόνας, βρίσκουμε τον μέσο όρο ίσο με 0.33.

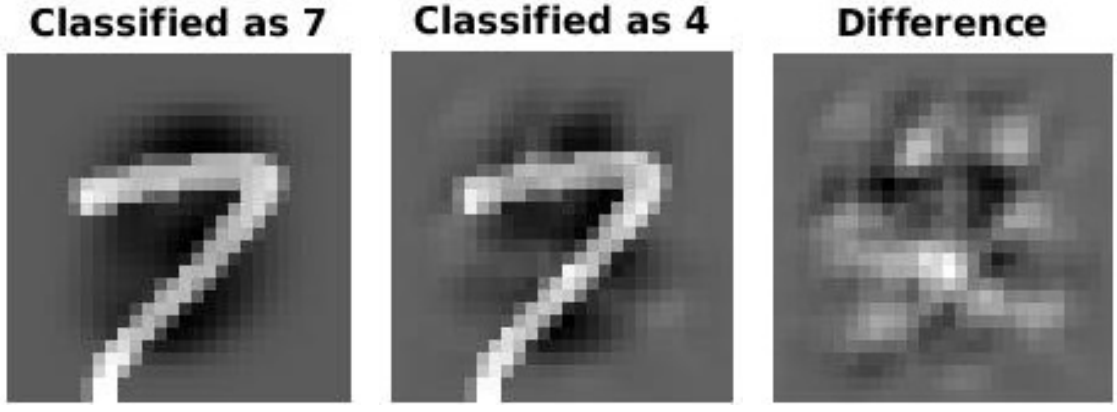
Ενδιαφέρον έχει και η μορφή της διαφορά την οποία αν προσθέσουμε στην αρχική εικόνα οδηγούμαστε στην επιθετική εικόνα. Συγκεκριμένα αν παρατηρήσουμε την εικόνα της διαφοράς βλέπουμε ότι δεν έχει την μορφή ενός γραμμικού συνδυασμού προτύπων από τις δύο κατηγορίες. Αυτό μπορεί να αποτυπωθεί και από το ίδιο το δίκτυο στο οποίο αν δοθεί σαν είσοδος μεμονωμένη η εικόνα της διαφορά, αυτή δεν ταξινομείτε ούτε στην πραγματική ούτε στην επιθετική κατηγορία. Αντίστοιχη παρατήρηση μπορούμε να κάνουμε και στο Σχήμα 2.5 όπου βλέπουμε την εφαρμογή της FGSM μεθόδου σε έγχρωμη εικόνα.

Οι επιθετικές αυτές εικόνες δεν εμφανίζονται σε μεμονωμένα σημεία αλλά βρίσκονται προς μια συγκεκριμένη διεύθυνση ως προς το αρχικό δείγμα που χρησιμοποιήθηκε για να τις παράξει. Έτσι αν βρούμε την διαφορά $\textbf{difference} = \boldsymbol{x}_{\text{adv}} - \boldsymbol{x}_0$ μεταξύ της αρχικής εικόνας $\boldsymbol{x}_0$ και της



(α) (β) (γ)

Σχήμα 4.1: (α) Αρχικό δείγμα ψηφίου /9/, (β) επιθετικό παράδειγμα που ταξινομείται ως /4/ και (γ) η διαφορά τους



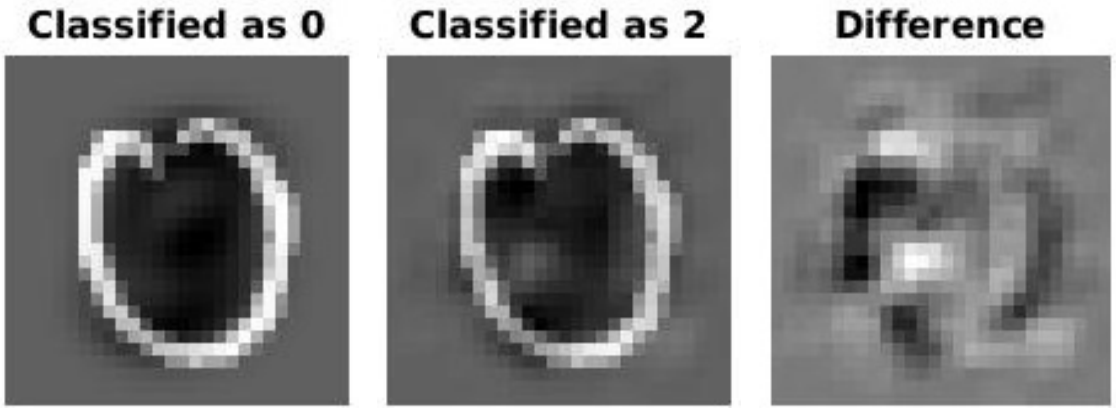
(α) (β) (γ)

Σχήμα 4.2: (α) Αρχικό δείγμα ψηφίου /7/, (β) επιθετικό παράδειγμα που ταξινομείται ως /4/ και (γ) η διαφορά τους

επιθετικής εικόνας $\boldsymbol{x}_{\text{adv}}$, τότε όλες οι εικόνες $\boldsymbol{x}_0 + \epsilon * \textbf{difference} \quad \epsilon \geq \epsilon_0$ αποτελούν επιθετικές εικόνες. Ακόμα παρατηρούμε ότι η κατεύθυνση στην οποία εμφανίζονται οι επιθετικές εικόνες είναι σχεδόν ίδια με την κατεύθυνση τού προσήμου του gradient του λάθους του δικτύου. Δηλαδή αν $J(f(\boldsymbol{x}), \ell)$ είναι το λάθος του δικτύου στο σημείο $\boldsymbol{x}$ ως προς την σωστή κατηγορία ℓ , ισχύει ότι:

$$\frac{\textbf{difference}}{\|\textbf{difference}\|_2} \approx \frac{\text{sgn}(\nabla_{\boldsymbol{x}} J(f(\boldsymbol{x}_0), \ell))}{\|\text{sgn}(\nabla_{\boldsymbol{x}} J(f(\boldsymbol{x}_0), \ell))\|_2}$$

Από την παρατήρηση αυτή προκύπτει η μέθοδος FGSM , η οποία εκμεταλλεύεται το γεγονός ότι μπορούμε να προσεγγίσουμε την επιθυμητή κατεύθυνση **difference** χρησιμοποιώντας το πρόσημο του gradient . Έτσι για το παράδειγμα του Σχήματος 4.1 βλέπουμε στο διάγραμμα του Σχήματος 4.4 πως αλλάζει η βεβαιότητα του δικτύου καθώς κινούμαστε στην κατεύθυνση που προκύπτει τόσο από το πρόσημου του *gradient* όσο και από το **difference** που προέκυψε από την διαδικασία ελαχιστοποίησης. Ενώ στο Σχήμα 4.5 φαίνεται η βεβαιότητα του



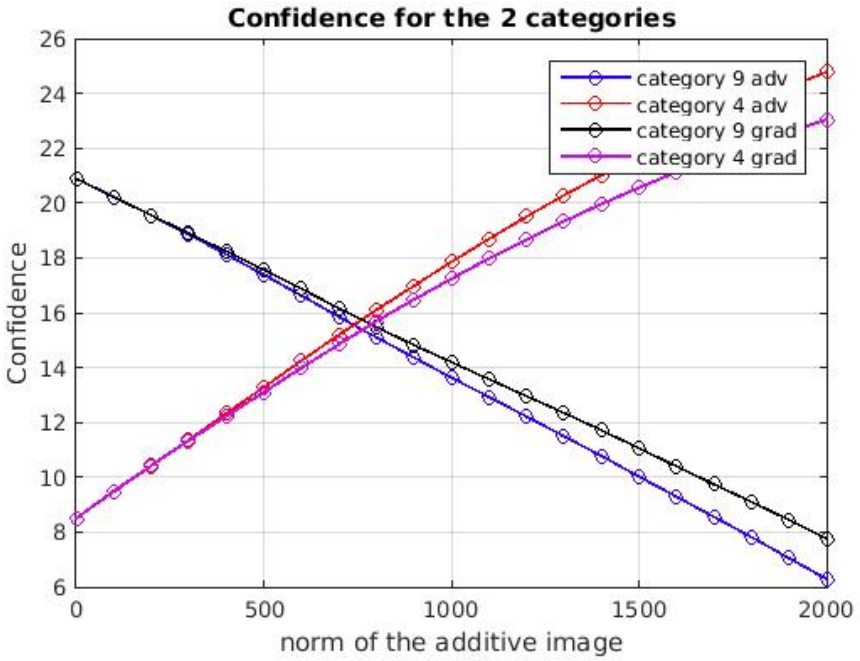
(α) (β) (γ)

Σχήμα 4.3: (α) Αρχικό δείγμα ψηφίου /0/, (β) επιθετικό παράδειγμα που ταξινομείται ως /2/ και (γ) η διαφορά τους

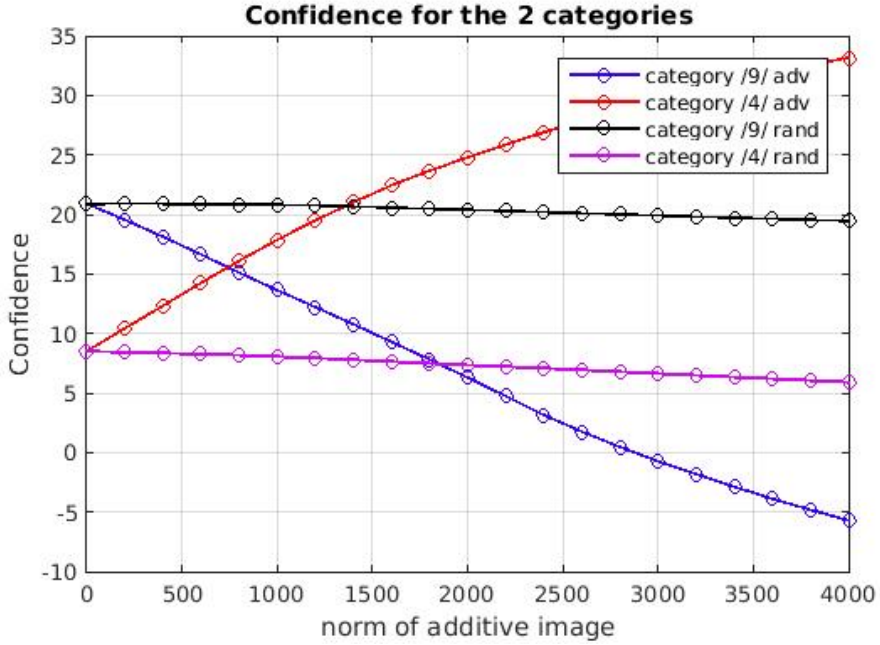
δικτύου καθώς κινούμαστε στην κατεύθυνση του **difference** και σε μια τυχαία κατεύθυνση που προέκυψε από *Gaussian* κατανομή.

Είναι φανερό ότι με την επίλυση του προβλήματος ελαχιστοποίησης, βρίσκουμε την κατεύθυνση η οποία πετυχαίνει την μεγαλύτερη μεταβολή στην βεβαιότητα του δικτύου. Παρόλο όμως που η μέθοδος FGSM δεν μας δίνει το καλύτερο αποτέλεσμα, τα αποτελέσματά της προσεγγίζουν πολύ κοντά τα αποτελέσματα της ελαχιστοποίησης, όπως αναφέρετε και στο Κεφάλαιο 2.3. Επιβεβαιώνουμε λοιπόν και πειραματικά για την συγκεκριμένη υλοποίηση συνελικτικού νευρωνικού, ότι η FGSM μας επιτρέπει να βρίσκουμε αποτελέσματα παρόμοια με την επίλυση του προβλήματος ελαχιστοποίησης με πολύ λιγότερους υπολογισμούς. Η διαφορά στους υπολογισμούς οφείλεται στο γεγονός ότι κατά την εκτέλεση της μεθόδου FGSM, εκτελείται μόνο μια φορά ο αλγόριθμος backpropagation για τον υπολογισμό της αρχικής κατεύθυνσης στην οποία θα αναζητήσουμε την επιθετική εικόνα, ενώ για την προσεγγιστική επίλυση του προβλήματος ελαχιστοποίησης (2.9) είναι αναγκαία η επαναληπτική εκτέλεση του backpropagation αλγορίθμου.

Επίσης όταν προσθέτουμε στην αρχική εικόνα τυχαίο θόρυβο δεν παρατηρούμε σχεδόν καμία μεταβολή στην βεβαιότητα του δικτύου ακόμα και όταν το μέτρο του θορύβου είναι διπλάσιο από το μέτρο της επιθετική κατεύθυνσης με την οποία πετυχαίνουμε αλλαγή της κατηγορίας αναγνώρισης. Η μόνη μεταβολή που προκαλεί ο τυχαίος θόρυβος, καθώς αυξάνεται το μέτρο του, είναι μια μικρή μείωση της βεβαιότητας του δικτύου και για τις δύο κατηγορίες. Αυτό έχει ως συνέπεια η αύξηση του συνόλου εκπαίδευσης με θορυβώδη εικόνες, ενώ βελτιώνει γενικά την γενίκευση του δικτύου, δεν βελτιώνει την συμπεριφορά του ενάντια σε επιθετικές εικόνες, κάτι που το πετυχαίνει το adversarial training όπου το σύνολο εκπαίδευσης αυξάνεται με προσθήκη επιθετικών εικόνων.



Σχήμα 4.4: Βεβαιότητα δικτύου καθώς κινούμαστε στην διεύθυνση του προσήμου του gradient($grad$) και σε αυτήν που προκύπτει από την διαδικασία ελαχιστοποίησης της βεβαιότητα για την σωστή κατηγορία(adv)



Σχήμα 4.5: Βεβαιότητα του δικτύου καθώς κινούμαστε σε μια τυχαία διεύθυνση($rand$) και σε αυτήν που προκύπτει από διαδικασία ελαχιστοποίησης της βεβαιότητας για την σωστή κατηγορία(adv)

4.2 Προσέγγιση επιθετικών κατευθύνσεων με χρήση του συνόλου εκπαίδευσης

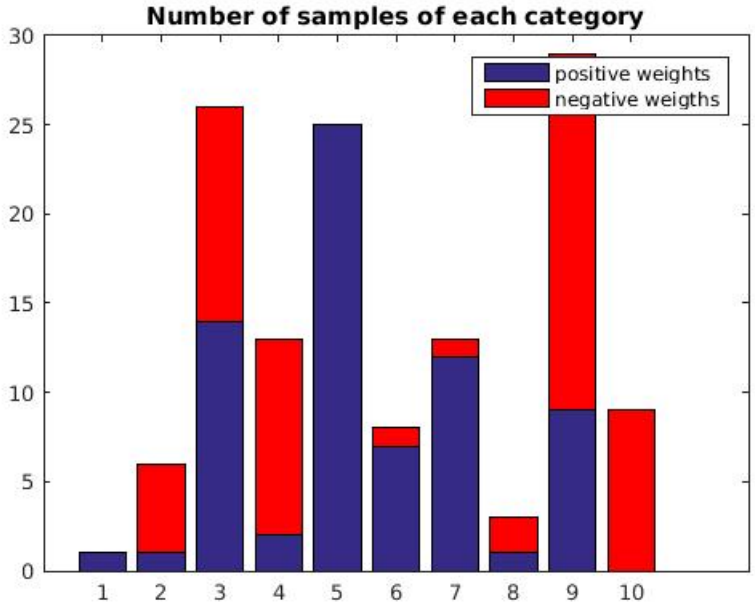
Γενικά για την εύρεση των επιθετικών παραδειγμάτων χρειάζεται να έχουμε στην διάθεση μας ένα εκπαιδευμένο δίκτυο για να εφαρμόσουμε σε αυτό κάποιες από τις μεθόδους που αναφέρθηκαν στο Κεφάλαιο 2.3. Ταυτόχρονα όμως τα επιθετικά παραδείγματα τα οποία έχουν προκύψει από ένα δίκτυο έχουν την δυνατότητα να επεκτείνονται, σε ένα ποσοστό, και σε άλλα δίκτυα, διαφορετικής αρχιτεκτονικής τα οποία λύνουν το ίδιο πρόβλημα. Αυτό μας δείχνει ότι τα επιθετικά παραδείγματα πέρα των μεμονωμένων αρχιτεκτονικών των διαφορετικών δικτύων, συνδέονται στενά και με το ίδιο το πρόβλημα και το σύνολο δεδομένων του. Έτσι επιθυμούμε να διερευνήσουμε αν μπορούμε να εντοπίσουμε μια δομή την οποία ακολουθούν τα επιθετικά παραδείγματα η οποία συνδέεται με το σύνολο εκπαίδευσης και όχι με ένα συγκεκριμένο δίκτυο το οποίο έχουμε εκπαιδεύσει.

Σε μια προσπάθεια να βρεθεί η δομή των κατευθύνσεων που δημιουργούν την λάθος ταξινόμηση σε περιοχές πολύ κοντά σε σωστά ταξινομημένα δείγματα, δοκιμάσαμε να εκφράσουμε τις κατευθύνσεις αυτές ως γραμμικό συνδυασμό των δειγμάτων εκπαίδευσης χρησιμοποιώντας την μέθοδο *Lasso*. Στο Σχήμα 4.6 βλέπουμε το ιστόγραμμα που αντιπροσωπεύει τον αριθμό των δειγμάτων εκπαίδευσης από κάθε κατηγορία που συμμετέχουν στον γραμμικό συνδυασμό για να παράγουν την ζητούμενη κατεύθυνση. Από το αποτέλεσμα αυτό δεν προκύπτει κάποια συγκεκριμένη δομή της κατεύθυνσης, παρά μόνο η παρατήρηση ότι τα δείγματα της σωστής κατηγορίας (κατηγορία /9/ με *label* 10) εμφανίζονται μόνο με αρνητικά βάρη ενώ τα δείγματα της λάθος κατηγορίας (κατηγορία /4/ με *label* 5) εμφανίζονται μόνο με θετικά βάρη.

Βασιζόμενοι στην παρατήρηση αυτή δημιουργούμε τυχαίες κατευθύνσεις (κανονικοποιημένες έτσι ώστε να έχουν μέτρο 1) όπου επιλέγουμε τυχαία δείγματα εκπαίδευσης και τους αναθέτουμε τυχαία βάρη φροντίζοντας να αναθέτουμε πάντα αρνητικά βάρη στα δείγματα της σωστής κατηγορίας και θετικά βάρη στα δείγματα της λάθος κατηγορίας. Στην συνέχεια αναφερόμαστε στις κατευθύνσεις αυτές ως τυχαίες επιθετικές κατευθύνσεις. Στο Σχήμα 4.7 φαίνεται η βεβαιότητα του δικτύου καθώς κινούμαστε στην τυχαία επιθετική κατεύθυνση. Βλέπουμε ότι ενώ δεν πετυχαίνει την μεταβολή στην βεβαιότητα που πετυχαίνει η κατεύθυνση που προέκυψε από την επίλυση του προβλήματος ελαχιστοποίησης, πετυχαίνει μια μεταβολή πολύ μεγαλύτερη από αυτήν της καθαρά τυχαίας κατεύθυνσης.

Στο Κεφάλαιο 2.4 παρουσιάζονται οι μέθοδοι Adversarial Training και Mixup Training, με τις οποίες εκπαιδεύουμε συνελικτικά νευρωνικά τα οποία εμφανίζουν καλύτερη συμπεριφορά ενάντια στις επιθετικές εικόνες. Και στις δύο αυτές μεθόδους ενισχύουμε το σύνολο εκπαίδευσης με επιπλέον εικόνες, έτσι ώστε το δίκτυο να μπορέσει να γενικεύσει καλύτερα. Καθώς με την μέθοδο που παρουσιάστηκε παραπάνω μπορούμε να δημιουργούμε εύκολα παραδείγματα που προσεγγίζουν την συμπεριφορά των επιθετικών εικόνων, επιθυμούμε να διερευνήσουμε αν ενισχύοντας το σύνολο εκπαίδευσης με τις εικόνες αυτές μπορούμε να βελτιώσουμε την συμπεριφορά του εκπαιδευμένου δικτύου ενάντια σε επιθετικές εικόνες.

Έτσι εκπαιδεύουμε ένα δίκτυο στο MNIST και μετά την 15η εποχή, εισάγουμε στο σύνολο εκπαίδευση εικόνες στις οποίες έχουν προστεθεί οι τυχαίες επιθετικές κατευθύνσεις.Στην συνέχεια χρησιμοποιώντας το δίκτυο αυτό, παράγουμε επιθετικές εικόνες και συγκρίνουμε το μέτρο του διανύσματος $\boldsymbol{r}$ το οποίο προσθέτουμε με το μέτρο των αντίστοιχων διανυσμάτων όταν παράγουμε τις αντίστοιχες επιθετικές εικόνες σε ένα δίκτυο που έχουμε εκτελέσει Adversarial Training, σε ένα δίκτυο στο οποίο έχουμε εκτελέσει Mixup Training και σε ένα δίκτυο στο οποίο έχουμε εκτελέσει απλή εκπαίδευση. Στον Πίνακα 4.1 εμφανίζονται οι μέσοι όροι των αποστάσεων για τα 4 δίκτυα για 2000 επιθετικές εικόνες.

