Ανακατασκευή οπτικών ερεθισμάτων από fMRI με την χρήση cDCGAN

[*]Εργασία στο μάθημα «Ανάλυση Βιο-δεδομένων»

Ευάγγελος Τσόγκας
Τμήμα Η.Μ.Μ.Υ
Εθνικό Μετσόβιο Πολυτεχνείο
Αθήνα, Ελλάδα
ΑΜ: 03400120
evaggelostsogkas@mail.ntua.gr

Φίλιππος Μαυρεπής
Τμήμα Η.Μ.Μ.Υ
Εθνικό Μετσόβιο Πολυτεχνείο
Αθήνα, Ελλάδα
ΑΜ: 03400098
philipposmavrepis@mail.ntua.gr

Κωνσταντίνος Βιλουράς
Τμήμα Η.Μ.Μ.Υ
Εθνικό Μετσόβιο Πολυτεχνείο
Αθήνα, Ελλάδα
ΑΜ: 03400086
konstantinosvilouras@mail.ntua.gr

Γεώργιος Πιπιλής
Τμήμα Η.Μ.Μ.Υ
Εθνικό Μετσόβιο Πολυτεχνείο
Αθήνα, Ελλάδα
ΑΜ: 03400117
georgiospipilis@mail.ntua.gr

***Περίληψη*—Η ανακατασκευή οπτικών ερεθισμάτων από δεδομένα fMRI είναι ένα από τα δυσκολότερα και πιο ενδιαφέροντα προβλήματα στην τομή των επιστημών της νευροβιολογίας και της βαθιάς μάθησης. Σε αυτή την εργασία, αναλύουμε τις υπάρχουσες προσεγγίσεις και τα αποτελέσματα τους. Επιπλέον, προτείνουμε ένα δικό μας μοντέλο το οποίο συναγωνίζεται, και ξεπερνά σε κάποιες περιπτώσεις, τα αποτελέσματα των μοντέρνων ερευνητικών προσεγγίσεων. Τέλος, ερευνούμε την πιστότητα των αποτελεσμάτων ορισμένων αρχιτεκτονικών και προτείνουμε μελλοντικές κατευθύνσεις που μπορεί να ακολουθήσει η έρευνα στο συγκεκριμένο τομέα.**

Λέξεις Κλειδιά - fMRI, Deep Learning, Brain, Generative Adversarial Networks, Autoencoders

I. Εισαγωγή

Η αποκωδικοποίηση του τρόπου με τον οποίο ο εγκέφαλός μας επεξεργάζεται τα οπτικά ερεθίσματα, αλλά και η ανακατασκευή των εν λόγω ερεθισμάτων είναι δύο από τις πιο ελκυστικές προκλήσεις στο χώρο της νευροαπεικόνισης και της νευροβιολογίας. Ενώ πριν από μερικές δεκαετίες αυτοί οι στόχοι μπορεί να έμοιαζαν ως σενάρια επιστημονικής φαντασίας, η ραγδαία ανάπτυξη της βαθιάς μάθησης (Deep Learning) και η εφαρμογή της στην ανακατασκευή εικόνων από δεδομένα Λειτουργικής Μαγνητικής Τομογραφίας (functional MRI ή fMRI) τους καθιστά άμεσα εφικτούς και, ταυτόχρονα, μας φέρνει ένα βήμα πιο κοντά στο να κατανοήσουμε τους μηχανισμούς όρασης και οπτικής επεξεργασίας του εγκεφάλου.

Ο στόχος της παρούσας εργασίας είναι πολυεπίπεδος. Αρχικά, ασχολούμαστε με την εξερεύνηση διαφόρων τεχνικών Deep Learning που εμφανίζονται στη μοντέρνα βιβλιογραφία, αλλά και την ανάλυση της αποτελεσματικότητάς τους σε εφαρμογές ανακατασκευής από δεδομένα fMRI (fMRI reconstruction). Πρακτικά, αφού ο ασθενής δει ένα σύνολο από εικόνες κατά τη διάρκεια μιας εξέτασης fMRI, στοχεύουμε στην ανακατασκευή των αρχικών εικόνων χρησιμοποιώντας τα αποτελέσματα της εξέτασης.