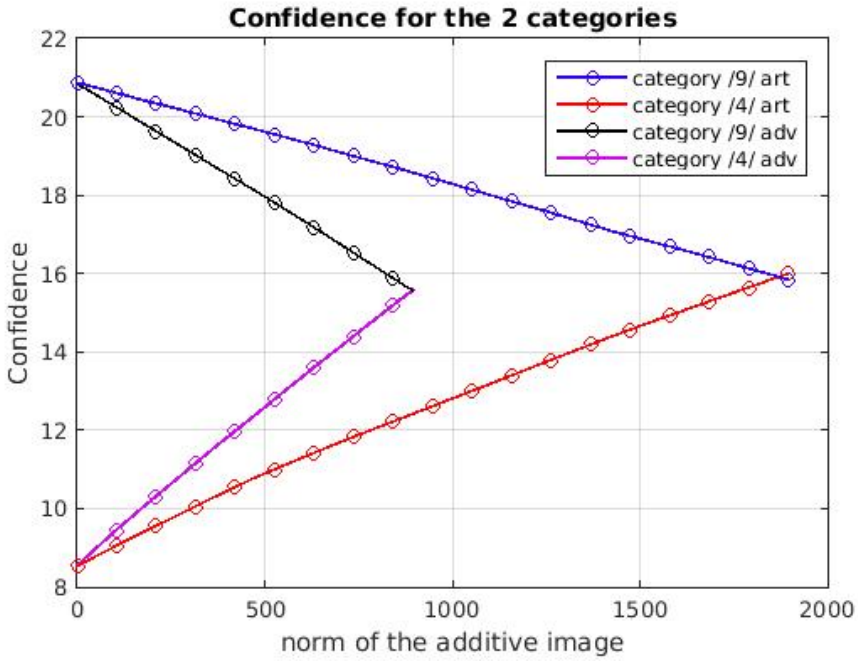


Σχήμα 4.6: Ιστόγραμμα των κατηγοριών των δειγμάτων που χρησιμοποιούνται για την προσέγγιση της επιθετικής κατεύθυνσης με την οποία το αρχικό δείγμα (κατηγορία /9/,label 10) ταξινομείται λανθασμένα ως /4/ (label 5)

	Μέσο μέτρο της διαφοράς
Απλή Εκπαίδευση	629.4
Adversarial Training	928.9
Mixup Training	815.8
Εκπαίδευση με τυχαίες επιθετικές κατευθύνσεις	789.18

Πίνακας 4.1: Μέσος όρος της L_2 νόρμας των διαφορών ανάμεσα στις επιθετικές εικόνες και τις πραγματικές εικόνες από τις οποίες παράγονται

Από τα αποτελέσματα του Πίνακα 4.1 βλέπουμε ότι και οι τρεις μέθοδοι καταφέρνουν να βελτιώσουν την συμπεριφορά του δικτύου σε σχέση με ένα δίκτυο στο οποίο έχει γίνει κανονική εκπαίδευση. Την καλύτερη συμπεριφορά έναντι σε επιθετικές εισόδους έχει το δίκτυο στο οποίο έχουμε εκτελέσει Adversarial Training. Η μέθοδος Mixup έχει την δεύτερη καλύτερη επίδοση η οποία είναι αρκετά κοντά στην επίδοση της μεθόδου που κάνει χρήση των τυχαίων επιθετικών κατευθύνσεων. Αυτές οι δύο μέθοδοι παρουσιάζουν αρκετές ομοιότητες, καθώς και στις δύο προστίθεται στην αρχική εικόνα ένα γραμμικός συνδυασμός από εικόνες του συνόλου εκπαίδευσης. Έτσι αυτές οι μέθοδοι παράγουν εικόνες, χωρίς την χρήση κάποιου εκπαιδευμένου δικτύου, οι οποίες όταν εισάγονται στο σύνολο εκπαίδευσης προσεγγίζουν την δράση που έχει η εισαγωγή επιθετικών εικόνων.



Σχήμα 4.7: Βεβαιότητα του δικτύου καθώς κινούμαστε στην τυχαία επιθετική κατεύθυνση(art), και στην διεύθυνση που προέκυψε από την διαδικασία ελαχιστοποίησης της βεβαιότητας για την σωστή κατηγορία(adv)

4.3 Γενίκευση επιθετικών εικόνων της FGSM σε διαφορετικά δίκτυα

Έχει επίσης ενδιαφέρον να μελετηθεί κατά πόσο η μέθοδος FGSM παράγει επιθετικές εικόνες οι οποίες γενικεύονται σε δίκτυα πέρα από τα αρχικό δίκτυο το οποίο τις παρήγαγε. Θέλουμε λοιπόν να καταγράψουμε το ποσοστό των επιθετικών εικόνων που παράγουμε με την FGSM για ένα νευρωνικό δίκτυο και συνεχίζουν να ταξινομούνται λανθασμένα σε ένα διαφορετικό νευρωνικό δίκτυο με διαφορετική αρχιτεκτονική και διαφορετικό σύνολο εκπαίδευσης. Τα δίκτυα που μελετάμε είναι εκπαιδευμένα στο MNIST σύνολο δεδομένων και έχουν τα παρακάτω χαρακτηριστικά

- Το πρώτο δίκτυο αποτελείται από 3 συνελικτικά επίπεδα,3 Relu επίπεδα, 3 Max Pooling επίπεδα τα οποία καταλήγουν να παράγουν ένα διάνυσμα 500 χαρακτηριστικών από το οποίο ένα πλήρως συνδεδεμένο επίπεδο δίνει 10 εξόδους, μια για κάθε ψηφίο
- Το δεύτερο δίκτυο διαφέρει από το πρώτο στο ότι έχει Average Pooling επίπεδα αντί για Max Pooling και το τελικό διάνυσμα χαρακτηριστικών αποτελείται από 400 χαρακτηριστικά
- Το τρίτο δίκτυο έχει την ίδια αρχιτεκτονική με το πρώτο αλλά μετά την 15η εποχή εκπαίδευσης έχουμε εκτελέσει Adversarial Training
- Το τέταρτο δίκτυο διαφέρει από το πρώτο στο ότι έχει Average Pooling επίπεδα και η εκπαίδευσή του έχει γίνει με την Mixup μέθοδο

Αφού λοιπόν εκπαιδεύσαμε τα παραπάνω νευρωνικά δίκτυα βρίσκουμε 2000 επιθετικές εικόνες για το πρώτο δίκτυο. Στην συνέχεια ελέγχουμε πόσες από αυτές τις εικόνες οδηγούνται σε

λάθος ταξινόμηση και από τους υπόλοιπους ταξινομητές. Έτσι παρακάτω παρουσιάζονται τα ποσοστά των επιθετικών εικόνων που μεταφέρονται στα αντίστοιχα δίκτυα

- Δεύτερο δίκτυο: 77.8%
- Τρίτο δίκτυο(Adversarial Training): 32.9%
- Τέταρτο δίκτυο (Mixup): 63.5%

Είναι φανερό ότι ανάμεσα σε δίκτυα τα οποία έχουν εκπαιδευτεί στο ίδιο σύνολο δεδομένων, το ποσοστό των επιθετικών εικόνων που μεταφέρονται είναι μεγάλο ακόμα και όταν οι αρχιτεκτονικές είναι διαφορετικές, όπως στην παραπάνω περίπτωση όπου είχαμε διαφορετικά Pooling επίπεδα. Την λιγότερη μεταφορά επιθετικών εικόνων την πετυχαίνει το δίκτυο στο οποίο έχει γίνει Adversarial Training, γεγονός που δείχνει την αποτελεσματικότητα της μεθόδου. Τέλος το Mixup δίκτυο δεν περιορίζει την μεταφορά στον ίδιο βαθμό με το τρίτο δίκτυο, καθώς παρά την πιο λεία συνάρτηση που η μέθοδος πετυχαίνει, υπάρχουν ακόμα περιοχές με επιθετικές εικόνες οι οποίες μπορούν να εντοπιστούν μέσω του πρώτου δικτύου, κάτι το οποίο εκμεταλλεύεται και η μέθοδος Black Box Attack η οποία παρουσιάστηκε στο Κεφάλαιο 2.3. Βέβαια αξίζει να σημειωθεί ότι δεδομένου του ελάχιστου επιπλέον υπολογιστικού κόστους που εισάγει η μέθοδος Mixup στην διαδικασία εκπαίδευσης η βελτίωση της μεταφοράς επιθετικών εικόνων που πετυχαίνει είναι σημαντική.

Κεφάλαιο 5

Προτεινόμενες μέθοδοι άμυνας κατά επιθετικών εικόνων

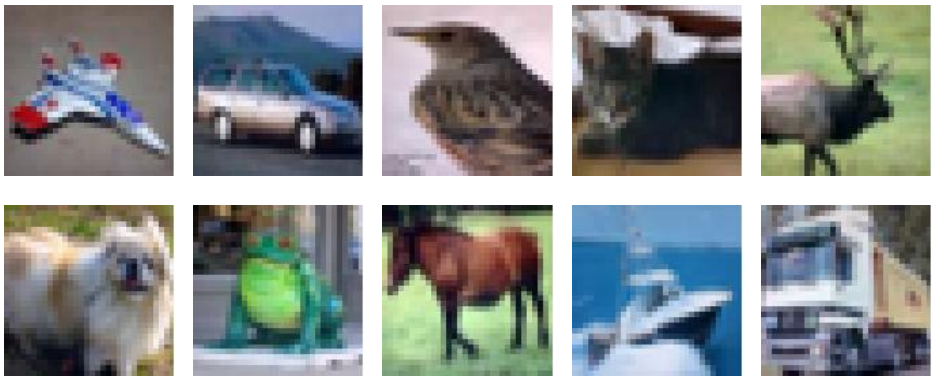
5.1 Τα σύνολα δεδομένων MNIST, CIFAR-10

Για τις τρεις μεθόδους εντοπισμού επιθετικών εικόνων που προτείνουμε, παρουσιάζουμε και τα αποτελέσματα εντοπισμού επιθετικών εικόνων για συνελικτικό νευρωνικό δίκτυο το οποίο έχει εκπαιδευτεί στο MNIST [21] σύνολο δεδομένων. Επίσης για τις μεθόδους εντοπισμού που προτείνονται στο Κεφάλαιο 5.4.2, παρουσιάζουμε τα αποτελέσματα εντοπισμού επιθετικών εικόνων για συνελικτικό νευρωνικό δίκτυο το οποίο έχει εκπαιδευτεί στο CIFAR-10 [17] σύνολο δεδομένων.

Το MNIST σύνολο δεδομένων αποτελείται από 70000, 28×28 ασπρόμαυρες εικόνες χειρόγραφων ψηφίων. Οι εικόνες του συνόλου αυτού έχουν ταξινομηθεί σε 10 διαφορετικές κατηγορίες (μια κατηγορία για κάθε ψηφίο). Στο Σχήμα 5.1 παρουσιάζονται παραδείγματα εικόνων του MNIST για τις 10 κατηγορίες ψηφίων.

Το CIFAR-10 σύνολο δεδομένων αποτελείται από 60000, 32×32 RGB εικόνες οι οποίες ανήκουν στο 10 διαφορετικές κατηγορίες. Οι εικόνες του συνόλου αυτού έχουν ταξινομηθεί στις κατηγορίες /αεροπλάνο/, /αυτοκίνητο/, /πτηνό/, /γάτα/, /ελάφι/, /σκύλος/, /βάτραχος/, /άλογο/, /πλοίο/, /φορτηγό/. Στο Σχήμα 5.2 παρουσιάζονται παραδείγματα εικόνων του CIFAR-10 για τις 10 κατηγορίες.

Σχήμα 5.1: Παραδείγματα εικόνων του MNIST για τις 10 κατηγορίες ψηφίων



Σχήμα 5.2: Παραδείγματα εικόνων του CIFAR-10 για τις 10 διαφορετικές κατηγορίες

5.2 Ομαλοποίηση του διανύσματος χαρακτηριστικών

Ακολουθώντας την διαδικασία *discrete regularization*, η οποία αναφέρεται στο [8], θεωρούμε ότι ένα διάνυσμα χαρακτηριστικών το οποίο έχει παραχθεί από μια εικόνα αποτελεί κόμβο σε έναν πλήρη μη κατευθυνόμενο γράφο.

Συγκεκριμένα μια εικόνα $\boldsymbol{x}$ αντιστοιχίζεται σε ένα κόμβο u ενός πλήρη μη κατευθυνόμενου γράφου με βάρη. Επίσης στους κόμβους του γράφου ορίζουμε μια συνάρτηση $f(u)$, όπου αν ο κόμβος u αντιστοιχίζεται στην εικόνα $\boldsymbol{x}$ η οποία έχει διάνυσμα χαρακτηριστικών $\boldsymbol{y}_x$,τότε $f(u) = \boldsymbol{y}_x$. Όπου ως διάνυσμα χαρακτηριστικών $\boldsymbol{y}_x$ της εικόνας $\boldsymbol{x}$ θεωρούμε την έξοδο του προτελευταίου επιπέδου του δικτύου, όταν θέτουμε σαν είσοδο την εικόνα αυτή.

Αν $w(u, v)$ το βάρος της ακμής που ενώνει τους κόμβους u,v τότε το μέτρο του gradient της f στον κόμβο u ορίζεται ως

$$|\nabla_w f_i(u)| = \sqrt{\sum_{v \neq u} w(u, v)(f_i(u) - f_i(v))^2} \tag{5.1}$$

Στην περίπτωσή μας συμβολίζουμε ως $f_i(u)$ το i-οστό στοιχείο του διανύσματος $f(u)$, όπου θεωρούμε ότι το $f(u)$ έχει m στοιχεία άρα $i = 1, 2, ..., m$. Χρησιμοποιώντας τον παραπάνω ορισμό στο [8] προτείνεται η παρακάτω επαναληπτική μέθοδος για την ομαλοποίηση της συνάρτηση f , όπου συμβολίζουμε ως $f^{(t)}$ την ομαλοποιημένη συνάρτηση f μετά την t επανάληψη.

$$f_i^{(0)} = f_i \tag{5.2}$$

$$f_i^{(t+1)}(v) = \frac{\lambda f_i^{(0)}(v) + \sum_{u \neq v} \gamma_i^{(t)}(u, v) f_i^{(t)}(u)}{\lambda + \sum_{u \neq v} \gamma_i^{(t)}(u, v)} \tag{5.3}$$

όπου το $\gamma_i(u, v)$ ονομάζεται p-Laplace τελεστής και ορίζεται ως εξής

$$\gamma_i^{(t)}(u, v) = w(u, v)(|\nabla_w f_i^{(t)}(u)|^{p-2} + |\nabla_w f_i^{(t)}(v)|^{p-2}) \tag{5.4}$$

ενώ τα p,λ είναι παράμετροι της μεθόδου που ρυθμίζουν τον βαθμό της ομαλοποίησης και την πιστότητα στην αρχική συνάρτηση αντίστοιχα.

Αξίζει επίσης να σημειωθεί ότι για $p \neq 2$ η τιμή του $\gamma_i(u, v)$ διαφέρει για κάθε στοιχείο $f_i(u)$ και δεν εξαρτάται από τα υπόλοιπα στοιχεία του διανύσματος. Με αυτόν τον τρόπο δεν λαμβάνεται υπόψιν η συσχέτιση μεταξύ των στοιχείων του διανύσματος κατά την ομαλοποίηση.

Έτσι προτείνεται, αντί για κάθε στοιχείο i του διανύσματος $f_i(u)$ να χρησιμοποιείται το μετρό του gradient της εξίσωσης (5.1), να χρησιμοποιείται ένα κοινό μέτρο που ορίζεται ως εξής

$$|\nabla_w f^{(t)}(u)| = \sqrt{\sum_{i=1}^{m} |\nabla_w f_i^{(t)}(u)|^2} \tag{5.5}$$

Η παραλλαγή αυτή οδηγεί σε καλύτερα αποτελέσματα στην παρούσα εφαρμογή που θα αναπτύξουμε στην συνέχεια.

Έτσι λαμβάνουμε διανύσματα χαρακτηριστικών από εικόνες του συνόλου εκπαίδευσης καθώς και από επιθετικές εικόνες, και δημιουργούμε έναν πλήρη γράφο με κόμβους τα διανύσματα αυτά και εφαρμόζουμε *regularization* για $p = 1$. Στην συνέχεια εκπαιδεύουμε το τελευταίο επίπεδο του δικτύου, το οποίο δέχεται σαν είσοδο τα διανύσματα χαρακτηριστικών, χρησιμοποιώντας τα νέα ομαλοποιημένα διανύσματα τόσο των εικόνων εκπαίδευσης όσο και ενός υποσυνόλου των επιθετικών εικόνων. Μετά την εκπαίδευση ελέγχουμε την ικανότητα του νέου δικτύου να ταξινομεί σωστά τις επιθετικές εικόνες, που δεν χρησιμοποιήθηκαν κατά την εκπαίδευση, χρησιμοποιώντας τα ομαλοποιημένα διανύσματα χαρακτηριστικών τους. Συγκεκριμένα ο τρόπος με τον οποίο η διαδικασία αυτή εντοπίζει επιθετικές εικόνες είναι ελέγχοντας την κατηγορία στην οποία ταξινομείται η είσοδος στο αρχικό δίκτυο και στο δίκτυο μετά την διαδικασία ομαλοποίησης, και αποφασίζοντας ότι η εικόνα είναι επιθετική στην περίπτωση που οι κατηγορίες αυτές διαφέρουν. Τα πειράματα πραγματοποιήθηκαν σε νευρωνικό δίκτυο που έχει εκπαιδευτεί στο MNIST σύνολο δεδομένων και οι επιθετικές εικόνες παράγονται με την μέθοδο FGSM που παρουσιάστηκε στο Κεφάλαιο 2.3

Μια σημαντική λεπτομέρεια της μεθόδου αποτελούν τα βάρη τα οποία έχουν οι ακμές του γράφου. Αρχικά αν $f(u_1)$ είναι το διάνυσμα χαρακτηριστικών της εικόνας $\boldsymbol{x}_1$ και $f(u_2)$ είναι το διάνυσμα χαρακτηριστικών της εικόνας $\boldsymbol{x}_2$. Τότε το βάρος της ακμής που ενώνει του κόμβους u_1,u_2 το θέτουμε

$$w(u_1, u_2) = e^{(-\frac{\|f(u_1)-f(u_2)\|_2^2}{\sigma^2})} \tag{5.6}$$

όπου η παράμετρος σ αποτελεί σταθερά για την οποία στο τέλος της ενότητας αναφέρουμε τις τιμές που χρησιμοποιήθηκαν στα πειράματα.

Ελέγχοντας 1000 επιθετικές και 1000 πραγματικές εικόνες η παραπάνω μέθοδος αναγνωρίζει τις επιθετικές εικόνες με Precision 89% και Recall 61.8% . Ως μέτρο σύγκρισης εφαρμόζω εκπαίδευση του τελευταίου επιπέδου του δικτύου χρησιμοποιώντας τα αρχικά διανύσματα, διαδικασία που αποτελεί ουσιαστικά adversarial training για το τελευταίο επίπεδου του δικτύου, και λαμβάνω ποσοστά Precision 86% και Recall 51%.

Αλλάζοντας το βάρος των ακμών έτσι ώστε να εξαρτάται από την ευκλείδεια απόσταση των εικόνων εισόδου και όχι των διανυσμάτων χαρακτηριστικών, δηλαδή θέτοντας τα βάρη

$$w(u_1, u_2) = e^{(-\frac{\|\boldsymbol{x}_1-\boldsymbol{x}_2\|_2^2}{\sigma^2})} \tag{5.7}$$

Λαμβάνουμε ποσοστό Precision 89.9% και Recall 63.5%.

Η χρήση της απόστασης των εικόνων εισόδου για τον καθορισμό των βαρών των ακμών του γράφου προκύπτει από το ότι επιθυμούμε να εκφράσουμε την ομοιότητα των δύο εικόνων η οποία μπορεί να χαθεί αν χρησιμοποιήσουμε τα διανύσματα χαρακτηριστικών, ειδικά στην περίπτωση που μια από τις εικόνες αυτές αποτελεί επιθετική είσοδο. Έτσι επιθυμούμε μετά την ομαλοποίηση όμοιες εικόνες να έχουν όμοια διανύσματα χαρακτηριστικών. Όμως η ευκλείδεια

απόσταση δουλεύει καλά για τον καθορισμό ομοιότητας μόνο όταν τα πρότυπα που θέλουμε να αναγνωρίσουμε έχουν κοινή κλίμακα και προσανατολισμό, με αποτέλεσμα να παρατηρούμε από το προηγούμενο πείραμα ότι αποτυγχάνει σε περιπτώσεις όπου οι εικόνες των ψηφίων έχουν υποστεί μικρή περιστροφή ή δεν είναι κεντραρισμένες.

Στις εξισώσεις (5.6),(5.7) ως απόσταση μεταξύ των διανυσμάτων χρησιμοποιώ την ευκλείδεια απόσταση . Καθώς όμως η απόσταση καθορίζει τα βάρη μεταξύ των εικόνων, είναι σημαντικό να εξετάσουμε αν μια διαφορετική απόσταση μεταξύ των διανυσμάτων χαρακτηριστικών ή των εικόνων εισόδου μας δίνει καλύτερα αποτελέσματα.
Έτσι δοκιμάζω τις παρακάτω συναρτήσεις απόστασης d

- απόσταση συνημιτόνου: που ορίζεται ως $d(\boldsymbol{x_1}, \boldsymbol{x_2}) = 1 - \frac{\langle \boldsymbol{x_1}, \boldsymbol{x_2} \rangle}{\|\boldsymbol{x_1}\|_2 \|\boldsymbol{x_2}\|_2}$.
- L_1 νόρμα της διαφοράς των διανυσμάτων άρα $d(\boldsymbol{x_1}, \boldsymbol{x_2}) = \|\boldsymbol{x_1} - \boldsymbol{x_2}\|_1$.

Στην περίπτωση όπου χρησιμοποιούμε την απόσταση μεταξύ του διανύσματος χαρακτηριστικών επεκτείνουμε την εξίσωση (5.6) και χρησιμοποιούμε την συνάρτηση d

$$w(u_1, u_2) = e^{(-\frac{d(f(u_1), f(u_2))^2}{\sigma^2})} \tag{5.8}$$

Παρόμοια, στην περίπτωση που χρησιμοποιούμε την απόσταση μεταξύ των εικόνων εισόδου επεκτείνουμε την εξίσωση (5.7) οπότε έχουμε

$$w(u_1, u_2) = e^{(-\frac{d(\boldsymbol{x}_1, \boldsymbol{x}_2)^2}{\sigma^2})} \tag{5.9}$$

Στον Πίνακα 5.1 παρουσιάζονται τα ποσοστά αναγνώρισης των επιθετικών εικόνων ανάμεσα σε 1000 πραγματικες και 1000 επιθετικές εικόνες για τις διαφορετικές αποστάσεις

	Ποσοστό Αναγνώρισης			
	Χρήση της εξίσωσης (5.8)		Χρήση της εξίσωσης (5.9)	
Απόσταση	Precision	Recall	Precision	Recall
L_2	89%	61.8%	89.9%	63.5%
Cosine	89.5%	63.1%	91%	65.2%
L_1	89.2%	62.2%	90.1%	62%
Ποσοστό Αναγνώρισης χωρίς Ομαλοποίηση				
	Precision=86%		Recall=51%	

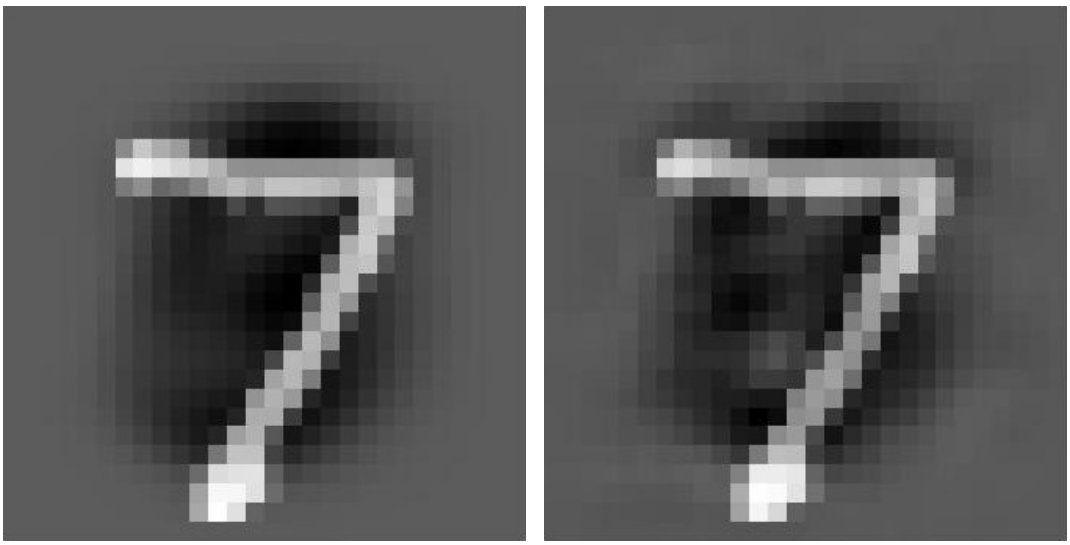
Πίνακας 5.1: Αποτελέσματα αναγνώρισης επιθετικών εικόνων μετά την ομαλοποίηση του διανύσματος χαρακτηριστικών όταν χρησιμοποιούμε διαφορετικές συναρτήσεις απόστασης

Τα διαφορετικά αποτελέσματα που παρουσιάζονται προκύπτουν όταν εκτελούμε την ομαλοποίηση που παρουσιάστηκε στην αρχή του κεφαλαίου με $p = 1$, $\lambda = 0.08$. Επίσης σημαντικό ρόλο παίζει και η παράμετρος σ που χρησιμοποιείται στις εξισώσεις (5.8),(5.9). Οπότε οι τιμές του σ που χρησιμοποιήθηκαν είναι

- Όταν χρησιμοποιείται η L_2 νόρμα έχουμε $\sigma = 120$ στην εξίσωση (5.8) και $\sigma = 1400$ στην εξίσωση (5.9)
- Όταν χρησιμοποιείται η απόσταση συνημιτόνου έχουμε $\sigma = 0.33$ στην εξίσωση (5.8) και $\sigma = 0.36$ στην εξίσωση (5.9)
- Όταν χρησιμοποιείται η L_1 νόρμα έχουμε $\sigma = 1500$ στην εξίσωση (5.8) και $\sigma = 19000$ στην εξίσωση (5.9)

5.3 Χρήση ιστογραμμάτων ενεργοποιήσεων του δικτύου για ανίχνευση επιθετικών εισόδων

Όταν δημιουργούμε επιθετικά παραδείγματα με χρήση της μεθόδου FSGM, παρατηρούμε ότι ο 'θόρυβος' που εισάγεται ενισχύει ορισμένες κορυφές των αποτελεσμάτων των συνελίξεων στα διάφορα επίπεδα του δικτύου, ενώ παράλληλα μειώνει την ενέργεια στις υπόλοιπες περιοχές. Έτσι παρατηρούμε ότι στις επιθετικές εισόδους οι έξοδοι των συνελικτικών επιπέδων έχουν μεγαλύτερο ποσοστό ενέργειας συσσωρευμένο στις κορυφές των εξόδων από ότι όταν εισάγουμε τις πραγματικές εισόδους από τις οποίες παράξαμε τα επιθετικά παραδείγματα.

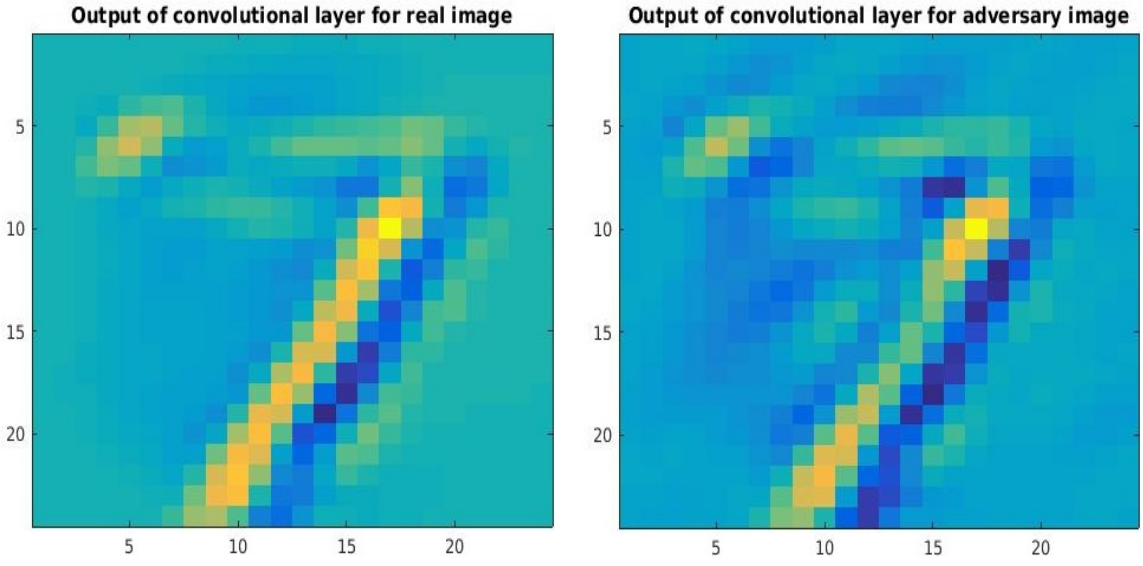


(α) (β)

Σχήμα 5.3: Παραδείγματα εικόνων εισόδου: (α) Πραγματική είσοδος που αναγνωρίζεται ως το ψηφίο /7/ (β) Επιθετική είσοδος που αναγνωρίζεται ως το ψηφίο /3/

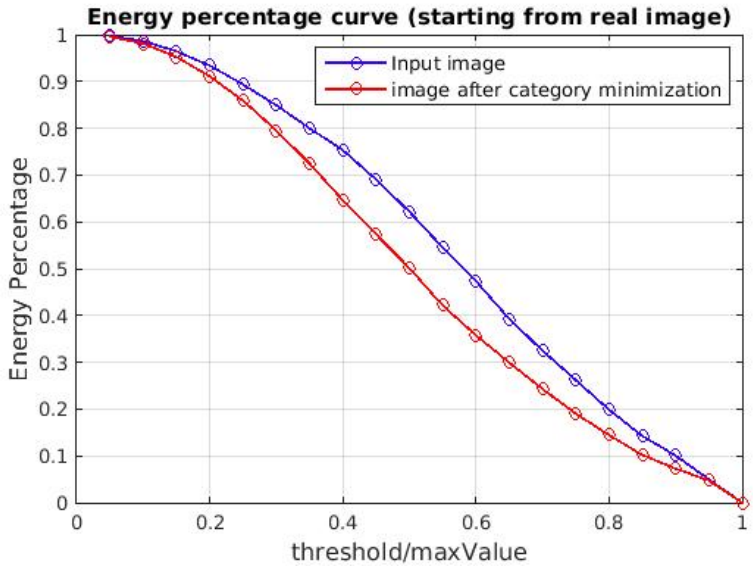
Μια διαδικασία που μπορούμε να ακολουθήσουμε προκειμένου να αξιολογήσουμε το ποσό της ενέργειας που είναι συσσωρευμένη στις κορυφές είναι να λάβουμε την έξοδο ενός συνελικτικού επιπέδου και να υπολογίσουμε την ενέργεια μόνο για τις τιμές που ξεπερνούν κατά απόλυτη τιμή ένα συγκεκριμένο κατώφλι. Μεταβάλλοντας το κατώφλι συναρτήσει της μέγιστης απόλυτης τιμής των εξόδων λαμβάνουμε μια καμπύλη. Στο διάγραμμα του Σχήματος 5.5 βλέπουμε την σύγκριση της καμπύλης για την πραγματική είσοδο και την είσοδο που έχει κατασκευαστεί από την FGSM μέθοδο. Βλέπουμε ότι η καμπύλη της πραγματικής εισόδου βρίσκεται πάνω από την καμπύλη της επιθετικής εισόδου.

Επίσης αν πάρουμε μια επιθετική είσοδο και ακολουθήσουμε την ίδια διαδικασία ,δηλαδή θεωρήσουμε ότι είναι πραγματική είσοδος και προσπαθήσουμε να κατασκευάσουμε μια επιθετική είσοδο, τότε η είσοδος που κατασκευάζουμε ταξινομείται στην σωστή κατηγορία. Καταγράφοντας την ενέργεια των εξόδων για διαφορετικά κατώφλια λαμβάνουμε τις καμπύλες του Σχήματος 5.6 όπου βλέπουμε ότι για την είσοδο που αναγνωρίζεται στην σωστή κατηγορία , σε αυτήν την περίπτωση την είσοδο που προκύπτει από την ελαχιστοποίηση της κατηγορίας της αρχικής εισόδου (κόκκινη καμπύλη), η καμπύλη είναι πάνω από την καμπύλη της επιθετικής εισόδου (μπλε καμπύλη).



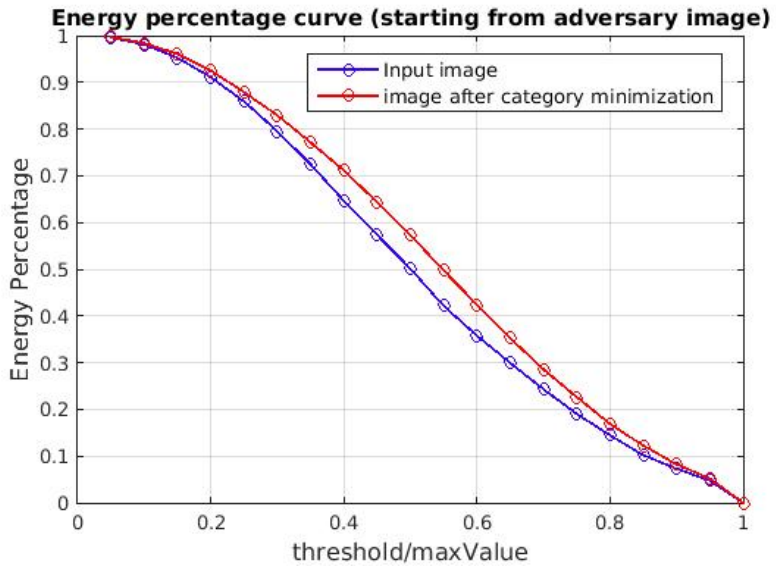
(α) (β)

Σχήμα 5.4: Παραδείγματα εξόδων του πρώτου συνελικτικού επιπέδου: (α) Όταν εισάγουμε την πραγματική είσοδο του Σχήματος 5.3α που αναγνωρίζεται ως /7/ (β) Όταν εισάγουμε την επιθετική είσοδο του Σχήματος 5.3β που αναγνωρίζεται ως /3/



Σχήμα 5.5: Ποσοστό της ενέργειας των τιμών των εξόδων του πρώτου συνελικτικού επιπέδου οι οποίες βρίσκονται πάνω από ορισμένο κατώφλι για την πραγματική είσοδο του Σχήματος 5.3α και την επιθετική είσοδο του Σχήματος 5.3β

Επομένως μεταξύ των επιθετικών και των πραγματικών εισόδων έχουμε διαφορετική κατανομή των τιμών των εξόδων των φίλτρων. Η διαφορά αυτή μπορεί να γίνει εμφανής από τα ιστογράμματα κατανομής των τιμών των εξόδων των φίλτρων. Στα Σχήματα 5.7, 5.8 βλέπουμε τα ιστόγραμμα των απόλυτων τιμών των εξόδων του πρώτου συνελικτού επιπέδου όταν έχουμε βάλει πραγματική είσοδο και επιθετική είσοδο. Στα ιστογράμματα αυτά βλέπουμε ότι στην περίπτωση των επιθετικών παραδειγμάτων τα σημεία της εικόνας όπου έχουμε ακραίες τιμές είναι λιγότερα από ότι τα αντίστοιχα σημεία των αρχικής πραγματικής εικόνας.



Σχήμα 5.6: Ποσοστό της ενέργειας των τιμών των εξόδων του πρώτου συνελικτικού επιπέδου οι οποίες βρίσκονται πάνω από ορισμένο κατώφλι για την επιθετική είσοδο του Σχήματος 5.3β και είσοδο που προκύπτει ελαχιστοποιώντας την επιθετική κατηγορία

Η διαφορά αυτή μεταξύ των ιστογραμμάτων πραγματικών και επιθετικών εισόδων μπορεί να χρησιμοποιηθεί ως χαρακτηριστικό για τον διαχωρισμό τους. Εκπαιδεύουμε λοιπόν έναν SVM ταξινομητή ο οποίος δέχεται ως εισόδους το ιστόγραμμα των εξόδων του πρώτου συνελικτικού επιπέδου και εξάγει μια εκτίμηση σχετικά με τα αν έχουμε πραγματικό ή επιθετικό παράδειγμα. Το ιστόγραμμα που χρησιμοποιήθηκε αποτελείται από 200 διαφορετικές ομάδες τιμών. Οπότε δοκιμάζοντας να αναγνωρίσουμε επιθετικές εικόνες ανάμεσα σε 2000 πραγματικές και 2000 επιθετικές εικόνες πετυχαίνουμε

$$\text{Precision} = 97.2\%$$

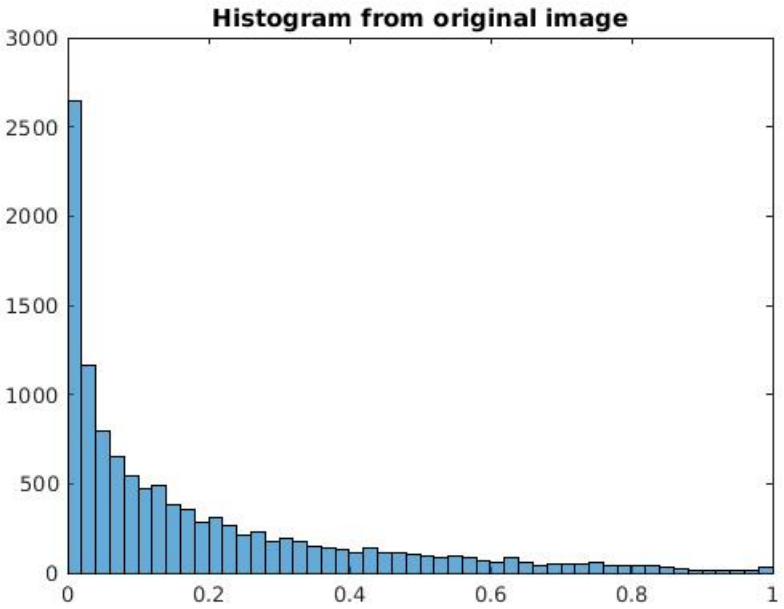
$$\text{Recall} = 97.8\%$$

Η μέθοδος αυτή είναι αρκετά ευαίσθητη αν προσθέσουμε τυχαίο θόρυβο στις εισόδους, οπότε με κατάλληλη εκπαίδευση σε θορυβώδη παραδείγματα πετυχαίνουμε

$$\text{Precision} = 69.3\%$$

$$\text{Recall} = 65.2\%$$

Η εισαγωγή Gaussian θορύβου έχει ως αποτέλεσμα να μεταβάλλονται οι τιμές των εξόδων με τυχαίο τρόπο, με αποτέλεσμα η μορφή ενός ιστογράμματος επιθετικής εισόδου να προσεγγίζει την μορφή ενός ιστογράμματος πραγματικής εισόδου. Αυτό σημαίνει ότι όταν χρησιμοποιούμε τα ιστογράμματα των εισόδων σαν διανύσματα χαρακτηριστικών για την αναγνώριση ή μη μιας εισόδου ως επιθετικής, στην περίπτωση των εικόνων με θόρυβο η διάκριση των διανυσμάτων στις δύο κατηγορίες είναι δυσκολότερη. Μια βελτίωση της προηγούμενης μεθόδου είναι δεδομένης μιας εισόδου στην οποία έχει προστεθεί θόρυβος να κινηθούμε προς την κατεύθυνση η οποία ενισχύει την κατηγορία στην οποία αναγνωρίζεται η αρχική είσοδος (reinforcement step).



Σχήμα 5.7: Ιστόγραμμα τιμών εξόδου του πρώτου συνελικτικού επιπέδου όταν εισάγουμε σαν είσοδο την πραγματική εικόνα του Σχήματος 5.3α

Δηλαδή για ένα συνελικτικό νευρωνικό δίκτυο που υλοποιεί την συνάρτηση f, αν έχουμε την είσοδο $\boldsymbol{x}$ η όποια ταξινομείται στην κατηγορία $f(\boldsymbol{x}) = \ell$, τότε λαμβάνουμε την νέα είσοδο

$$\boldsymbol{x}_{new} = \boldsymbol{x} - \epsilon * \frac{\nabla_x J(f(\boldsymbol{x}), \ell)}{\|\nabla_x J(f(\boldsymbol{x}), \ell)\|_2} \|\boldsymbol{x}\|_2 \tag{5.10}$$

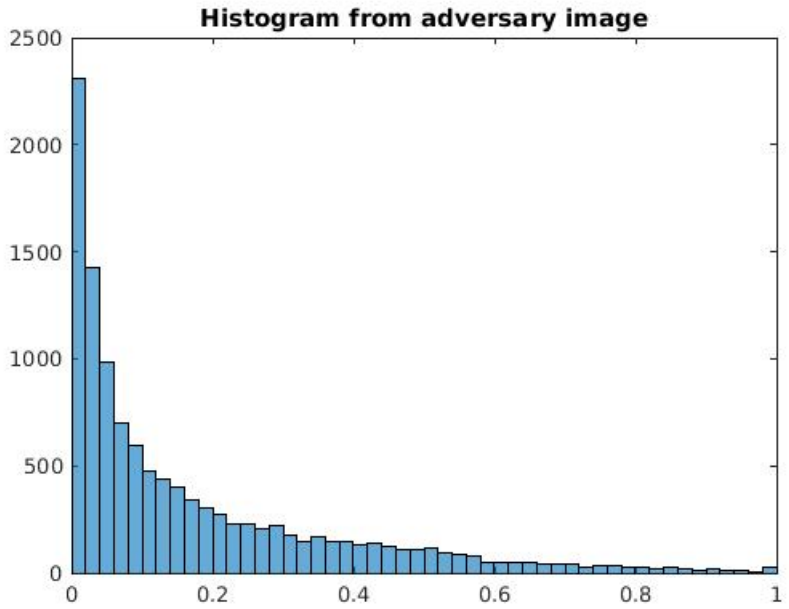
όπου $J(f(\boldsymbol{x}), \ell)$ είναι το λάθος της εξόδου του ταξινομητή για είσοδο $\boldsymbol{x}$, δεδομένης επιθυμητής εξόδου ℓ

Στην περίπτωση όπου η αρχική είσοδος $\boldsymbol{x}$ είναι επιθετική, το $\boldsymbol{x}_{new}$ θα ενισχύει την επιθετική κατηγορία, οπότε το ιστόγραμμα του $\boldsymbol{x}_{new}$ θα έχει πιο έντονα χαρακτηριστικά ιστογράμματος επιθετικής εισόδου από ότι το ιστόγραμμα του $\boldsymbol{x}$. Ενώ στην περίπτωση όπου το $\boldsymbol{x}$ είναι πραγματική είσοδος θα συμβαίνει το αντίθετο. Έτσι όταν λαμβάνουμε τα ιστογράμματα του $\boldsymbol{x}$, $\boldsymbol{x}_{new}$ σαν ένα ενιαίο διάνυσμα χαρακτηριστικών ο συνδυασμός των δύο ιστογραμμάτων μας δίνει περισσότερη πληροφορία για το είδος της εισόδου.