Ταυτόχρονα, στοχεύουμε στον πειραματισμό με νέες τεχνικές και σύνολα δεδομένων, αλλά και στην επέκταση μοντέλων που ήδη χρησιμοποιούνται ευρέως στους ερευνητικούς κύκλους. Επιπροσθέτως, αναλύουμε την επίδραση συγκεκριμένων παραμέτρων (όπως ο αριθμός από voxels που χρησιμοποιεί το μοντέλο σε κάθε επανάληψη) στην αποτελεσματικότητα των διαφόρων αρχιτεκτονικών. Τέλος, προσεγγίζουμε με κριτική σκέψη ορισμένα από τα μοντέλα που έχουν προταθεί τα τελευταία χρόνια και τα εξετάζουμε από τη σκοπιά ενός μηχανικού βαθιάς μάθησης ως προς την ορθότητα, αλλά και την αποτελεσματικότητά τους.

Η βασική μέθοδος στην οποία επικεντρωνόμαστε εμείς, αλλά και ένα πολύ μεγάλο κομμάτι της ερευνητικής κοινότητας, αφορά τα Generative Adversarial Networks (GANs), μια κατηγορία αρχιτεκτονικών βαθιάς μάθησης που προτάθηκε από τον I. Goodfellow [1] το 2014. Η μέθοδος αυτή αποσκοπεί στην έμμεση εκπαίδευση ενός δικτύου (το οποίο ονομάζουμε generator) μέσα από τη συμμετοχή του σε ένα παίγνιο μηδενικού αθροίσματος (zero sum game) ενάντια σε ένα δίκτυο κριτή (το οποίο ονομάζουμε discriminator). Η μεγάλη διαφορά με άλλες αρχιτεκτονικές παραγωγής εικόνας είναι πως, στην περίπτωση των GANs, το δίκτυο που εκπαιδεύουμε (ο generator) δεν αποσκοπεί στο να ελαχιστοποιήσει κάποια μετρική σφάλματος σε σχέση με την εικόνα στόχο, αλλά στο να παραπλανήσει τον discriminator ώστε αυτός να μη μπορεί να κρίνει πιθανοτικά εάν η εικόνα

που δημιουργήθηκε από τον generator είναι πραγματική ή τεχνητή.

Αντίστοιχα, όπως αναφέρθηκε παραπάνω, η απεικονιστική μέθοδος στην οποία επικεντρωνόμαστε είναι το fMRI. Οι νευρώνες του εγκεφάλου δε διαθέτουν εσωτερικές πηγές ενέργειας και, συνεπώς, βασίζονται στο αίμα για την οξυγόνωση τους. Τα σήματα του fMRI, μέσω μιας τεχνικής γνωστής ως Blood Oxygen Level Dependent (BOLD) imaging, μπορούν να καταγράψουν τη σχετική δραστηριότητα των περιοχών του εγκεφάλου μέσω της παρατήρησης των μεταβολών των επιπέδων οξυγόνου των νευρώνων. Με αυτό τον τρόπο, η χρήση του fMRI μας επιτρέπει να δημιουργήσουμε μια συσχέτιση μεταξύ της εγκεφαλικής δραστηριότητας και του οπτικού ερεθίσματος στο οποίο έχει εκτεθεί ο ασθενής και, συνεπώς, να πάρουμε μια ιδέα για το πώς αντιδρά ο ανθρώπινος νους σε ορισμένα ερεθίσματα.

Ο κλινικός αντίκτυπος είναι άμεσα ορατός, καθώς η επίλυση του συγκεκριμένου προβλήματος θα οδηγήσει στη δημιουργία πρωτοποριακών εφαρμογών στον κλάδο της βιοϊατρικής και της νευροεπιστήμης. Συγκεκριμένα, η δυνατότητα αποκωδικοποίησης των λειτουργιών του ανθρώπινου εγκεφάλου θα βοηθήσει στη βαθύτερη κατανόηση της ανθρώπινης όρασης και κατ΄ επέκταση της ανθρώπινης αντίληψης γενικότερα, γεγονός το οποίο θα δημιουργήσει αλυσιδωτές αντιδράσεις σε πληθώρα επιστημονικών κλάδων, όπως για παράδειγμα το deep learning. Τέλος, μια χαρακτηριστική εφαρμογή που μπορεί να προκύψει αφορά το “mind reading", δηλαδή τη δυνατότητα να διαβάζουμε τις σκέψεις ενός ανθρώπου σε πραγματικό χρόνο. Κάτι τέτοιο θα μπορούσε να συνεισφέρει στη διάγνωση νευροεκφυλιστικών ασθενειών ή άλλων αντίστοιχων παθήσεων, ωστόσο είναι εύλογο να υπάρχουν και κατάλληλα μέτρα ασφαλείας τα οποία θα προστατεύουν τους χρήστες-ασθενείς από κακόβουλες πηγές.