Μια ακόμη λεπτομέρεια η οποία οδηγεί σε καλύτερα αποτελέσματα, είναι για την έξοδο ενός επιπέδου το οποίο έχει n κανάλια να λαμβάνουμε ένα ιστόγραμμα για κάθε κανάλι ξεχωριστά. Οπότε το διάνυσμα χαρακτηριστικών είναι η συνένωση των ιστογραμμάτων των n καναλιών, σε αντίθεση με την προηγούμενη μέθοδο όπου λαμβάναμε ένα ιστόγραμμα που περιλάμβανε όλες τις τιμές των εξόδων του επιπέδου από όλα τα κανάλια.

Για την τροποποιημένη μέθοδο εκτελώ πείραμα στο σύνολο δεδομένων MNIST όπου προσπαθούμε να αναγνωρίσουμε επιθετικές εικόνες ανάμεσα σε 2000 πραγματικές και 2000 επιθετικές εικόνες. Στην περίπτωση αυτή χρησιμοποιήθηκαν οι έξοδοι του πρώτου συνελικτικού επιπέδου και το διάνυσμα χαρακτηριστικών αποτελείται από την συνένωση 20 ιστογραμμάτων, κάθε ένα από τα οποία αποτελείται από 20 διαφορετικές ομάδες τιμών.

Στον Πίνακα 5.2 παρουσιάζονται τα αποτελέσματα τόσο της τροποποιημένης μεθόδου όσο και της αρχικής μεθόδου, όταν έχουμε εικόνες χωρίς θόρυβο και όταν έχουμε εικόνες όπου έχει προστεθεί Gaussian θόρυβος με μηδενική μέση τιμή και τυπική απόκλιση 15. Παρατηρούμε



Σχήμα 5.8: Ιστόγραμμα τιμών εξόδου του πρώτου συνελικτικού επιπέδου όταν εισάγουμε σαν είσοδο την επιθετική εικόνα του Σχήματος 5.3β

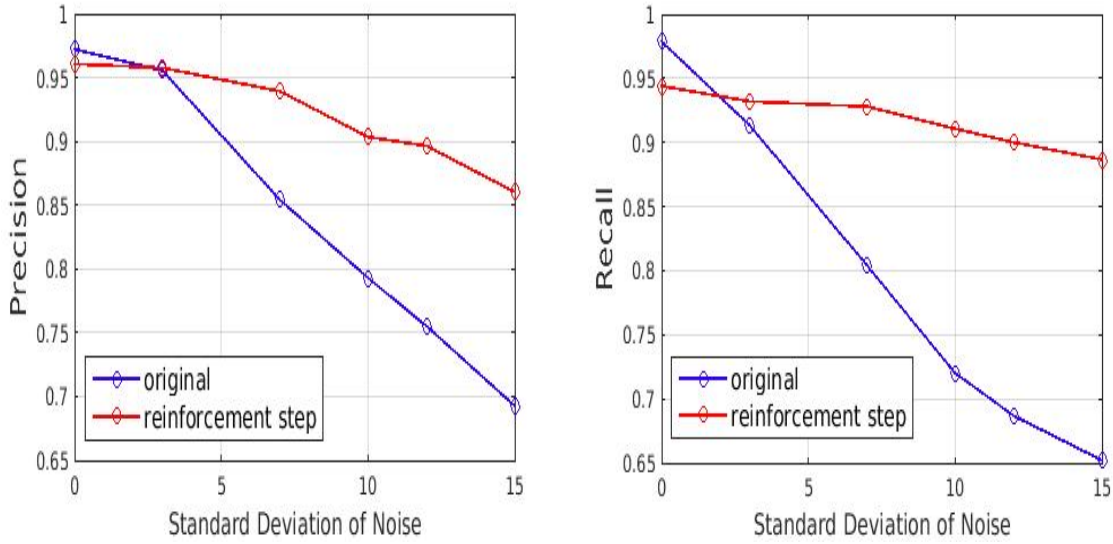
Without Noise		
	Precision	Recall
Original	97.2%	97.8%
Reinforcement Step	96%	94.4%
With Noise with standard deviation 15		
Original	69.3%	65.2%
Reinforcement Step	86%	88.6%

Πίνακας 5.2: Αποτελέσματα των δύο μεθόδων ιστογραμμάτων για διαφορετική ποσότητα θορύβου

ότι τα αποτελέσματα της τροποποιημένης μεθόδου βελτιώνονται στις θορυβώδης εικόνες, ενώ στις εικόνες στις οποίες δεν έχει προστεθεί θόρυβος πετυχαίνουμε αποτελέσματα αρκετά κοντά στην αρχική μέθοδο

Επίσης στο Σχήμα 5.9 βλέπουμε πως μεταβάλλονται τα αποτελέσματα των δύο μεθόδων συναρτήσει της τυπικής απόκλισης του θορύβου που προσθέτουμε. Όπου βλέπουμε ότι με την δεύτερη μέθοδο (reinforcement step), ενώ χωρίς θόρυβο έχουμε μικρότερα ποσοστά Precision και Recall από ότι με την αρχική μέθοδο, η προσθήκη θορύβου έχει ως αποτέλεσμα μικρότερη μεταβολή των επιδόσεων της σε σχέση με την μεταβολή των επιδόσεων της αρχικής μεθόδου. Έτσι μετά από δεδομένο κατώφλι της τυπικής απόκλισης του θορύβου, η δεύτερη μέθοδος υπερτερεί της πρώτης μεθόδου.

Τέλος θέλουμε να εξετάσουμε αν τα μοτίβα που αναγνωρίζει το SVM για τον διαχωρισμό ανάμεσα σε επιθετικές και πραγματικές εικόνες περιορίζονται από τον τρόπο παραγωγής των επιθετικών εικόνων, δηλαδή στην περίπτωσή μας δουλεύει μόνο για την FGSM μέθοδο. Έτσι αφού έχουμε εκπαιδεύσει το SVM για 2000 πραγματικές και 2000 επιθετικές εικόνες οι οποίες έχουν παραχθεί από την FGSM μέθοδο, προσπαθούμε να αναγνωρίσουμε 2000 επιθετικές



(α) (β)

Σχήμα 5.9: Ποσοστά αναγνώρισης των δύο μεθόδων ιστογραμμάτων συναρτήσει της τυπικής απόκλισης του προστιθέμενου θορύβου (α) Μεταβολή του Precision των δύο μεθόδων (β) Μεταβολή του Recall των δύο μεθόδων

εικόνες οι οποίες έχουν παραχθεί από την μέθοδο DeepFool που παρουσιάστηκε στο Κεφάλαιο 2.3. Έτσι λαμβάνουμε Precision = 88.2% και Recall = 87.1% για τις επιθετικές εικόνες χωρίς θόρυβο, και $Precision = 80.3\%$ και $Recall = 80.9\%$ για επιθετικές εικόνες όπου έχουμε προσθέσει Gaussian θόρυβο με τυπική απόκλιση 15.

Βλέπουμε ότι παρόλο που δεν πετυχαίνουμε το αποτέλεσμα που πετυχαίνουμε με τις εικόνες της FGSM μεθόδου, μπορούμε ακόμα να αναγνωρίζουμε σε μεγάλο ποσοστό τις επιθετικές εικόνες, γεγονός που δείχει ότι η μέθοδος μπορεί να επεκταθεί και σε περισσότερες από μια μεθόδους παραγωγής επιθετικών εικόνων. Επίσης δοκιμάζουμε να εκπαιδεύσουμε ένα SVM σε 2000 πραγματικές και 2000 επιθετικές εικόνες οι οποίες αποτελούνται από ένα μείγμα εικόνων από την μέθοδο FGSM και την μέθοδο DeepFool , καθώς επίσης και ένα SVM όπου οι επιθετικές εικόνες παράγονται μόνο από την DeepFool μέθοδο. Στις περιπτώσεις αυτές παρατηρούμε ότι και οι δύο διαδικασίες εκπαίδευσης οδηγούν σε χειρότερα αποτελέσματα καθώς μειώνονται και το Precision και το Recall. Καταλήγουμε λοιπόν στο ότι, δεδομένου ότι η παραγωγή της εικόνας $\boldsymbol{x}_{new}$ γίνεται με την εξίσωση (5.10), η μέθοδος πετυχαίνει τα καλύτερα αποτελέσματα όταν οι επιθετικές εικόνες του συνόλου εκπαίδευσης έχουν παραχθεί από την FGSM μέθοδο, ανεξάρτητα από το αν οι επιθετικές εικόνες τις οποίες θέλουμε να αναγνωρίσουμε παράγονται από την FGSM ή την DeepFool μέθοδο.

Αν θεωρήσουμε ότι έχουμε εκτελέσει την εκπαίδευση του SVM, το υπολογιστικό κόστος της μεθόδου για να αποφασίσουμε αν μια νέα είσοδος είναι επιθετική ή πραγματική, είναι πολύ μικρό. Καθώς αφού εισάγουμε την είσοδο και λάβουμε τις εξόδους του νευρωνικού, η εξαγωγή των ιστογραμμάτων μπορεί να γίνει σε γραμμικό χρόνο ως προς το μέγεθος του διανύσματος εξόδου του επιπέδου που χρησιμοποιούμε για να την δημιουργία των ιστογραμμάτων. Επίσης η πολυπλοκότητα εκτέλεσης του SVM είναι $O(sn)$, όπου s είναι ο αριθμός των support vectors και n είναι το μέγεθος της εισόδου, όπου στην πρώτη μέθοδο είναι το ιστόγραμμα με μέγεθος $n = 200$ ενώ στην δεύτερη μέθοδο είναι το διάνυσμα που προκύπτει από την συνένωση των διαφορετικών ιστογραμμάτων όπου έχει διάσταση $n = 800$.

5.4 Χρήση της υπολειπόμενης από τον ταξινομητή εικόνας για ανίχνευση επιθετικών εισόδων

5.4.1 Η έννοια της υπολειπόμενης εικόνας

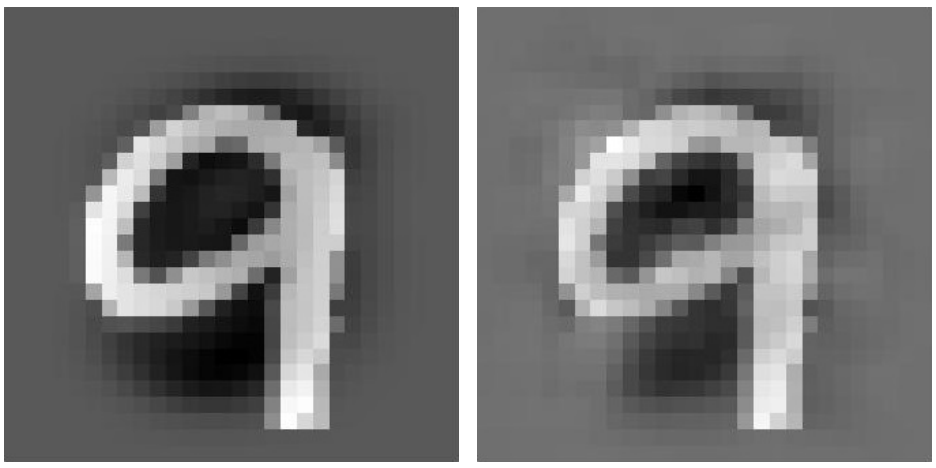
Η λειτουργία του νευρωνικού βασίζεται στην εξαγωγή ενός διανύσματος χαρακτηριστικών το οποίο στην συνέχεια μπορεί να χρησιμοποιηθεί για να ταξινομήσει την είσοδο σε μια από τις πιθανές κατηγορίες. Θέλουμε να βρούμε την εικόνα που ενισχύει το μέτρο του διανύσματος χαρακτηριστικών χωρίς να αλλάζει την κατευθυνσή του. Δηλαδή αν f είναι η συνάρτηση που παράγει το διάνυσμα χαρακτηριστικών , $\boldsymbol{x}_0$ η αρχική είσοδος και $f(\boldsymbol{x}_0) = \boldsymbol{p}$ το διάνυσμα χαρακτηριστικών, θέλουμε να βρούμε την κατεύθυνση $\boldsymbol{\delta}$ όπου

$$f(\boldsymbol{x}_0 + \boldsymbol{\delta}) = (1 + \epsilon)\boldsymbol{p} \quad \epsilon > 0$$

Αυτό το καταφέρνουμε εκτελώντας *backpropagation* και αρχικοποιώντας έτσι ώστε η μεταβολή της συνάρτησης f να ισούται με $\boldsymbol{p}$.

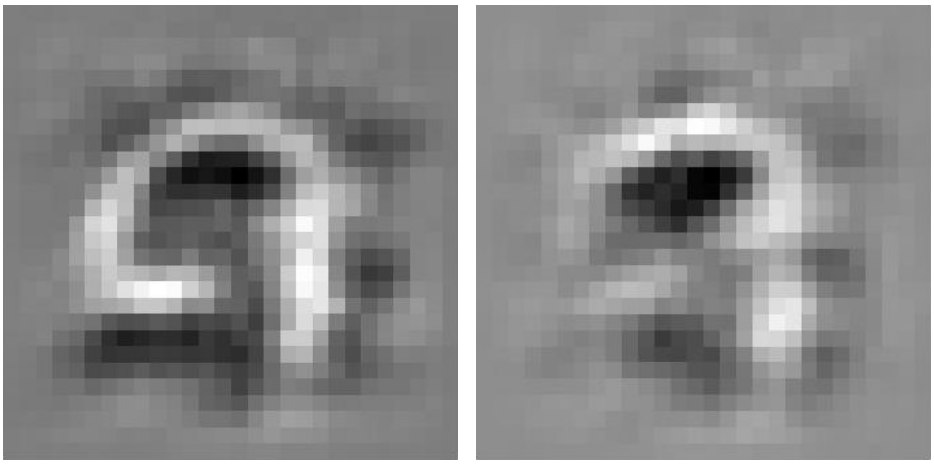
Στην ενότητα αυτή παράλληλα με την ανάπτυξη της έννοιας της υπολειπόμενης εικόνας χρησιμοποιείται ως παράδειγμα ένα συνελικτικό νευρωνικό δίκτυο το οποίο έχει εκπαιδευτεί στο MNIST σύνολο δεδομένων. Έτσι στην περίπτωση του δικτύου αυτού, το διάνυσμα χαρακτηριστικών είναι η έξοδος του προτελευταίου επιπέδου του δικτύου που αποτελείται από 500 χαρακτηριστικά.

Εκτελώντας λοιπόν backpropagation και ενισχύοντας το διάνυσμα χαρακτηριστικών λαμβάνουμε τις εικόνες 5.11α, 5.11β οι οποίες παράγονται από τις εισόδους 5.10α, 5.10β . Οπότε παρατηρούμε ότι οι κατευθύνσεις που παράγονται από την παράγωγο δείχνουν κατά κάποιον τρόπο τα κυριότερα μέρη του προτύπου εισόδου που οδηγούν το δίκτυο να καταλήξει στο διάνυσμα χαρακτηριστικών. Έτσι βλέπουμε ότι με είσοδο η οποία ταξινομείται σωστά ως /9/ η παράγωγος έχει πράγματι την μορφή του ψηφίου /9/, ενώ με την είσοδο η οποία ταξινομείται λάθος ως /7/ η παράγωγος έχει μορφή που πλησιάζει το ψηφίο /7/.



(α) (β)

Σχήμα 5.10: Παραδείγματα εικόνων εισόδου: (α) Πραγματική είσοδος που αναγνωρίζεται ως το ψηφίο /9/ (β) Επιθετική είσοδος που αναγνωρίζεται ως το ψηφίο /7/



(α) (β)

Σχήμα 5.11: Κατευθύνσεις που προκύπτουν από το backpropagation και ενισχύουν το μέτρο του διανύσματος χαρακτηριστικών: (α) Κατεύθυνση η οποία ενισχύει το διάνυσμα χαρακτηριστικών που προκύπτει με είσοδο την εικόνα 5.10α που ταξινομείται ως /9/ (β) Κατεύθυνση η οποία ενισχύει το διάνυσμα χαρακτηριστικών που προκύπτει με είσοδο την εικόνα 5.10β που ταξινομείται ως /7/

Συγκρίνοντας το *gradient* που προκύπτει από το *backpropagation* με την εικόνα εισόδου μπορούμε να εξάγουμε συμπεράσματα για το είδος της εισόδου. Η σύγκριση προκύπτει από το γεγονός ότι τα μη γραμμικά στοιχεία του δικτύου είναι σε περιοχές γραμμικά. Έτσι αφού καθοριστούν οι ενεργοποιήσεις των μη γραμμικών στοιχείων, σε μια περιοχή γύρω από την εικόνα εισόδου $\boldsymbol{x}_0$ το νευρωνικό συμπεριφέρεται σαν γραμμικός ταξινομητής. Έτσι για κάθε χαρακτηριστικό f_i μπορούμε να βρούμε ένα διάνυσμα $\boldsymbol{w}_i$ και μια τιμή b_i έτσι ώστε $f_i = \boldsymbol{w}_i^T \boldsymbol{x}_0 + b_i$. Οπότε αν έχουν καθοριστεί οι ενεργοποιήσεις των μη γραμμικών στοιχείων, το διάνυσμα χαρακτηριστικών $\boldsymbol{f}_{x_0}$ της εικόνας εισόδου $\boldsymbol{x}_0$ μπορεί να προκύψει ως εξής

$$\boldsymbol{f}_{x_0} = \begin{bmatrix} \boldsymbol{w}_1^T \\ \boldsymbol{w}_2^T \\ \vdots \\ \boldsymbol{w}_n^T \end{bmatrix} \boldsymbol{x}_0 + \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} \tag{5.11}$$

Το νευρωνικό λοιπόν αντιλαμβάνεται την εικόνα $\boldsymbol{x}_0$ ως το διάνυσμα χαρακτηριστικών $\boldsymbol{f}_{x_0}$. Έτσι την ομοιότητα μιας νέας εικόνας $\boldsymbol{x}$ με διάνυσμα χαρακτηριστικών $\boldsymbol{f}$ με την εικόνα $\boldsymbol{x}_0$ (θεωρώντας ότι $\|\boldsymbol{x}\|_2 = \|\boldsymbol{x}_0\|_2$) το νευρωνικό την αντιλαμβάνεται ως το εσωτερικό γινόμενο $\boldsymbol{f}_{x_0}^T \boldsymbol{f}$. Αν οι δύο εικόνες ενεργοποιούν τις γραμμικότητες με τον ίδιο τρόπο, η ομοιότητα είναι

$$\boldsymbol{f}_{x_0}^T \boldsymbol{f} = \boldsymbol{f}_{x_0}^T \boldsymbol{W} \boldsymbol{x} + \boldsymbol{f}_{x_0}^T \boldsymbol{b} \tag{5.12}$$

Άρα το εσωτερικό γινόμενο των χαρακτηριστικών εξαρτάται από το εσωτερικό γινόμενο της εικόνας $\boldsymbol{x}$ με την εικόνα $\boldsymbol{f}_{x_0}^T \boldsymbol{W}$. Έτσι η εικόνα $\boldsymbol{f}_{x_0}^T \boldsymbol{W}$ μπορεί να ερμηνευτεί ως η πληροφορία που εξάγεται από το νευρωνικό για την εικόνα $\boldsymbol{x}_0$ (δεδομένων των ενεργοποιήσεων των μη γραμμικών στοιχείων). Την $\boldsymbol{f}_{x_0}^T \boldsymbol{W}$ την λαμβάνουμε με την διαδικασία του *backpropagation* που περιγράφτηκε στην αρχή της ενότητας.

Αν παρατηρήσουμε τις εικόνες $\boldsymbol{f}_{x_0}^T \boldsymbol{W}$, ιδιαίτερα αυτή που προκύπτει από την επιθετική είσοδο, παρατηρούμε ότι τα σημεία τα οποία αντιλαμβάνεται ο ταξινομητής σαν το πρότυπο

εισόδου δεν ταυτίζονται απόλυτα με αυτά του πραγματικού προτύπου εισόδου. Έτσι υπάρχει ένα μέρος της εισόδου το οποίο αγνοείται από τον ταξινομητή και στην περίπτωση της επιθετικής εισόδου είναι φανερό ότι το μέρος της εισόδου που αγνοείται παίζει μεγάλο ρόλο στον διαχωρισμό ανάμεσα στην πραγματική κατηγορία /9/ και στην λάθος κατηγορία /7/. Παρακάτω παρουσιάζουμε αναλυτικότερα μια διαδικασία που μας δίνει το μέρος της εικόνας εισόδου που αγνοείται από τον ταξινομητή και το οποίο το ονομάζουμε υπολειπόμενη εικόνα

Θεωρώ ως y_i την βεβαιότητα του νευρωνικού ότι η εικόνα εισόδου ανήκει στην κατηγορία c_i, δηλαδή είναι η έξοδος του τελευταίου επιπέδου μετά την εξαγωγή του διανύσματος χαρακτηριστικών. Οπότε παρόμοια με προηγουμένως, δεδομένων των ενεργοποιήσεων των μη γραμμικών στοιχείων, μπορούμε να βρούμε πίνακα $\boldsymbol{A}$ και διάνυσμα $\boldsymbol{d}$ για τα οποία ισχύει ότι

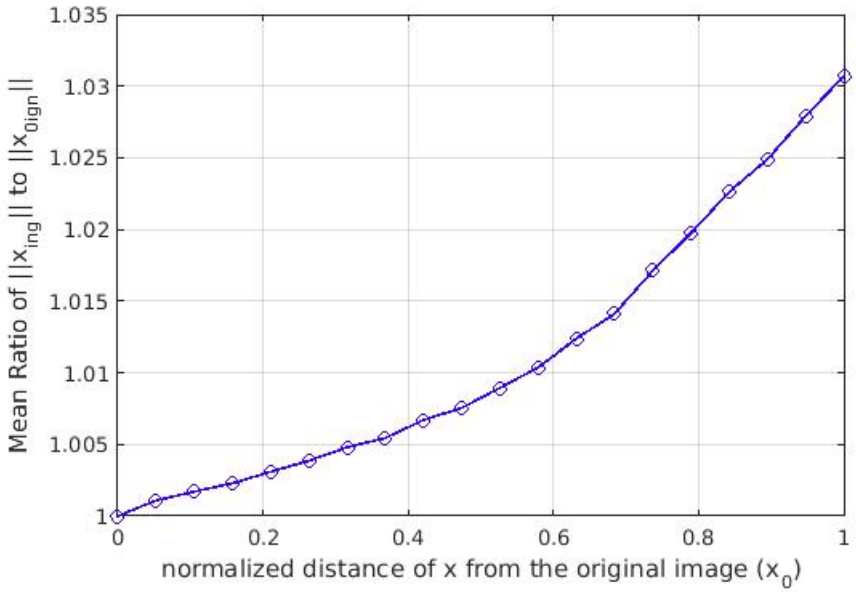
$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{d} \tag{5.13}$$

Αν προβάλουμε το $\boldsymbol{x}$ στον μηδενοχώρο του $\boldsymbol{A}$ λαμβάνουμε την εικόνα $\boldsymbol{x}_{ign}$, οπότε $\boldsymbol{x} = \boldsymbol{x}_p + \boldsymbol{x}_{ign}$. Έτσι έχουμε την εικόνα $\boldsymbol{x}_p$ την οποία αντιλαμβάνεται το νευρωνικό και στην οποία βασίζεται για να αποφασίσει την κατηγορία της εισόδου και την εικόνα $\boldsymbol{x}_{ign}$, την οποία αγνοεί. Παρατηρώντας την $\boldsymbol{x}_p$ βλέπουμε ότι στην περίπτωση επιθετικής εισόδου προσεγγίζει την λάθος κατηγορία αναγνώρισης και επομένως η $\boldsymbol{x}_{ign}$ μας δίνει πληροφορία για τις 'τρύπες' και τα 'γεμίσματα' που εισάγονται στην αντίληψη του νευρωνικού και το οδηγούν στην λάθος ταξινόμηση.