Η δομή του υπολοίπου της εργασίας περιγράφεται παρακάτω. Στην δεύτερη ενότητα (II) αναφέρονται προγενέστερες προσπάθειες που έχουν επιχειρηθεί για την επίλυση αυτού του προβλήματος μαζί με τα σημαντικότερα αποτελέσματά τους. Στην ενότητα τρία (III) περιγράφεται η φύση και τα χαρακτηριστικά των δεδομένων που χρησιμοποιήθηκαν στη φάση της ανάλυσης και της ανάπτυξης. Έπειτα, στην ενότητα τέσσερα (IV), εξηγούνται οι μέθοδοι και τεχνικές που επιστρατεύτηκαν για την ανάπτυξη του προτεινόμενου μοντέλου καθώς και η γενική διατύπωση του προβλήματος στα πλαίσια των GANs. Ύστερα, φαίνονται συγκεντρωτικά στην ενότητα πέντε (V) οι μετρικές που χρησιμοποιήθηκαν για την αξιολόγηση των αποτελεσμάτων καθώς και τα αποτελέσματα καθ΄ αυτά για τα διάφορα σύνολα δεδομένων. Επιπροσθέτως, στην ενότητα έξι (VI) αναπτύσσονται υπό μορφή συζήτησης τα σύμπερασματα και πιθανές μελλοντικές επεκτάσεις της προσέγγισης που ακολουθήθηκε. Τέλος, στην ενότητα εφτά (VII) καταγράφονται οι ευχαριστίες μας για την συνδρομή της διδάσκουσας, ενώ ακολούθως στην ενότητα οχτώ (VIII) υπάρχει η πλήρης βιβλιογραφία που διερευνήθηκε.

II. Βιβλιογραφική Επισκόπιση

Η εξερεύνηση του συνδέσμου μεταξύ της εγκεφαλικής δραστηριότητας στον οπτικό φλοιό (visual cortex) και των οπτικών ερεθισμάτων είναι ένα θέμα που έχει απασχολήσει την επιστημονική κοινότητα για πολλές δεκαετίες [2]. Τα τελευταία είκοσι χρόνια, έχουν προταθεί ισχυρές συσχετίσεις μεταξύ του λειτουργικού MRI και των οπτικών ερεθισμάτων που έλαβε ο ασθενής, σε βαθμό που ορισμένοι ερευνητές, θεωρούν τις αρχικές εικόνες-ερεθίσματα ως κάτι το οποίο μπορεί να αποκωδικοποιηθεί πλήρως από τα αποτελέσματα του fMRI [3], [4].

Το πρόβλημα της αποκωδικοποίησης αυτής έχει προσεγγιστεί με πληθώρα τρόπων. Ένα μεγάλο ποσοστό από αυτούς αφορά τη χρήση γραμμικών μοντέλων με μεγάλη έμφαση στη χρήση γραμμικής παλινδρόμησης [5], [6]. Σε αυτές τις τεχνικές, οι ερευνητές αποσπούν features μέσω των Gabor wavelets [3] από τις εικόνες που έχουν χρησιμοποιηθεί ως ερεθίσματα και δημιουργούν μια γραμμική αντιστοίχιση μεταξύ αυτών και των δεδομένων του fMRI. Αυτή η προσέγγιση με τεχνικές επεξεργασίας σήματος οδηγεί σε καλή ανακατασκευή των low-level χαρακτηριστικών των εικόνων, αλλά ελλιπή αποτελεσματικότητα όταν αντιμετωπίζουμε πιο σύνθετες και πιο φυσικές εικόνες [7], [8], [9], [10].

Φυσικά, όπως συνέβη και σε πολλούς άλλους κλάδους της επιστήμης, η άνθιση του deep learning έφερε στον κλάδο της νευροβιολογίας αρχιτεκτονικές όπως τα συνελικτικά νευρωνικά δίκτυα (Convolutional Neural Networks, CNNs) οι οποίες σταδιακά αντικατέστησαν τα γραμμικά μοντέλα στις εφαρμογές ανακατασκευής fMRI [8]. Υπό αυτή την προσέγγιση, στόχος του συνελικτικού δικτύου είναι να προσαρμόσει τα βάρη των νευρώνων του, έτσι ώστε να υπάρχει όσο το δυνατόν μικρότερο σφάλμα ανάμεσα στα σήματα που παράγει το fMRI και τα features της εικόνας στα οποία επικεντρώνεται το CNN.