Ιδανικά ο ταξινομητής για να λάβει απόφαση σχετικά με το πρότυπο εισόδου θα πρέπει να αντιλαμβάνεται το σύνολο του προτύπου, και επομένως η υπολειπόμενη εικόνα $\boldsymbol{x}_{ign}$ να έχει σχεδόν μηδενικό μέτρο. Στην πραγματικότητα όμως αυτό δεν συμβαίνει καθώς ακόμα και στα πρότυπα εκπαίδευσης το μέτρο της υπολειπόμενης εικόνας $\boldsymbol{x}_{ign}$ παραμένει μεγάλο. Όμως αυτό που μπορεί να παρατηρηθεί είναι ότι ανάμεσα σε ζεύγη πραγματικών εικόνων και επιθετικών εικόνων που έχουν παραχθεί από τις πραγματικές , το μέτρο της υπολειπόμενης εικόνας στις πραγματικές εικόνες είναι μικρότερο από το μέτρο της υπολειπόμενης εικόνας στις επιθετικές εικόνες. Το αποτέλεσμα αυτό είναι φανερό στο διάγραμμα του σχήματος 5.12 . Παρόλο που η παρατήρηση αυτή ισχύει για το μεγαλύτερο ποσοστό εικόνων, το χαρακτηριστικό αυτό δεν μπορεί να χρησιμοποιηθεί για τον εύκολο διαχωρισμό ανάμεσα στις πραγματικές και επιθετικές εικόνες, καθώς το μέτρο του $\boldsymbol{x}_{ign}$ μεταβάλλεται σημαντικά μεταξύ των πραγματικών εικόνων, πολύ περισσότερο από την μεταβολή που παρατηρείται μεταξύ πραγματικής και επιθετικής εικόνας.

Με βάση τις παραπάνω παρατηρήσεις επιθυμούμε μετά τον καθορισμό των ενεργοποιήσεων των μη γραμμικών στοιχείων, ο γραμμικός ταξινομητής που προκύπτει να αντιλαμβάνεται συνολικά το πρότυπο εισόδου της εικόνας $\boldsymbol{x}_0$. Δηλαδή επιθυμούμε η εικόνα $\boldsymbol{f}_{x_0}^T \boldsymbol{W}$ να προσεγγίσει το πρότυπο εισόδου ή αντίστοιχα να μειώσουμε το μέτρο της υπολειπόμενης εικόνας $\boldsymbol{x}_{0ign}$, χωρίς όμως να μεταβάλουμε την πραγματική κατηγορία της εικόνας εισόδου. Για να επιτύχουμε το παραπάνω μπορούμε να προσπαθήσουμε να μεταβάλουμε τις παραμέτρους του δικτύου ή να μεταβάλουμε την εικόνα εισόδου φροντίζοντας να μην μεταβάλουμε την πραγματική κατηγορία. Σε αυτό και στο επόμενο κεφάλαιο διερευνούμε τρόπους με τους οποίους μπορούμε να πετύχουμε τον παραπάνω στόχο μεταβάλλοντας την εικόνα εισόδου.

Κατά την μεταβολή της εικόνας εισόδου, αυτό που επιθυμούμε να πετύχουμε είναι να εισάγουμε την πληροφορία της εικόνας $\boldsymbol{x}_{0ign}$ στον ταξινομητή με τρόπο που δεν θα την αγνοεί. Αν απλά προσθέσουμε στην αρχική εικόνα $\boldsymbol{x}_0$ την εικόνα που αγνοείται $\boldsymbol{x}_{0ign}$, το αποτέλεσμα είναι να μην υπάρχει σχεδόν καθόλου αλλαγή στην απόφαση του ταξινομητή καθώς στην



Σχήμα 5.12: Μέση τιμή του λόγου του μέτρου της υπολοιπόμενης εικόνας $\boldsymbol{x}_{ign}$ προ το μέτρο της υπολειπόμενης εικόνας της πραγματικής εισόδου $\boldsymbol{x}_{0ign}$, καθώς το $\boldsymbol{x}$ κινείται από την πραγματική είσοδο $\boldsymbol{x}_0$ προς την επιθετική είσοδο

περιοχή γύρω από την αρχική εικόνα ο ταξινομητής συμπεριφέρεται σαν γραμμικός ταξινομητής στον οποίο προσθέτουμε είσοδο κάθετη στην επιφάνεια απόφασης.

Για την εικόνα εισόδου $\boldsymbol{x}$ μπορούμε να βρούμε με χρήση backpropagation το gradient του ταξινομητή, το οποίο προκαλεί την μεγαλύτερη μεταβολή στην βεβαιότητα του ταξινομητή για την κατηγορία στην οποία ταξινομείται η εικόνα $\boldsymbol{x}$. Επομένως θέλουμε να προσθέσουμε είσοδο που προσεγγίζει όσο γίνεται περισσότερο την διεύθυνση του *gradient* αλλά παράλληλα περιέχει την πληροφορία της υπολειπόμενης εικόνας $\boldsymbol{x}_{0ign}$. Το πρόβλημα που προκύπτει είναι ότι η εικόνα $\boldsymbol{x}_{0ign}$ είναι κάθετη στο *gradient*. Παρατηρούμε ότι αν εφαρμόσουμε *regularization* στην $\boldsymbol{x}_{0ign}$ τότε η εικόνα που προκύπτει ενώ διατηρεί την οπτική πληροφορία της $\boldsymbol{x}_{0ign}$ δεν είναι πλέον κάθετη στην διεύθυνση του *gradient*. Μπορούμε λοιπόν για κάθε *pixel* της εικόνα να λαμβάνουμε την κατεύθυνση του *gradient* που συμφωνεί με την κατεύθυνση της κανονικοποιημένης υπολειπόμενης εικόνας.

Έτσι αν $R(\boldsymbol{x})$ είναι η συνάρτηση *regularization* και **grad** είναι το *gradient* του ταξινομητή, η τελική εικόνα $\boldsymbol{x}_{add}$ που προσθέτω στην εικόνα εισόδου $\boldsymbol{x}$ για να ενσωματώσω την χαμένη πληροφορία της υπολειπόμενης εικόνας είναι :

$$\boldsymbol{x}_r = R(\boldsymbol{x}_{ign}) \tag{5.14}$$

$$\boldsymbol{x}_{add} = sgn(\boldsymbol{x}_r \odot \mathbf{grad}) \odot \mathbf{grad} \tag{5.15}$$

όπου με $\odot$ συμβολίζω τον πολλαπλασιασμό στοιχείο με στοιχείο ανάμεσα στα διανύσματα. Μια ακόμα μέθοδος για να υπολογίζω τις εικόνες $\boldsymbol{x}_{ign}$, $\boldsymbol{x}_p$ είναι μέσω του ψευδοαντίστροφου πίνακα του $\boldsymbol{A}$, δηλαδή τον $\boldsymbol{A}^{\dagger}$. Συγκεκριμένα γνωρίζοντας την έξοδο $\boldsymbol{y}$ από την εξίσωση (5.13) υπολογίζω το $\boldsymbol{x}_p$ ως εξής:

$$\boldsymbol{x}_p = \boldsymbol{A}^{\dagger}\boldsymbol{y} \tag{5.16}$$

οπότε

$$\boldsymbol{x}_{ign} = \boldsymbol{x} - \boldsymbol{A}^{\dagger}\boldsymbol{y} \tag{5.17}$$

Αυτό μας επιτρέπει αλλάζοντας το διάνυσμα εξόδου να βρίσκουμε διαφορετικά $\boldsymbol{x}_{ign}$.

5.4.2 Πειραματικές μέθοδοι για ανίχνευση επιθετικών εισόδων με χρήση της υπολειπόμενης εικόνας

Στην ενότητα αυτή παρουσιάζονται ορισμένες μέθοδοι ανίχνευσης επιθετικών εισόδων που χρησιμοποιούν την έννοια της υπολειπόμενης εικόνας που αναπτύχθηκε στο Κεφάλαιο 5.4.1

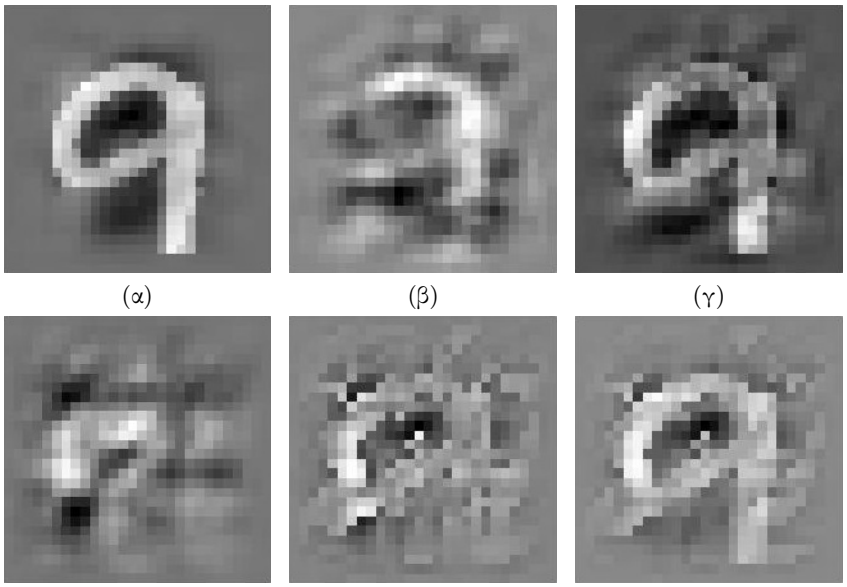
Μέθοδος Α

Στην πρώτη μέθοδο υπολογίζουμε επαναληπτικά την εικόνα $\boldsymbol{x}_{add}$, η οποία υπολογίζεται σύμφωνα με την εξίσωση (5.15), και την προσθέτουμε στην αρχική εικόνα. Έτσι η μέθοδος περιλαμβάνει τα παρακάτω βήματα:

1. Θέτω την αρχική εικόνα εισόδου ως την τρέχουσα εικόνα, $\boldsymbol{x}_{cur}(0) = \boldsymbol{x}_0$
2. Για την $\boldsymbol{x}_{cur}(t)$ υπολογίζω την $\boldsymbol{x}_{add}$ σύμφωνα με τις (5.14),(5.15), και την προσθέτω στην τρέχουσα εικόνα οπότε λαμβάνω την $\boldsymbol{x}_{cur}(t+1) = \boldsymbol{x}_{cur}(t) + \epsilon * \boldsymbol{x}_{add}$
3. Λαμβάνω την εκτίμηση του ταξινομητή για την τρέχουσα εικόνα $\boldsymbol{y}_{cur}(t+1)$
4. Αν έχω ξεπεράσει τον μέγιστο αριθμό επαναλήψεων ($t > T_{\max}$) τερματίζω την διαδικασία, διαφορετικά συνεχίζω από το βήμα 2 για $t = t + 1$

Μετά τον τερματισμό της μεθόδου αποφασίζω ότι η αρχική εικόνα εισόδου είναι επιθετική αν η τελική εικόνα ταξινομείται σε διαφορετική κατηγορία από την αρχική εικόνα. Στο συγκεκριμένο πείραμα χρησιμοποιώ total variation regularization ως την συνάρτηση R που αναφέρεται στην εξίσωση (5.14) . Επίσης κατά την εκτέλεση του πειράματος παρατηρώ ότι αν εκτελέσουμε προ-επεξεργασία της αρχικής εικόνας πριν εκτελέσουμε την μέθοδο τότε τα αποτελέσματα βελτιώνονται σημαντικά. Στην συγκεκριμένη περίπτωση εκτελώ tophat μετασχηματισμό της αρχικής εικόνας.

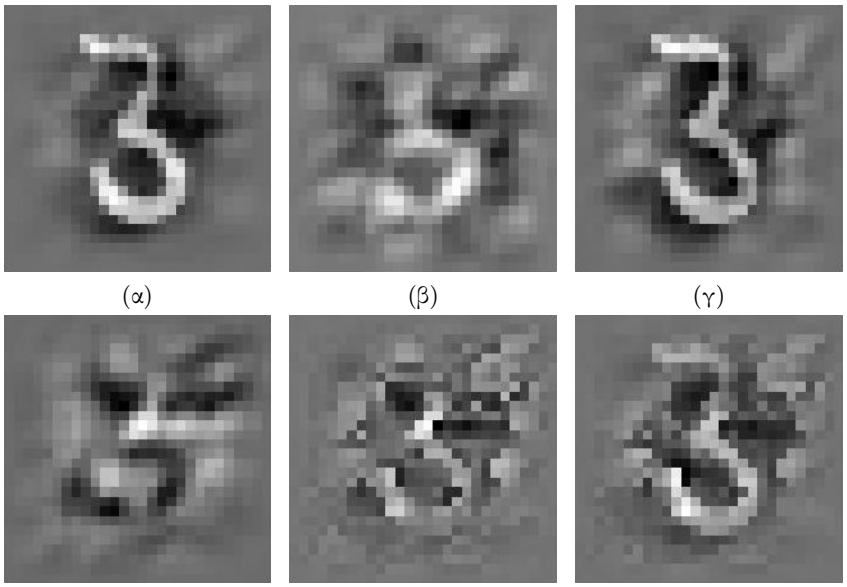
Στα Σχήματα 5.13, 5.15 βλέπουμε παραδείγματα όπου η διαδικασία δουλεύει σωστά ενώ στα Σχήματα 5.14, 5.16 βλέπουμε προβληματικές περιπτώσεις. Στην περίπτωση της επιθετικής εικόνας που αναγνωρίζεται από την μέθοδο στο Σχήμα 5.13, βλέπουμε την σωστή λειτουργία της μεθόδου όπου η εικόνα που προσθέτουμε στην αρχική, ενισχύει το μέρος του αρχικού προτύπου που αγνοείται. Αντίθετα στην περίπτωση της επιθετικής εικόνας που δεν μπορούμε να αναγνωρίσουμε στο Σχήμα 5.14, βλέπουμε ότι η εικόνα που προσθέτουμε προσεγγίζει περισσότερο την διεύθυνση του gradient που μεγιστοποιεί την λάθος κατηγορία. Αυτό οφείλεται στο γεγονός ότι στην εικόνα $\boldsymbol{x}_p$ που αντιλαμβάνεται ο ταξινομητής, τα σημεία υψηλής έντασης δεν πλησιάζουν την τιμή της ένταση που έχουν στην αρχική εικόνα με αποτέλεσμα να εμφανίζονται με υψηλή ένταση και στην εικόνα που αγνοείται. Το πρόβλημα αυτό το επιλύει σε έναν βαθμό η μέθοδος Β που παρουσιάζουμε στην συνέχεια. Τέλος είναι φανερό ότι η εικόνα που προστίθεται είναι θορυβώδης και παρουσιάζει κορυφές και σε σημεία που δεν ανήκουν στο πρότυπο, όπως στο Σχήμα 5.16 το οποίο καταλήγει να μην αναγνωρίζεται με επιτυχία ως πραγματική εικόνα.



(α) (β) (γ)

(δ) (ε) (ϝ)

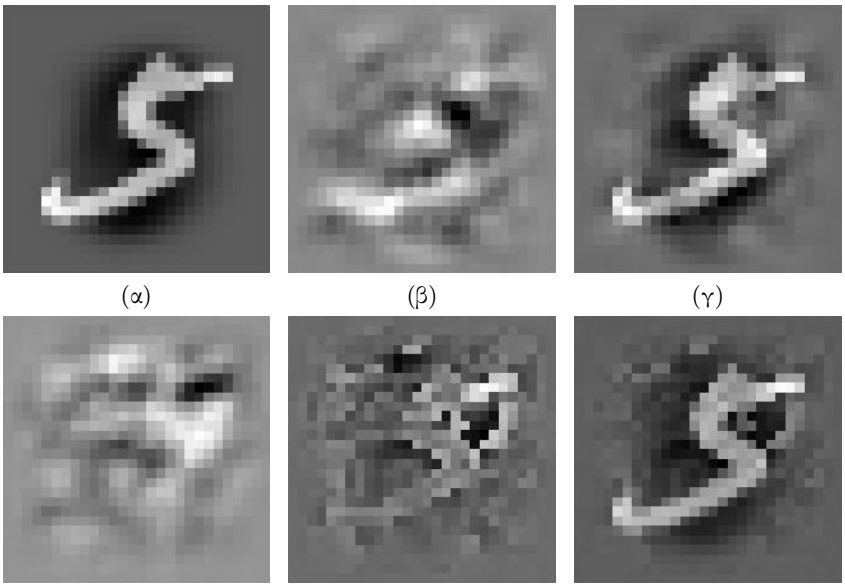
Σχήμα 5.13: Παράδειγμα επιθετικής εισόδου που εντοπίζει η μέθοδος A : (α) Αρχική είσοδος που αναγνωρίζεται ως το ψηφίο /7/ (β)Εικόνα που αντιλαμβάνεται ο ταξινομητής (γ) Προβολή της αρχικής εικόνας στον μηδενοχώρο του ταξινομητή (δ) Διεύθυνση παραγώγου που ελαχιστοποιεί την κατηγορία /7/ (ε)Εικόνα που προσθέτουμε στην αρχική εικόνα (ϝ) Τελική εικόνα που δεν αναγνωρίζεται πλέον ως /7/



(α) (β) (γ)

(δ) (ε) (ϝ)

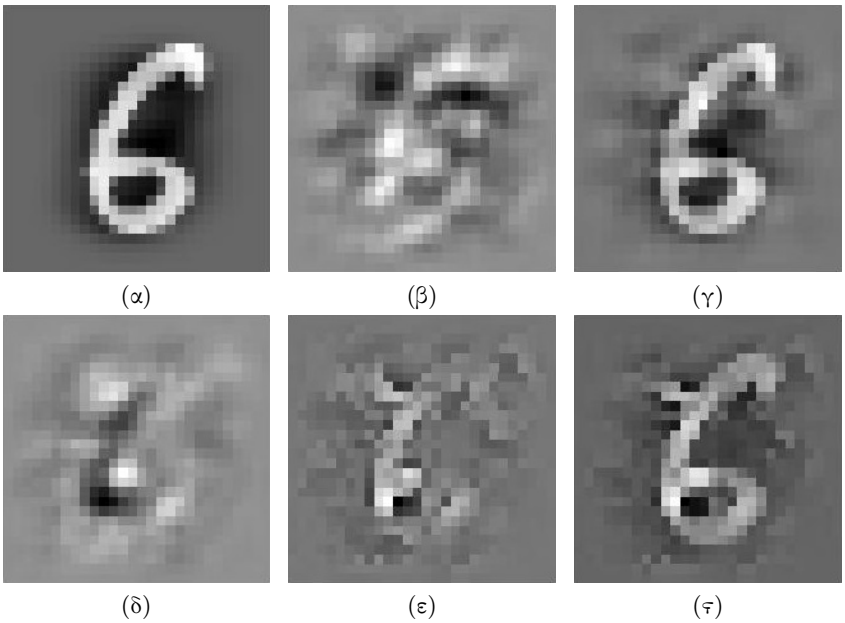
Σχήμα 5.14: Παράδειγμα επιθετικής εισόδου που δεν εντοπίζει η μέθοδος A : (α) Αρχική είσοδος που αναγνωρίζεται ως το ψηφίο /5/ (β)Εικόνα που αντιλαμβάνεται ο ταξινομητής (γ) Προβολή της αρχικής εικόνας στον μηδενοχώρο του ταξινομητή (δ) Διεύθυνση παραγώγου που ελαχιστοποιεί την κατηγορία /5/ (ε)Εικόνα που προσθέτουμε στην αρχική εικόνα (ϝ) Τελική εικόνα που αναγνωρίζεται ακόμα ως /5/



(α) (β) (γ)

(δ) (ε) (ϛ)

Σχήμα 5.15: Παράδειγμα πραγματικής εισόδου που ανιχνεύεται με επιτυχία ως πραγματική από την μέθοδο A : (α) Αρχική είσοδος που αναγνωρίζεται ως το ψηφίο /5/ (β)Εικόνα που αντιλαμβάνεται ο ταξινομητής (γ) Προβολή της αρχικής εικόνας στον μηδενοχώρο του ταξινομητή (δ) Διεύθυνση παραγώγου που ελαχιστοποιεί την κατηγορία /5/ (ε)Εικόνα που προσθέτουμε στην αρχική εικόνα (ϛ) Τελική εικόνα που αναγνωρίζεται ακόμα ως /5/



(α) (β) (γ)

(δ) (ε) (ϛ)

Σχήμα 5.16: Παράδειγμα πραγματικής εισόδου που ανιχνεύεται ως επιθετική από την μέθοδο A: (α) Αρχική είσοδος που αναγνωρίζεται ως το ψηφίο /6/ (β)Εικόνα που αντιλαμβάνεται ο ταξινομητής (γ) Προβολή της αρχικής εικόνας στον μηδενοχώρο του ταξινομητή (δ) Διεύθυνση παραγώγου που ελαχιστοποιεί την κατηγορία /6/ (ε)Εικόνα που προσθέτουμε στην αρχική εικόνα (ϛ) Τελική εικόνα που δεν αναγνωρίζεται πλέον ως /6/

Μέθοδος Β

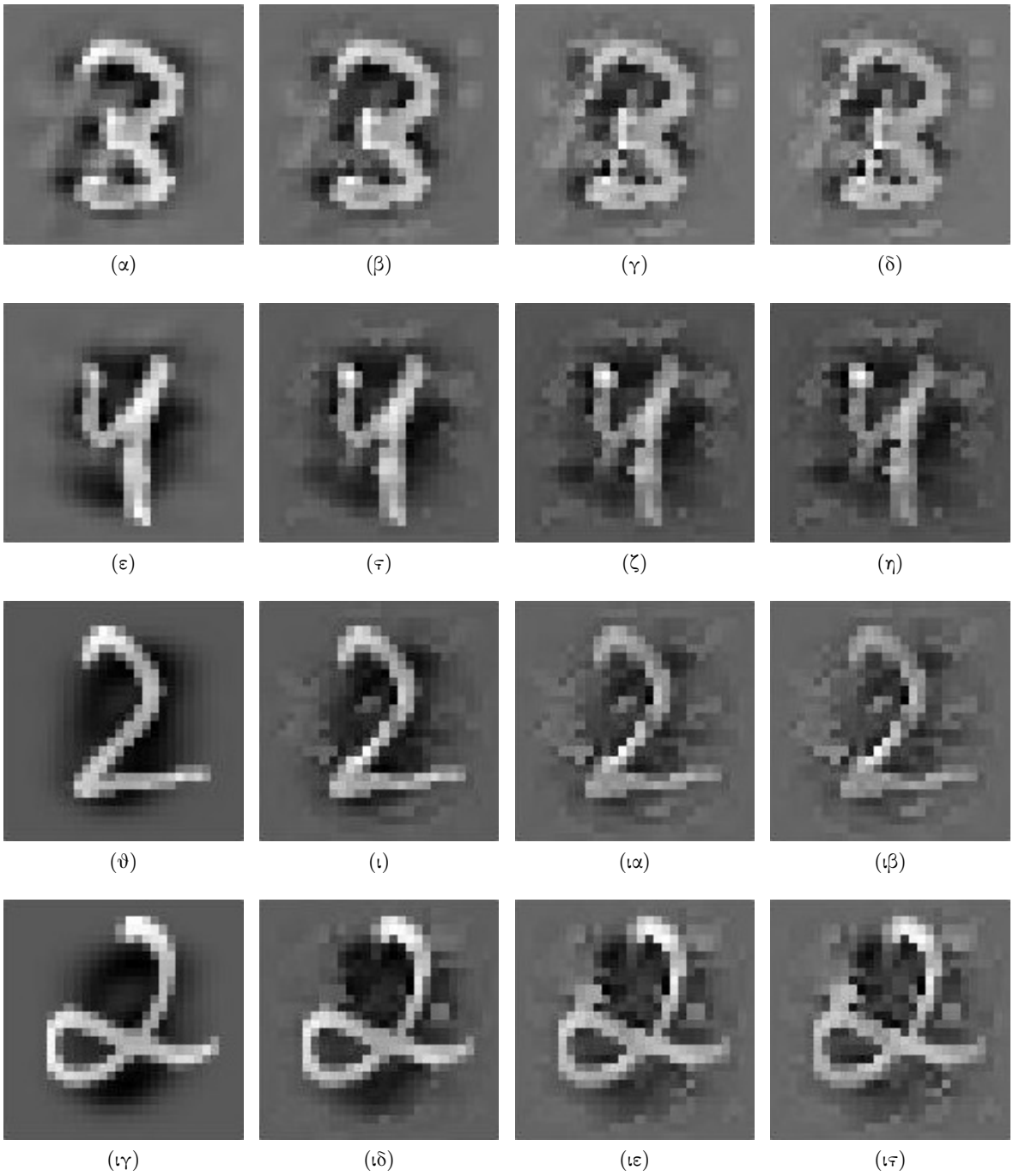
Στην μέθοδο αυτή ακολουθώ τα βήματα της μεθόδου Α με την μόνη διαφορά ότι υπολογίζω την $\boldsymbol{x}_{ign}$ χρησιμοποιώντας τις εξισώσεις (5.16),(5.17) αφού μεταβάλω το διάνυσμα εξόδου $\boldsymbol{y}$.

Συγκεκριμένα για να βελτιώσουμε την επίδοση της μεθόδου Α, επιθυμούμε η εικόνα $\boldsymbol{x}_p$ να τείνει σε μια καθαρή εικόνα ενός προτύπου που αντιστοιχεί στην κατηγορία που ταξινομείται η $\boldsymbol{x}$, είτε αυτή είναι πραγματική είτε επιθετική είσοδος. Δεδομένου ότι η έξοδος του τελευταίου επιπέδου του δικτύου καθορίζει τα χαρακτηριστικά της εισόδου που αντιλαμβάνεται ο ταξινομητής , επιθυμούμε η έξοδος αυτή να αντιστοιχεί σε χαρακτηριστικά πραγματικών προτύπων της κατηγορίας στην οποία ταξινομείται το πρότυπο. Αυτό έχει ως συνέπεια τα σημεία υψηλής έντασης της $\boldsymbol{x}_p$ να προσεγγίζουν την ένταση που έχουν στο αρχικό πρότυπο με αποτέλεσμα να μην εμφανίζονται με υψηλή ένταση στην αγνοούμενη εικόνα $\boldsymbol{x}_{ign}$. Ο καλύτερος διαχωρισμός των σημείων με υψηλή ένταση μεταξύ της $\boldsymbol{x}_p$ και της $\boldsymbol{x}_{ign}$ αντιμετωπίζει ορισμένα από τα προβλήματα της μεθόδου Α που παρουσιάστηκαν παραπάνω.

Σύμφωνα με το [10] η έξοδος των επιθετικών παραδειγμάτων βρίσκεται εκτός του $manifold$ στο οποίο βρίσκονται οι έξοδοι των πραγματικών εισόδων. Έτσι επιθυμώ κάθε $\boldsymbol{y}$ να το προβάλω στο $manifold$ στο οποίο βρίσκονται οι έξοδοι των προτύπων εκπαίδευσης. Αυτό το υλοποιώ λαμβάνοντας τις εξόδους των προτύπων εκπαίδευσης και εκτελώντας ένα απλό $kmeans$ με τον οποίο βρίσκω τα κέντρα των $clusters$ που δημιουργούν οι έξοδοι των πραγματικών προτύπων στον χώρο εξόδου. Στην συνέχεια για κάθε $\boldsymbol{y}$ λαμβάνω το $\boldsymbol{y}_{\text{cent}}$ το οποίο είναι το πλησιέστερο κέντρο στο αρχικό διάνυσμα $\boldsymbol{y}$. Αν αντικαταστήσω στις εξισώσεις (5.16),(5.17) το $\boldsymbol{y}$ με το $\boldsymbol{y}_{\text{cent}}$ λαμβάνω την νέα υπολειπόμενη εικόνα την οποία συμβολίζω ως $\widetilde{\boldsymbol{x}}_{ign}$.

$$\widetilde{\boldsymbol{x}}_{ign} = \boldsymbol{x} - \boldsymbol{A}^{\dagger}\boldsymbol{y}_{\text{cent}} \tag{5.18}$$

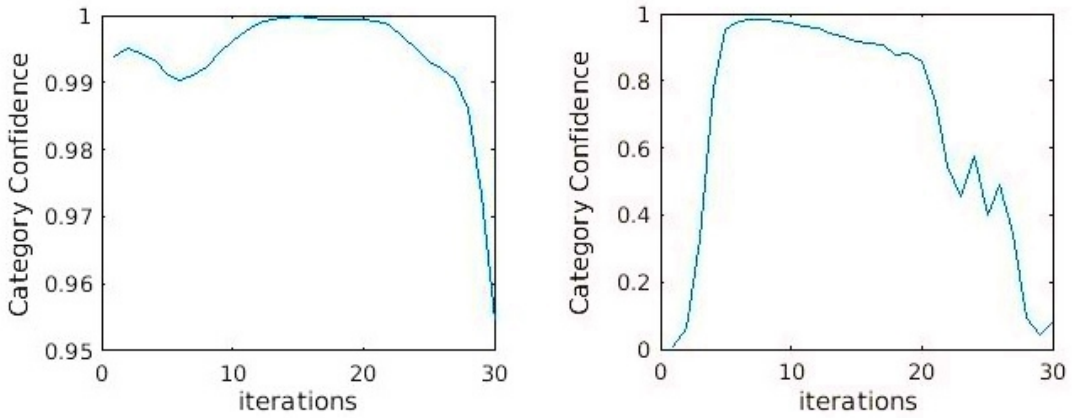
Ακολουθώ λοιπόν τα βήματα που περιγράφτηκαν στην μέθοδο Α , αντικαθιστώντας στην εξίσωση (5.14) το $\boldsymbol{x}_{ign}$ με το νέο $\widetilde{\boldsymbol{x}}_{ign}$. Στο Σχήμα 5.17 παρουσιάζονται παραδείγματα της εξέλιξης της μεθόδου σε διαφορετικό αριθμό επαναλήψεων για διαφορετικά πρότυπα. Ένα βασικό μειονέκτημα της μεθόδου που γίνεται εμφανές είναι ότι μετά από έναν δεδομένο αριθμό επαναλήψεων οι εικόνες γίνονται αρκετά θορυβώδης και αρχίζουν να εμφανίζονται σημεία υψηλής έντασης τα οποία δεν συμβαδίζουν με την μορφή του αρχικού προτύπου.



Σχήμα 5.17: Σε κάθε γραμμή φαίνεται πως η μέθοδος B μεταβάλει επαναληπτικά την αρχική εικόνα. Συγκεκριμένα στην πρώτη στήλη παρουσιάζονται οι αρχικές εικόνες, στην δεύτερη οι εικόνες μετά από 4 επαναλήψεις, στην τρίτη οι εικόνες μετά από 8 επαναλήψεις και στην τέταρτη οι εικόνες μετά από 10 επαναλήψεις. Η εικόνα 5.17α είναι επιθετική είσοδος που αναγνωρίζεται από την μέθοδο, η εικόνα 5.17ε είναι επιθετική είσοδος που δεν αναγνωρίζεται από την μέθοδο, η εικόνα 5.17ϑ είναι πραγματική εικόνα η οποία δεν αναγνωρίζεται ως επιθετική από την μέθοδο, η εικόνα 5.17ιγ είναι πραγματική εικόνα η οποία αναγνωρίζεται ως επιθετική από την μέθοδο.

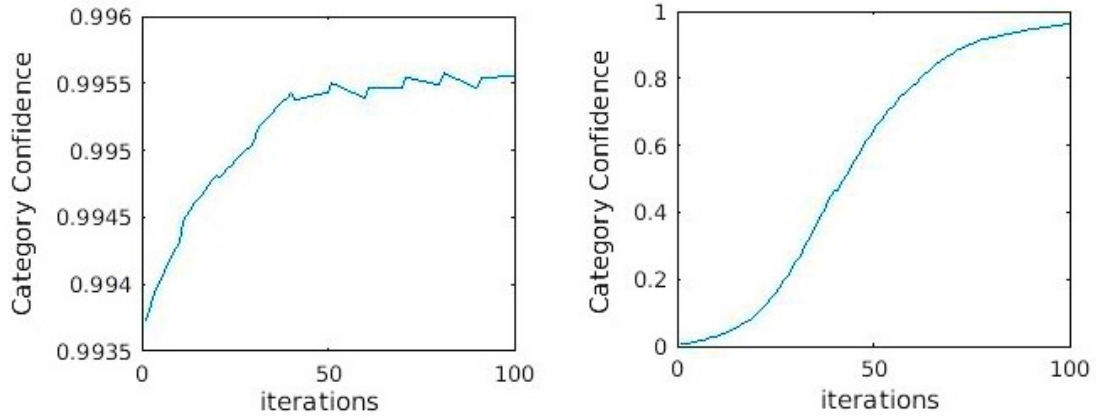
Μέθοδος Γ

Οι παραπάνω επαναληπτικές μέθοδοι έχουν το μειονέκτημα ότι δεν συγκλίνουν σε μια τελική εικόνα, οπότε μετά από έναν αριθμό επαναλήψεων αρχίζουν και αποκλίνουν από την κατηγορία στην οποία θα έπρεπε ιδανικά να οδηγηθούν από την διαδικασία. Αυτό είναι φανερό στα διαγράμματα του Σχήματος 5.18 όπου βλέπουμε την βεβαιότητα του ταξινομητή για την σωστή κατηγορία ενός ψηφίου τόσο στην περίπτωση της πραγματικής εισόδου όσο και στην περίπτωση επιθετικής εισόδου. Έτσι είναι αναγκαίο να βρίσκουμε έναν μέγιστο αριθμό επαναλήψεων ο οποίος είναι αρκετά μεγάλος έτσι ώστε να παρατηρηθεί η μεταβολή που επιθυμούμε αλλά και αρκετά μικρός ώστε να μην προλάβει να αποκλίνει το αποτέλεσμα οδηγώντας σε λάθος συμπεράσματα.



(α) (β)

Σχήμα 5.18: Βεβαιότητα του ταξινομητή για την σωστή κατηγορία με χρήση της μεθόδου Β: (α) Πραγματική είσοδος που αναγνωρίζεται ως το ψηφίο /8/ (β) Επιθετική είσοδο που η σωστή κατηγορία αναγνώρισης είναι το ψηφίο /8/



(α) (β)

Σχήμα 5.19: Βεβαιότητα του ταξινομητή για την σωστή κατηγορία με χρήση της μεθόδου Γ: (α) Πραγματική είσοδος που αναγνωρίζεται ως το ψηφίο /8/ (β) Επιθετική είσοδο που η σωστή κατηγορία αναγνώρισης είναι το ψηφίο /8/

Παρατηρώ ότι αν μεταβάλω τον τρόπο με τον οποίο υπολογίζω τα $\boldsymbol{x}_{ign}$ και τα $\boldsymbol{x}_{add}$ μπορώ να πετύχω καλύτερη συμπεριφορά όσο αφορά την σύγκλιση. Τα τροποποιημένα $\boldsymbol{x}_{ign}$,$\boldsymbol{x}_{add}$ για την μέθοδο αυτή τα συμβολίζω ως $\hat{\boldsymbol{x}}_{ign}$,$\hat{\boldsymbol{x}}_{add}$ και τα υπολογίζω ως εξής:

$$\hat{\boldsymbol{x}}_{ign} = \frac{(\boldsymbol{x} \odot \boldsymbol{x}_0) - (\boldsymbol{x} \odot \boldsymbol{x}_p)}{\|\boldsymbol{x}\|_2} \tag{5.19}$$

$$\hat{\boldsymbol{x}}_{add} = |\hat{\boldsymbol{x}}_{ign} \odot \mathbf{grad}| \odot \mathrm{sgn}(\hat{\boldsymbol{x}}_{ign}) \tag{5.20}$$

όπου με $\odot$ συμβολίζω τον πολλαπλασιασμό στοιχείο με στοιχείο ανάμεσα στα διανύσματα.

Οι εξισώσεις αυτές διαφοροποιούνται από τις αρχικές από το ότι δίνουμε μεγαλύτερη βαρύτητα στις διαφορές της $\boldsymbol{x}$ με την $\boldsymbol{x}_p$ σε σημεία όπου η ένταση της αρχικής εικόνας είναι μεγαλύτερη. Χρησιμοποιώντας τις εξισώσεις αυτές η επαναληπτική διαδικασία συγκλίνει σε ένα αποτέλεσμα, οπότε δεν είναι τόσο ευαίσθητη στον μέγιστο αριθμό επαναλήψεων, ο οποίος αρκεί να είναι αρκετά μεγάλος έτσι ώστε να προλάβει να γίνει η σύγκλιση της μεθόδου.

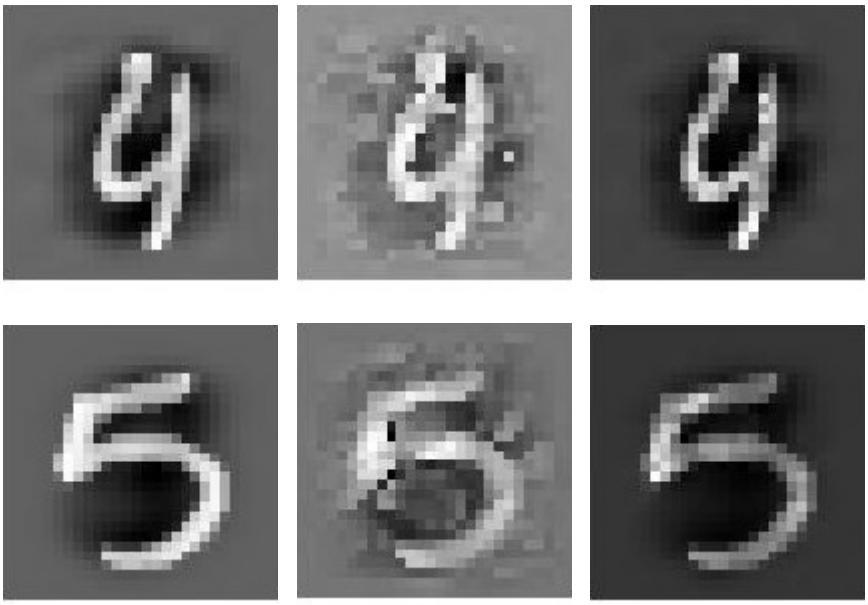
Ένα παράδειγμα της σύγκλισης φαίνεται στο διάγραμμα του Σχήματος 5.19 όπου βλέπουμε τα αντίστοιχα με το διάγραμμα του Σχήματος 5.18 αποτελέσματα, αλλά χρησιμοποιώντας τις εξισώσεις (5.19),(5.20) .

Μια ακόμα λεπτομέρεια η οποία βελτιώνει τα αποτελέσματα της μεθόδου είναι ο περιορισμός της εικόνας που παράγεται επαναληπτικά από την μέθοδο μέσα σε ένα εύρος τιμών. Αυτό προστατεύει την διαδικασία από σημεία της εικόνας τα οποία λαμβάνουν πολύ μεγάλες ή πολύ μικρές τιμές και οδηγούν την μέθοδο σε λανθασμένα αποτελέσματα. Έτσι έχοντας επιλέξει μια ελάχιστη τιμή v_{min} και μια μέγιστη τιμή v_{max}, για να ανανεώσουμε στην t επανάληψη το pixel i της εικόνας $\boldsymbol{x}_{cur}$,το οποίο συμβολίζουμε ως $x_{cur}(t)[i]$, εκτελούμε την πράξη

$$x_{cur}(t+1)[i] = \max(\min(x_{cur}(t)[i] + \epsilon * \hat{x}_{add}[i], v_{max}), v_{min})$$

Συνδυάζοντας τις δύο αυτές παραλλαγές της αρχικής μεθόδου πετυχαίνω καλύτερη σύγκλιση και παράλληλα τα αποτελέσματα αναγνώρισης δεν εξαρτώνται από τον αριθμό επαναλήψεων, μια παράμετρος που επηρέαζε σημαντικά τις προηγούμενες δύο μεθόδους. Στο Σχήμα 5.20 βλέπουμε την σύγκριση των μεθόδων Β,Γ. Είναι φανερό ότι η μέθοδος Γ εισάγει λιγότερο θόρυβο στις εικόνες σε σχέση με τον θόρυβο που εισάγεται στην Β και την οδηγεί σε λανθασμένα αποτελέσματα μετά από δεδομένο αριθμό επαναλήψεων. Επίσης έχει ενδιαφέρον να παρατηρήσουμε ότι σε αντίθεση με την μέθοδο Β, οι κορυφές έντασης που εισάγονται στο τελικό αποτέλεσμα της Γ είναι μόνο αυτές που συμβαδίζουν με την αρχική μορφή του προτύπου.

Το κύριο υπολογιστικό κόστος των παραπάνω μεθόδων αποτελεί ο υπολογισμός της υπολειπόμενης εικόνας $\boldsymbol{x}_{ign}$ ($\widetilde{\boldsymbol{x}}_{ign}$, $\hat{\boldsymbol{x}}_{ign}$ για τις μεθόδους Β,Γ αντίστοιχα). Για τον υπολογισμός αυτόν αρχικά είναι απαραίτητος ο υπολογισμός του πίνακα $\boldsymbol{A}$ της εξίσωσης (5.13). Ο τρόπος που υλοποιούμε τον υπολογισμό αυτόν είναι με εκτέλεση του αλγορίθμου backpropagation για κάθε μια από τις εξόδους του δικτύου, καθώς το gradient της εξόδου i προς το διάνυσμα εισόδου είναι ίσο με την i γραμμή του πίνακα $\boldsymbol{A}$. Γενικά για ένα πλήρως συνδεδεμένο νευρωνικό η πολυπλοκότητα του backpropagation είναι γραμμική ως προς τον αριθμό των παραμέτρων του δικτύου, δηλαδή $O(W_L)$ όπου W_L είναι ο αριθμός των παραμέτρων. Βέβαια στην περίπτωση όπου έχουμε ένα συνελικτικό νευρωνικό η πολυπλοκότητα ενός συνελικτικού επιπέδου δεν ισούται με τον αριθμό των παραμέτρων των φίλτρων, αλλά καθώς το συνελικτικό νευρωνικό μπορεί να υλοποιηθεί με ένα πλήρως συνδεδεμένο νευρωνικό με κατάλληλη ανάθεση των βαρών, θεωρούμε ότι η πολυπλοκότητα παραμένει $O(W_L)$ όπου W_L αυτή την φορά είναι ο αριθμός των μή μηδενικών βαρών του πλήρως συνδεδεμένου δικτύου που προσεγγίζει το συνελικτικό νευρωνικό. Αφού λοιπόν εκτελούμε το backpropagation για κάθε έξοδο, η συνολική πολυπλοκότητα είναι $O(MW_L)$, όπου M είναι το μέγεθος του διανύσματος εξόδου.



(α) (β) (γ)

Σχήμα 5.20: Σύγκριση των οπτικών αποτελεσμάτων μεταξύ των μεθόδων Β,Γ : (α) Αρχικές εικόνες (β) Εικόνες μετά το τερματισμό της Β μεθόδου (γ) Εικόνες μετά τον τερματισμό της Γ μεθόδου

Η πολυπλοκότητα αυτή αναφέρεται στην σειριακή εκτέλεση του backpropagation , γεγονός το οποίο δεν αντικατοπτρίζει την πραγματική επίδοση της μεθόδου αφού ο αλγόριθμος είναι σε μεγάλο βαθμό παραλληλοποιήσιμος με αποτέλεσμα να μπορεί να τρέξει αποδοτικά σε συστήματα παράλληλης επεξεργασίας.

Μετά τον υπολογισμός του πίνακα $\boldsymbol{A}$, για τον υπολογισμό του $\boldsymbol{x}_{ign}$ πρέπει είτε να υπολογιστεί ο μηδενοχώρος του $\boldsymbol{A}$ είτε να βρεθεί ο $\boldsymbol{A}^{\dagger}$. Και οι δύο υπολογισμοί μπορεί να γίνουν με χρήση του αλγορίθμου SVD , με πολυπλοκότητα $O(\max(N^2M, NM^2))$ όπου N το μέγεθος του διανύσματος εισόδου και M το μέγεθος του διανύσματος εξόδου. Συνολικά οι παραπάνω υπολογισμοί γίνονται στην χειρότερη περίπτωση T φορές , όπου T είναι ο μέγιστος αριθμός επαναλήψεων ο οποίος διαφέρει ανάμεσα στις μεθόδους. Έτσι η συνολική πολυπλοκότητα είναι

$$O(TMW_L) + O(T\max(N^2M, NM^2))$$

Παρόλο που η παραπάνω πολυπλοκότητα δεν αντικατοπτρίζει την πραγματική επίδοση που μπορεί να έχει μια παράλληλη υλοποίηση της μεθόδου,το υπολογιστικό της κόστος είναι αρκετά μεγαλύτερο σε σύγκριση με τις προηγούμενες μεθόδους που παρουσιάστηκαν, εξαιτίας της επαναληπτικής εκτέλεση των βημάτων της μεθόδου T φορές.

5.4.3 Αποτελέσματα ανίχνευσης επιθετικών εισόδων σε MNIST , CIFAR-10

Αρχικά εκτελέσαμε πείραμα ανίχνευσης επιθετικών εισόδων στο συνελικτικό νευρωνικό δίκτυο που έχει εκπαιδευτεί στο MNIST για αναγνώριση ασπρόμαυρων εικόνων ψηφίων.

Από το σύνολο εικόνων που δεν χρησιμοποιήθηκε κατα την εκπαίδευση επιλέγουμε αυτές οι οποίες ταξινομούνται σωστά από το εκπαιδευμένο δίκτυο και με κάθε μια από αυτές δοκιμάζουμε να παράξουμε μια επιθετική είσοδο. Έτσι δημιουργούμε ένα σύνολο από 2000

πραγματικές εισόδους και 2000 επιθετικές εισόδους. Από το σύνολο αυτό προσπαθούμε να εντοπίσουμε τις επιθετικές εισόδους χρησιμοποιώντας τις μεθόδους που αναπτύχθηκαν στο προηγούμενο κεφάλαιο.

Οι παράμετροι που χρησιμοποιήθηκαν για το πείραμα στο MNIST σύνολο δεδομένων είναι

- Για την μέθοδο Α χρησιμοποιήθηκε $\epsilon = 0.1$, $T_{\max} = 10$
- Για την μέθοδο Β χρησιμοποιήθηκε $\epsilon = 0.1$, $T_{\max} = 10$, ενώ για την εκτέλεση του kmeans για την εύρεση των clusters των διανυσμάτων χαρακτηριστικών θεωρώ ότι υπάρχουν 200 κέντρα
- Για την μέθοδο Γ χρησιμοποιήθηκε $\epsilon = 1$, $T_{\max} = 80$

Την ίδια διαδικασία ακολουθούμε και για το συνελικτικό δίκτυο που έχει εκπαιδευτεί στο CIFAR-10 . Ομοίως με το MNIST χρησιμοποιούμε 2000 πραγματικές εισόδους και 2000 επιθετικές εισόδους και προσπαθούμε να εντοπίσουμε τις επιθετικές εισόδους. Στο Σχήμα 5.21 βλέπουμε ορισμένα παραδείγματα της μεθόδου Α σε εικόνες του CIFAR-10.

Οι παράμετροι που χρησιμοποιήθηκαν για το πείραμα στο CIFAR-10 σύνολο δεδομένων είναι

- Για την μέθοδο Α χρησιμοποιήθηκε $\epsilon = 0.8$, $T_{\max} = 12$
- Για την μέθοδο Β χρησιμοποιήθηκε $\epsilon = 0.6$, $T_{\max} = 14$, ενώ για την εκτέλεση του kmeans για την εύρεση των clusters των διανυσμάτων χαρακτηριστικών θεωρώ ότι υπάρχουν 300 κέντρα
- Για την μέθοδο Γ χρησιμοποιήθηκε $\epsilon = 1$, $T_{\max} = 120$

Συνολικά τα αποτελέσματα για τα δύο σύνολα δεδομένων και τα δύο εκπαιδευμένα δίκτυα παρουσιάζονται στον Πίνακα 5.3

$MNIST$		
	Precision	Recall
Μέθοδος Α	78.1%	76.2%
Μέθοδος Β	86.1%	85.6%
Μέθοδος Γ	87.9%	88.1%
$CIFAR-10$		
Μέθοδος Α	65.4%	70%
Μέθοδος Β	75.7%	72.1%
Μέθοδος Γ	73.6%	71.3%

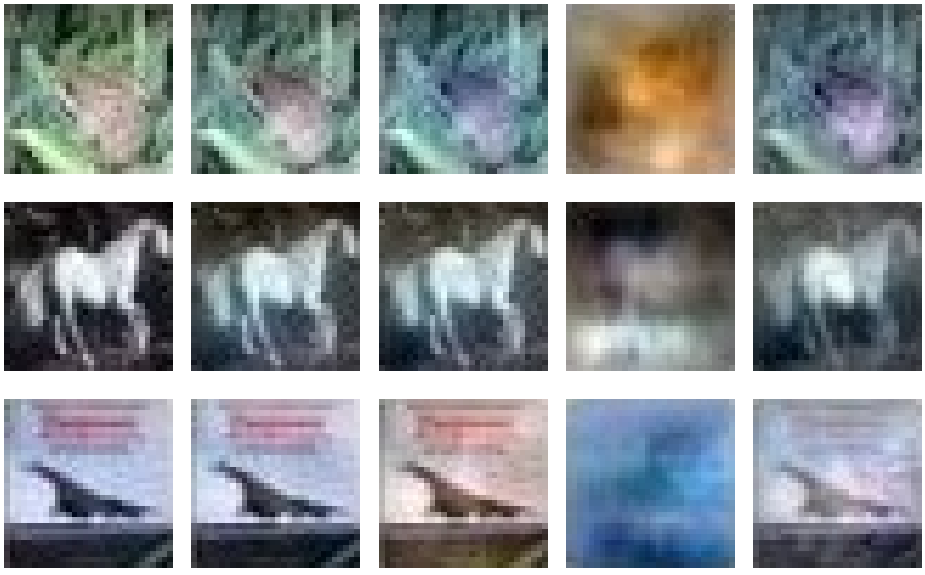
Πίνακας 5.3: Αποτελέσματα των μεθόδων του κεφαλαίου 5.4.2 στο σύνολο δεδομένων MNIST και CIFAR-10

Είναι φανερό και για τα δύο σύνολα δεδομένων η χρήση της μεθόδου Β, χρησιμοποιώντας το τροποποιημένο $\widetilde{\boldsymbol{x}}_{ign}$ που προκύπτει από την εξίσωση (5.18), μας οδηγεί σε καλύτερα αποτελέσματα από ότι η αρχική μέθοδος Α. Ενώ όσο αφορά την σύγκριση μεταξύ των μεθόδων Β,Γ παρατηρείται ότι τα αποτελέσματά τους είναι αρκετά κοντά, με την μέθοδο Γ να πετυχαίνει την καλύτερη επίδοση στο MNIST σύνολο δεδομένων και την Β να πετυχαίνει καλύτερη επίδοση στο CIFAR-10 σύνολο δεδομένων. Αυτό όμως που πρέπει να ληφθεί επίσης υπόψιν είναι ότι για να πετύχουμε με την μέθοδο Β τα αποτελέσματα που παρουσιάζονται πρέπει να βρούμε τις

παραμέτρους με τις οποίες έχουμε την επιθυμητή μεταβολή χωρίς όμως να αποκλίνουμε από το αποτέλεσμα, κάτι το οποίο δεν είναι αναγκαίο για την Γ μέθοδο που εμφανίζει πιο ευσταθή συμπεριφορά.

Επίσης επιθυμούμε να εξετάσουμε τα αποτελέσματα αναγνώρισης της μεθόδου όταν οι επιθετικές εικόνες έχουν παραχθεί με την μέθοδο DeepFool αντί της μεθόδου FGSM . Έτσι εκτελούμε τα προηγούμενα πειράματα για τις ίδιες πραγματικές εικόνες και για 2000 επιθετικές που προκύπτουν από την DeepFool μέθοδο, όπου για να τις παράξουμε χρησιμοποιούμε τις ίδιες πραγματικές εικόνες που χρησιμοποιήσαμε και στα προηγούμενα πειράματα. Τα αποτελέσματα που προκύπτουν μας δίνουν ποσοστά αναγνώρισης, τα οποία διαφέρουν από αυτά του Πίνακα 5.3 από 0.2% έως 2.6%. Η μικρή μεταβολή των αποτελεσμάτων μας δείχνει ότι οι μέθοδοι του κεφαλαίου λειτουργούν με πανομοιότυπο τρόπο ανεξάρτητα από το αν οι επιθετικές εικόνες έχουν προκύψει από την FGSM ή την DeepFool μέθοδο.

Αξίζει βέβαια να σημειωθεί ότι η ανεξαρτησία που παρατηρούμε στα αποτελέσματα των μεθόδων, αναφέρεται συγκεκριμένα για τις δύο μεθόδους FGSM,DeepFool. Γενικά είναι αρκετά δύσκολο να καταλήξουμε σε μέθοδο η οποία είναι πλήρως ανεξάρτητη από την διαδικασία παραγωγής των επιθετικών εικόνων. Αυτό γιατί ο χρήστης ο οποίος παράγει τις επιθετικές εισόδους, μπορεί πάντα έχοντας γνώση της στρατηγικής άμυνας να αλλάξει την μέθοδο παραγωγής έτσι ώστε να στοχεύσει την συγκεκριμένη στρατηγική. Ένα τέτοιο παράδειγμα είναι η μέθοδος C&W που αναφέρουμε στο Κεφάλαιο 2.3 η οποία αναπτύχθηκε με σκοπό να ξεπεράσει το Defense Distillitation, και στην συνέχεια όπως αναφέρεται στο [4] αναπτύχθηκε σε διαφορετικές παραλλαγές όπου κάθε μια στόχευε συγκεκριμένες στρατηγικές άμυνας.



Σχήμα 5.21: Στην (α) στήλη βλέπουμε εικόνες από το CIFAR-10 οι οποίες ταξινομούνται σωστά. Στην στήλη (β) βλέπουμε εικόνες οι οποίες έχουν παραχθεί από αυτές της στήλης (α) και ταξινομούνται λανθασμένα. Στην στήλη (γ) βλέπουμε το μέρος των εικόνων το οποίο δεδομένων των ενεργοποιήσεων των μη γραμμικών στοιχείων τις αγνοεί ο ταξινομητής. Στην στήλη (δ) βλέπουμε το μέρος των εικόνων που λαμβάνεται υπόψιν στην απόφαση. Στην στήλη (ε) βλέπουμε τις τελικές εικόνες οι οποίες παράγονται από την μέθοδο Α. Η επιθετική εικόνα της τελευταίας γραμμής δεν εντοπίζεται με επιτυχία ως επιθετική είσοδο και συνεχίζει να ταξινομείται στην κατηγορία /πλοίο/

5.5 Σύγκριση και συνδυασμός των τριών προτεινόμενων μεθόδων άμυνας

Κατά την σύγκριση των τριών μεθόδων του κεφαλαίου 5 θα αναφερόμαστε στις αντίστοιχες μεθόδους ως

- Μέθοδος Ομαλοποίησης : για την μέθοδο του Κεφαλαίου 5.2 που εκτελεί ομαλοποίηση του διανύσματος χαρακτηριστικών
- Μέθοδος Ιστογραμμάτων : για την μέθοδο του Κεφαλαίου 5.3 που χρησιμοποιεί τα ιστογράμματα των εξόδων του νευρωνικού για την ανίχνευση των επιθετικών εικόνων
- Μέθοδος Υπολειπόμενης Εικόνας: για τις μεθόδους του Κεφαλαίου 5.4.2 που χρησιμοποιούν την υπολειπόμενη εικόνα, και συγκεκριμένα για την τελευταία μέθοδο του κεφαλαίου (Μέθοδος Γ)

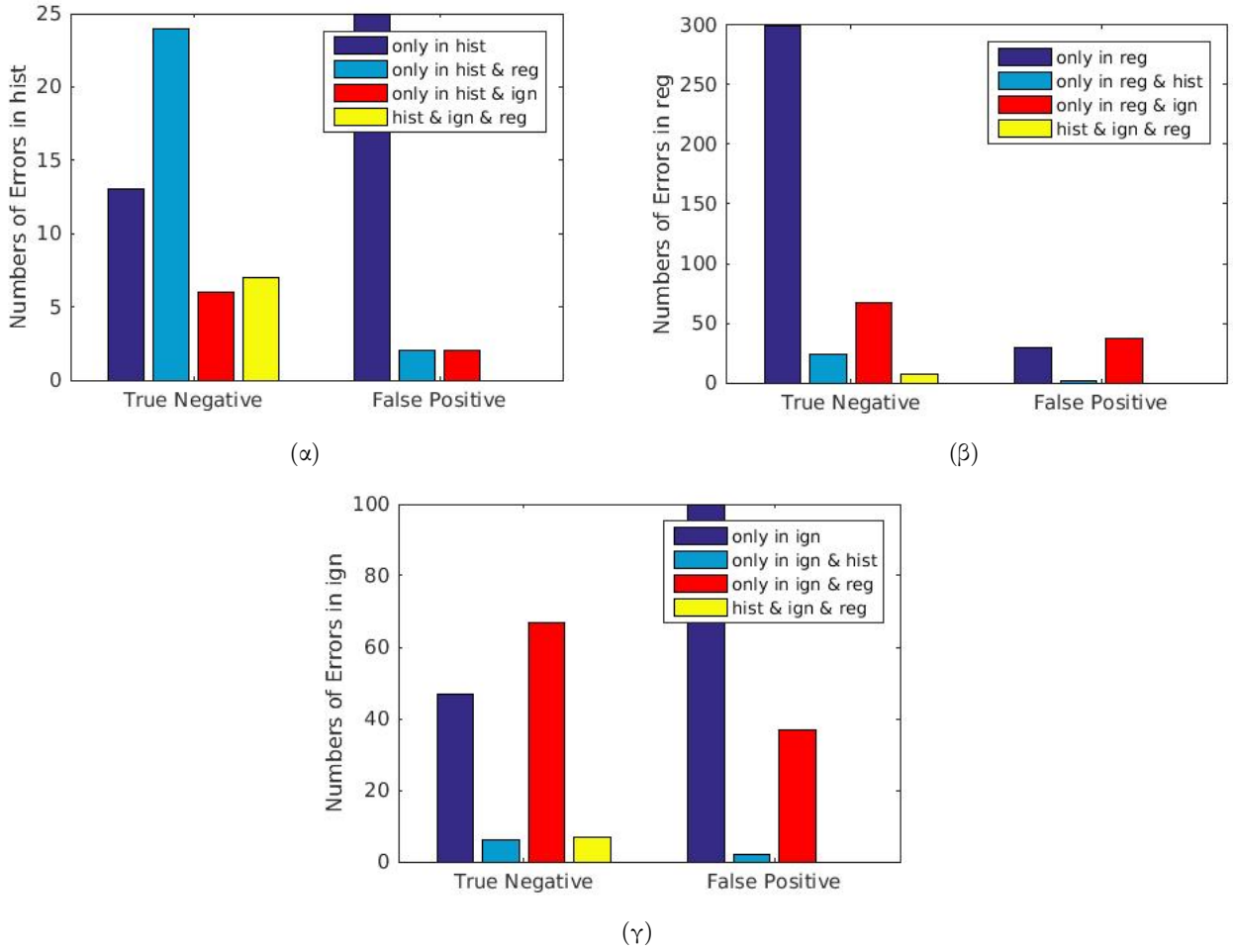
Από τις 3 διαφορετικές προσεγγίσεις που παρουσιάστηκαν στο κεφάλαιο αυτό, τα καλύτερα αποτελέσματα τα πετυχαίνει η μέθοδος ιστογραμμάτων. Ανάμεσα στις άλλες δύο μεθόδους, καλύτερα αποτελέσματα πετυχαίνουμε με τις μεθόδους υπολειπόμενης εικόνας.

Παρόλο όμως που η μέθοδο ιστογραμμάτων πετυχαίνει την καλύτερη επίδοση, είναι αρκετά ευαίσθητη στο σύνολο εικόνων στο οποίο θα εκπαιδευτεί. Καθώς όπως δείχνουμε στο Κεφάλαιο 5.3 τα αποτελέσματά της μεταβάλλονται σημαντικά ανάλογα με την ποσότητα του θορύβου που εισάγεται στις εικόνες καθώς και από την διαδικασία με την οποία έχουν παραχθεί οι επιθετικές εικόνες τις οποίες καλείται να ανιχνεύσει. Αντίθετα στις άλλες δύο μεθόδους τα αποτελέσματα δεν εξαρτώνται σε τόσο μεγάλο βαθμό από την μέθοδο με την οποία έχουμε παράξει τις επιθετικές εικόνες. Συγκεκριμένα η ομαλοποίηση του διανύσματος χαρακτηριστικών βασίζεται στην 'οπτική' απόσταση μεταξύ των διαφορετικών εικόνων, και παρόλο που εξαρτάται από τον τρόπο με τον οποίο ορίζεται αυτή, οι σχετικές αποστάσεις μεταξύ πραγματικών και επιθετικών εικόνων δεν εξαρτώνται σε μεγάλο βαθμό από την μέθοδο κατασκευής των επιθετικών εικόνων. Επίσης όπως αναφέρουμε και στο Κεφάλαιο 5.4.3 οι μέθοδοι υπολειπόμενης εικόνας έχουν την ίδια επίδοση τόσο για επιθετικές εικόνες που παράχθηκαν από την μέθοδο FGSM όσο και για εικόνες που παράχθηκαν από την μέθοδο DeepFool .

Επίσης έχει ενδιαφέρον να μελετήσουμε το ποσοστό των λάθος αναγνωρίσεων που είναι κοινές ανάμεσα στις μεθόδους. Έτσι εκτελούμε και τις 3 μεθόδους στο ίδιο σύνολο 1000 επιθετικών και 1000 πραγματικών εικόνων και καταγράφουμε τα κοινά λάθη ανάμεσα στις μεθόδους. Τα αποτελέσματα εμφανίζονται στο Σχήμα 5.22.

Στο σχήμα αυτό βλέπουμε ότι και οι 3 μέθοδοι εμφανίζουν ένα ποσοστό από λάθη τα οποία είναι μοναδικά σε αυτές. Επίσης το ποσοστό των επιθετικών εικόνων το οποίο δεν αναγνωρίζεται από καμία από τις 3 μεθόδους είναι πολύ μικρό (0.7%), ενώ ο αριθμός των πραγματικών εικόνων που αναγνωρίζονται λανθασμένα από όλες τις μεθόδους είναι μηδενικός. Ένα ακόμα αποτέλεσμα που εμφανίζει ενδιαφέρον είναι το ότι παρόλο που η μέθοδος ιστογραμμάτων (hist) έχει τον μικρότερο αριθμό από πραγματικές εικόνες τις οποίες αναγνωρίζει εσφαλμένα ως επιθετικές, οι εικόνες αυτές σε μεγάλη πλειοψηφία αναγνωρίζονται σωστά από τις άλλες μεθόδους. Αντίθετα για τις άλλες δύο μεθόδους βλέπουμε ότι υπάρχει μεγάλος αριθμός από πραγματικές εικόνες οι οποίες αναγνωρίζονται εσφαλμένα και από τις δύο μεθόδους.

Καθώς λοιπόν ένα ποσοστό των λαθών των μεθόδων γίνονται σε εικόνες όπου οι άλλες δύο μέθοδοι δεν οδηγούνται σε λάθη, επιθυμούμε να διερευνήσουμε αν συνδυάζοντας τα αποτελέσματα και των τριών μεθόδων μπορούμε να βελτιώσουμε το τελικό αποτέλεσμα αναγνώρισης. Για την μέθοδο ομαλοποίησης συμβολίζουμε με $R(\boldsymbol{x})$ το αποτέλεσμα της μεθόδου



Σχήμα 5.22: Κατανομή των κοινών λαθών ανάμεσα στις τρεις προτεινόμενες μεθόδους : (α) Κατανομή των λαθών της μεθόδου ιστογραμμάτων (hist) (β) Κατανομή των λαθών της μεθόδου ομαλοποίησης(reg) (γ) Κατανομή των λαθών της μεθόδου υπολειπόμενης εικόνας (ign)

για την εικόνα $\boldsymbol{x}$, το οποίο λαμβάνει την boolean τιμή 1 αν η εικόνα αναγνωρίζεται ως επιθετική. Αντίστοιχα συμβολίζουμε με $H(\boldsymbol{x})$ το αποτέλεσμα της μεθόδου ιστογραμμάτων , με $I(\boldsymbol{x})$ το αποτέλεσμα της μεθόδου υπολειπόμενης εικόνας και με $T(\boldsymbol{x})$ την τελική απόφαση σχετικά με το αν η εικόνα $\boldsymbol{x}$ είναι επιθετική. Επίσης συμβολίζουμε με $\wedge$ το λογικό AND και με $\vee$ το λογικό OR. Αρχικά αν αναγνωρίζουμε μια εικόνα ως επιθετική μόνο όταν αναγνωρίζεται ως επιθετική και από τις 3 μεθόδους, δηλαδή $T(\boldsymbol{x}) = (R(\boldsymbol{x}) \wedge H(\boldsymbol{x}) \wedge I(\boldsymbol{x}))$, τότε αναμένουμε να πετύχουμε το μεγαλύτερο δυνατό Precision, αλλά παράλληλα το ποσοστό του Recall να γίνει μικρότερο από το μικρότερο ποσοστό των τριών μεθόδων ξεχωριστά. Αντίθετα αν αναγνωρίζουμε μια εικόνα ως επιθετική όταν αναγνωρίζεται ως επιθετική τουλάχιστον από μια μέθοδο, δηλαδή $T(\boldsymbol{x}) = (R(\boldsymbol{x}) \vee H(\boldsymbol{x}) \vee I(\boldsymbol{x}))$, τότε αναμένουμε να πετύχουμε το μεγαλύτερο δυνατό Recall αλλά παράλληλα να μειώσουμε σημαντικά το Precision. Τέλος αν αναγνωρίζουμε ως επιθετική μια εικόνα που αναγνωρίζεται τουλάχιστον από 2 μεθόδους ως επιθετική, δηλαδή $T(\boldsymbol{x}) = ((H(\boldsymbol{x}) \wedge R(\boldsymbol{x})) \vee (H(\boldsymbol{x}) \wedge I(\boldsymbol{x})) \vee (R(\boldsymbol{x}) \wedge I(\boldsymbol{x}))$, τότε αναμένουμε να λάβουμε ενδιάμεσα αποτελέσματα και για τα δύο ποσοστά που μελετάμε.

Προκειμένου να αυξήσουμε το Precision,χωρίς να μειώσουμε σημαντικά το Recall μπορούμε να εκμεταλλευτούμε την παρατήρηση ότι οι πραγματικές εικόνες οι οποίες αναγνωρίζονται

ως επιθετικές από την μέθοδο ιστογραμμάτων, σε μεγάλο ποσοστό αναγνωρίζονται σωστά από τις δύο άλλες μεθόδους. Έτσι αποφασίζουμε ότι μια εικόνα είναι επιθετική αν αναγνωρίζεται επιθετική από την μέθοδο ιστογραμμάτων και τουλάχιστον από μια από τις άλλες δύο μεθόδους , δηλαδή $T(\boldsymbol{x}) = (H(\boldsymbol{x}) \wedge (R(\boldsymbol{x}) \vee I(\boldsymbol{x})))$

Αποτελέσματα συνδυασμού μεθόδων		
T	Precision	Recall
$R \wedge H \wedge I$	100%	53.7%
$R \vee H \vee I$	83.6%	99.3%
$((R \wedge H) \vee (R \wedge I) \vee (H \wedge I))$	95.6%	89.6%
$H \wedge (R \vee I)$	99.5%	88.3%

Πίνακας 5.4: Αποτελέσματα από διαφορετικούς συνδυασμούς των αποτελεσμάτων των μεθόδων του Κεφαλαίου 5

Στον Πίνακα 5.4 βλέπουμε τα αποτελέσματα που προκύπτουν από τους διάφορους συνδυασμούς των μεθόδων. Παρατηρούμε ότι αν απαιτούμε όλες οι μέθοδοι να αναγνωρίζουν την εικόνα ως επιθετική για να αποφασίσουμε ότι είναι επιθετική πετυχαίνουμε 100% Precision αλλά το Recall πέφτει στο 53.7%. Παρόλο που στην συγκεκριμένη περίπτωση το Recall ποσοστό είναι πολύ μικρό, για μια εφαρμογή η οποία δέχεται σαν είσοδο σε μεγάλη πλειοψηφία πραγματικές εικόνες τα ποσοστά αυτά μπορεί να είναι ικανοποιητικά. Αντίστοιχα όταν $T(x) = (R(\boldsymbol{x}) \vee H(\boldsymbol{x}) \vee I(\boldsymbol{x}))$, το αυξημένο Recall ποσοστό μπορεί να είναι χρήσιμο σε μια εφαρμογή όπου η αναγνώριση των επιθετικών εικόνων είναι κρίσιμη. Τέλος βλέπουμε ότι με την $T(x) = (H(\boldsymbol{x}) \wedge (R(\boldsymbol{x}) \vee I(\boldsymbol{x})))$ καταφέρνουμε πράγματι να πετύχουμε πολύ μεγάλο Precision χωρίς όμως να μειώσουμε το ποσοστό του Recall στο βαθμό που το μειώσαμε με τον πρώτο συνδυασμό των μεθόδων.

Κεφάλαιο 6

Συμπεράσματα και Συμβολές της Εργασίας

6.1 Συμπεράσματα και Συμβολές

Καταλήγοντας , με την εργασία αυτή διερευνήσαμε μια βασική αδυναμία την οποία παρουσιάζουν οι παραδοσιακές αρχιτεκτονικές και μέθοδοι εκπαίδευσης των νευρωνικών δικτύων, τα επιθετικά παραδείγματα. Αρχικά συσχετίσαμε την συμπεριφορά ενός νευρωνικού δικτύου ενάντια σε επιθετικά παραδείγματα με την σταθερά Lipschitz των συναρτήσεων εξόδου του και αναλύσαμε βασικά χαρακτηριστικά των επιθετικών παραδειγμάτων και των μεθόδων άμυνας ενάντια σε αυτά. Στην συνέχεια προτείναμε τρεις μεθόδους εντοπισμού επιθετικών παραδειγμάτων, κάθε μια από τις οποίες προσεγγίζει το πρόβλημα από διαφορετική κατεύθυνση και δείξαμε ότι ο συνδυασμός των μεθόδων αυτών μεταξύ τους ή πιθανόν και με άλλες μεθόδους που παρουσιάζονται στην βιβλιογραφία μπορεί να οδηγήσει σε βελτίωση των αποτελεσμάτων εντοπισμού επιθετικών παραδειγμάτων.

Συγκεκριμένα στην εργασία αυτή:

- Δείξαμε ότι μικρότερη σταθερά Lipschitz σε ένα νευρωνικό δίκτυο μπορεί να το οδηγήσει σε καλύτερη συμπεριφορά ενάντια σε επιθετικά παραδείγματα. Επίσης αφού αναπτύξαμε μια μέθοδο για υπολογισμό μιας σταθεράς Lipschitz του νευρωνικού δικτύου , δείξαμε ότι η διαδικασία του adversarial training οδηγεί στην εκπαίδευση ενός δικτύου με μικρότερη σταθερά Lipschitz από την σταθερά ενός δικτύου που έχει εκπαιδευτεί χρησιμοποιώντας το αρχικό σύνολο εκπαίδευσης.

- Αναλύσαμε την εικόνα που προστίθεται σε μια πραγματική εικόνα από την Fast Gradient Sign Method (FGSM) για να παραχθεί μια επιθετική εικόνα, και προτείναμε την τυχαία επιθετική κατεύθυνση με την οποία προσεγγίζουμε την εικόνα που προστίθεται με κατάλληλο συνδυασμό τυχαίων προτύπων του συνόλου εκπαίδευσης.

- Δείξαμε ότι όταν ενισχύσουμε το σύνολο εκπαίδευσης με πρότυπα στα οποία έχουν προστεθεί τυχαίες επιθετικές κατευθύνσεις, το εκπαιδευμένο δίκτυο έχει καλύτερη συμπεριφορά ενάντια σε επιθετικές εικόνες, χωρίς όμως να πετυχαίνουμε τα ίδια αποτελέσματα που πετυχαίνουν παρόμοιες μέθοδοι που προτείνονται στην βιβλιογραφία.

- Εξετάσαμε τον τρόπο με τον οποίο διαφορετικές μέθοδοι άμυνας επηρεάζουν την ικανότητα των επιθετικών εικόνων να επεκτείνονται σε διαφορετικά νευρωνικά δίκτυα, και καταλήξαμε στο ότι όλες οι μέθοδοι, σε διαφορετικό βαθμό η κάθε μια, πετυχαίνουν τον

περιορισμό του αριθμού των επιθετικών εικόνων οι οποίες επεκτείνονται σε διαφορετικά νευρωνικά δίκτυα.

- Προτείναμε την μέθοδο ομαλοποίηση για τον εντοπισμό επιθετικών εικόνων. Στην μέθοδο αυτή εκτελούμε ομαλοποίηση του διανύσματος χαρακτηριστικών που προκύπτει από το προτελευταίο επίπεδο του δικτύου και εκπαιδεύουμε το τελευταίο επίπεδο έτσι ώστε να αντιστοιχίζει το ομαλοποιημένο διάνυσμα στην σωστή κατηγορία τόσο για πραγματικές όσο και για επιθετικές εικόνες.

- Προτείναμε την μέθοδο ιστογραμμάτων για τον εντοπισμό επιθετικών εικόνων. Στην μέθοδο αυτή χρησιμοποιούμε τα ιστογράμματα των τιμών των εξόδων των ενδιάμεσων επιπέδων του δικτύου για να δημιουργήσουμε ένα διάνυσμα χαρακτηριστικών με βάση το οποίο ένας SVM ταξινομητής αποφασίζει σχετικά με το είδος της εισόδου.

- Παρουσιάσαμε την έννοια της υπολειπόμενης εικόνας η οποία περιέχει πληροφορία σχετικά με τα μέρη του προτύπου που αγνοούνται από τον νευρωνικό δίκτυο. Στην συνέχεια χρησιμοποιήσαμε την υπολειπόμενη εικόνα για να υλοποιήσουμε ένα σύστημα εντοπισμού επιθετικών εικόνων που βασίζεται στην μεταβολή της αρχική εικόνας έτσι ώστε να ελαχιστοποιηθεί το μέρος του αρχικού προτύπου που αγνοείται.

- Αφού συγκρίναμε τα αποτελέσματα των τριών μεθόδων, καταλήξαμε στο ότι η μέθοδος ιστογραμμάτων πετυχαίνει τα καλύτερα αποτελέσματα μεταξύ των μεθόδων αυτών. Παράλληλα παρουσιάσαμε την δυνατότητα συνδυασμού των τριών μεθόδων με την οποία επιτυγχάνεται βελτίωση των συνολικών αποτελεσμάτων εντοπισμού.

6.2 Κατευθύνσεις για μελλοντική έρευνα

Οι μέθοδοι εντοπισμού επιθετικών εικόνων που αναπτύσσονται στην εργασία αυτή αποτελούν προτάσεις για την επίλυση του προβλήματος οι οποίες εμφανίζουν αρκετά υποσχόμενα ποσοστά εντοπισμού . Ορισμένες κατευθύνσεις που προκύπτουν για μελλοντική έρευνα και περαιτέρω ανάπτυξη των μεθόδων που παρουσιάζονται είναι οι παρακάτω:

- Για την μέθοδο των ιστογραμμάτων του κεφαλαίου 5.3, έχει ενδιαφέρον να αναπτυχθεί ένας συστηματικός τρόπος επιλογής των επιπέδων του δικτύου των οποίων οι έξοδοι χρησιμοποιούνται για την εξαγωγή των ιστογραμμάτων, καθώς στην παρούσα εργασία η επιλογή των επιπέδων έγινε μετά από πειραματικές δοκιμές.

- Στο κεφάλαιο 5.4.1 αναπτύχθηκε η έννοια της υπολειπόμενης εικόνας η οποία στην συνέχεια χρησιμοποιήθηκε στο κεφάλαιο 5.4.2 για την κατασκευή μεθόδων για τον εντοπισμό επιθετικών εικόνων. Οι μέθοδοι αυτές στοχεύουν στην εισαγωγή της πληροφορίας της υπολειπόμενης εικόνας στο νευρωνικό δίκτυο, στόχος που στην εργασία αυτή επιτυγχάνεται με τις εξισώσεις (5.15),(5.20). Όμως όπως αναφέρουμε και στο αντίστοιχο κεφάλαιο, εκτός από την μεταβολή της εικόνας εισόδου μπορούμε να πετύχουμε την μείωση του μέτρου της υπολειπόμενης εικόνας με μεταβολή των παραμέτρων του δικτύου. Έτσι έχει ενδιαφέρον η περαιτέρω διερεύνηση τεχνικών με τις οποίες μπορεί να χρησιμοποιηθεί η πληροφορία της υπολειπόμενης εικόνας για την βελτίωση των αποτελεσμάτων των νευρωνικών δικτύων τόσο με την μεταβολή της εισόδου του δικτύου όσο και με την μεταβολή των παραμέτρων του δικτύου.

Βιβλιογραφία

[1] H. Anderson, J. Woodbridge and B. Filar, "DeepDGA: Adversarially-Tuned Domain Generation and Detection", in *Proceedings of the 2016 ACM Workshop on Artificial Intelligence and Security*, 2016, pp. 13-21.

[2] J. Ba and R. Caruana, "Do Deep Nets Really Need to be Deep?", in *Proceedings of the 27th International Conference on Neural Information Processing Systems*, 2014, pp. 2654-2662.

[3] C. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, New York, Springer, 2006.

[4] N. Carlini and D. Wagner, "Towards Evaluating the Robustness of Neural Networks", in *Proceedings of 2017 IEEE Symposium on Security and Privacy*, 2017, pp. 39-57.

[5] P. Chen, H. Zhang, Y. Sharma, J. Yi and C. Hsieh, "ZOO: Zeroth Order Optimization Based Black-box Attacks to Deep Neural Networks without Training Substitute Models", in *Proceedings of the 2017 ACM Workshop on Artificial Intelligence and Security*, 2017, pp. 15-26.

[6] M. Cisse, P. Bojanowski, E. Grave, Y. Dauphin and N. Usunier, "Parseval Networks: Improving Robustness to Adversarial Examples", in *Proceedings of International Conference on Machine Learning*, 2017, pp. 854-863.

[7] J. Dai, Y. Lu and Y. Wu, "Generative modeling of convolutional neural networks", *Statistics and Its Interface*, vol. 9, no. 4, pp. 485-496, 2016.

[8] A. Elmoataz, O. Lezoray and S. Bougleux, "Nonlocal Discrete Regularization on Weighted Graphs: A Framework for Image and Manifold Processing", *IEEE Transactions on Image Processing*, vol. 17, no. 7, pp. 1047-1060, 2008.

[9] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno and D. Song, "Robust Physical-World Attacks on Deep Learning Visual Classification", in *Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* , 2018, pp. 1625-1634.

[10] R. Feinman, R. Curtin, S. Shintre and A. Gardner, "Detecting Adversarial Samples from Artifacts", *arXiv preprint arXiv:1703.00410*, 2017.

[11] I. Goodfellow, J. Shlens and C. Szegedy, "Explaining and Harnessing Adversarial Examples", *arXiv preprint arXiv:1412.6572*, 2014.

[12] D. Gopinath, G. Katz, C. Pasareanu and C. Barret, "DeepSafe: A Data-driven Approach for Checking Adversarial Robustness in Neural Networks", *arXiv preprint arXiv:1710.00486*, 2017.

[13] K. Grosse, N. Papernot, P. Manoharan, M. Backes and P. McDanieland, "Adversarial Examples for Malware Detection", in *Proceedings of European Symposium on Research in Computer Security*, 2017, pp. 62-79.

[14] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition", in *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770-778.

[15] G.Hinton, N.Srivastava, A.Krizhevsky, I.Sutskever and R.Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors.", *arXiv preprint arXiv:1207.0580*, 2012.

[16] G. Katz, C. Barrett, D. Dill, K. Julian and M. Kochenderfer, "Reluplex: An Efficient SMT Solver for Verifying Deep Neural Networks", in *Proceedings of the 29th International Conference on Computer-Aided Verification*, 2017, pp. 97-117.

[17] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images", Computer Science Department, University of Toronto, Technical Report, 2009.

[18] A. Krizhevsky, I. Sutskever and G. Hinton, "ImageNet classification with deep convolutional neural networks", *Communications of the ACM*, vol. 60, no. 6, pp. 84-90, 2017.

[19] A. Kurakin, I. Goodfellow and S. Bengio, "Adversarial examples in the physical world", *arXiv preprint arXiv:1607.02533*, 2016.

[20] Y. LeCun, Y. Bengio and G. Hinton, "Deep learning", *Nature*, vol. 521, no. 7553, pp. 436-444, 2015.

[21] Y. LeCun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition", *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278-2324, 1998.

[22] J. Lu, T. Issaranon and D. Forsyth, "SafetyNet: Detecting and Rejecting Adversarial Examples Robustly", in *Proceedings of International Conference on Computer Vision*, 2017, pp. 446-454.

[23] S. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674-693, 1989.

[24] S. Mallat, "Group Invariant Scattering", *Communications on Pure and Applied Mathematics*, vol. 65, no. 10, pp. 1331-1398, 2012.

[25] S. Mallat, "Understanding deep convolutional networks", *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, 2016.

[26] D. Meng and H. Chen, "MagNet: A Two-Pronged Defense against Adversarial Examples", in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017, pp. 135-147.

[27] J. Metzen, T. Genewein, V. Fisher and B. Bischoff, "On Detecting Adversarial Perturbations", in *Proceedings of the 5th International Conference on Learning Representations*, 2017.

[28] J. Metzen, M. Kumar, T. Brox and V. Fischer, "Universal Adversarial Perturbations Against Semantic Image Segmentation", in *Proceedings of International Conference on Computer Vision*, 2017, pp. 2774-2783.

[29] S. Moosavi-Dezfooli, A. Fawzi, O. Fawzi and P. Frossard, "Universal Adversarial Perturbations", in *Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 86-94.

[30] S. Moosavi-Dezfooli, A. Fawzi and P. Frossard, "DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks", in *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2574–2582.

[31] N. Narodytska and S. Kasiviswanathan, "Simple Black-Box Adversarial Attacks on Deep Neural Networks", in *Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 1310 - 1318.

[32] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, B. Celik and A. Swami, "Practical Black-Box Attacks against Machine Learning", in *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, 2017, pp. 506-519.

[33] N. Papernot, P. McDaniel, X. Wu, S. Jha and A. Swami, "Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks", in *Proceedings of 2016 IEEE Symposium on Security and Privacy*, 2016, pp. 582-597.

[34] A. Rozsa, E. Rudd and T. Boult, "Adversarial Diversity and Hard Positive Generation", in *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 410 - 417.

[35] T.Salimans, A.Karpathy, X.Chen and D.Kingma, "PixelCNN++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications.", in *Proceedings of the 5th International Conference on Learning Representations*, 2017.

[36] M. Sharif, S. Bhagavatula, L. Bauer and M. Reiter, "Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition", in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016, pp. 1528-1540.

[37] P. Simard, D. Steinkraus and J. Platt, "Best practices for convolutional neural networks applied to visual document analysis", in *Proceedings of International Conference on Document Analysis and Recognition*, 2003, pp. 958 - 963.

[38] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition", *arXiv preprint arXiv:1409.1556*, 2014.

[39] Y. Song, T. Kim, S. Nowozin, S. Ermon and N. Kushman, "PixelDefend: Leveraging Generative Models to Understand and Defend against Adversarial Examples.", in *Proceedings of the 6th International Conference on Learning Representations*, 2018.

[40] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow and R. Fergus, "Intriguing properties of neural networks", *arXiv preprint arXiv:1312.6199*, 2013.

[41] Z. Wang, A. Bovik, H. Sheikh and E. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity", *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600-612, 2004.

[42] T. Wiatowski, M. Tschannen, A. Stanic, P. Grohs and H. Bölcskei, "Discrete Deep Feature Extraction: A Theory and New Architectures", in *Proceedings of International Conference on Machine Learning*, 2016, pp. 2149-2158.

[43] C. Xie, J. Wang, Z. Zhang, Y. Zhou, L. Xie and A. Yuille, "Adversarial Examples for Semantic Segmentation and Object Detection", in *Proceedings of International Conference on Computer Vision*, 2017, pp. 1378-1387.

[44] H. Zhang, M. Cisse, Y. Dauphin and D. Lopez-Paz, "Mixup: Beyond Empirical Risk Minimization", in *Proceedings of the 6th International Conference on Learning Representations*, 2018.