Επιπλέον, τα εν λόγω CNN μπορούν να είναι προεκπαιδευμένα πάνω σε ένα σύνολο από εικόνες του πραγματικού κόσμου. Το γεγονός αυτό, σε συνδυασμό με το ότι πλέον αναλύουμε την εικόνα με βάση τις περιοχές και τα features της και όχι μόνο με βάση τις τιμές του κάθε μεμονωμένου pixel (όπως κάναμε στη γραμμική παλινδρόμηση), οδηγεί σε συνολικά καλύτερα αποτελέσματα απ΄ό,τι οδηγούν τα γραμμικά μοντέλα [11]. Το συμπέρασμα αυτό ενισχύεται και απο την ανακάλυψη πως, όπως έδειξαν οι Horikawa και Kamitani [12], υπάρχει ομοιογένεια μεταξύ των ιεραρχικών αναπαραστάσεων του εγκεφάλου και των αντιστοίχων στους νευρώνες ενός deep neural network.

Πλέον, λόγω της ραγδαίας ανάπτυξης του κλάδου του deep learning, οι ερευνητές έχουν στρέψει το ενδιαφέρον τους σε μοντέλα end-to-end, τα οποία αποτελούν και το state-of-the-art του εν λόγω προβλήματος. Όλες οι σύγχρονες προσεγγίσεις μπορούν να κατηγοριοποιηθούν σε δύο ευρύτερους τομείς με βάση τον τύπο του μοντέλου που χρησιμοποιείται και σχετίζονται είτε με deep autoencoders είτε με GANs, ενώ συχνά εμφανίζεται και ο συνδυασμός των δύο αυτών τεχνικών. Στην περίπτωση των autoencoders, το

στάδιο του encoding αφορά την ένθεση της εισόδου του Encoder σε ένα νέο χώρο (latent space) με σκοπό τη δημιουργία μιας αναπαράστασης χαμηλότερης διάστασης, η οποία στη συνέχεια προωθείται στον Decoder, ο οποίος αναλαμβάνει να ανακατασκευάσει το αρχικό οπτικό ερέθισμα. Αντίθετα, στην περίπτωση των GANs, ο Generator αναλαμβάνει εξ ολοκλήρου την ανακατασκευή του ερεθίσματος, ενώ ο Discriminator πρέπει να είναι σε θέση να διακρίνει την ανακατασκευασμένη εικόνα από την πραγματική. Τέλος, σημειώνεται ότι ο ρόλος του Discriminator είναι σημαντικός μόνο κατά τη διάρκεια της εκπαίδευσης του δικτύου, οπότε δεν συμμετέχει στη διαδικασία του testing.

Συγκεκριμένα, στο [13] το τελικό μοντέλο αποτελείται από τρία μέρη: τον Encoder, τον Denoiser και το conditional GAN (cGAN). Στον Encoder εκτελείται η διαδικασία της εξαγωγής χαρακτηριστικών, ενώ παράλληλα εφαρμόζεται και ένας τύπος attention με την ονομασία feature-weighted receptive field, ο οποίος πρακτικά αναλαμβάνει να σταθμίσει τη σημαντικότητα/συνεισφορά του κάθε feature map ξεχωριστά. Στη συνέχεια, χρησιμοποιείται ένας denoising autoencoder (DAE) για τη μείωση των διαστάσεων και την αποθορυβοποίηση των fMRI, τα οποία ως γνωστό έχουν πολύ χαμηλό signal-to-noise ratio (SNR). Τέλος, το cGAN εκπαιδεύεται στο embedding που προκύπτει από την έξοδο του DAE, με αποτέλεσμα να βελτιώνεται η ευρωστία του μοντέλου ως προς το θόρυβο που εντοπίζεται στα raw fMRI σήματα.

Επιπλέον, στο [14] η ανακατασκευή γίνεται με end-to-end τρόπο μέσω ενός GAN, το οποίο εκπαιδεύεται λαμβάνοντας ως είσοδο το αρχικό σήμα fMRI. Η μόνη διαφορά είναι ότι το δίκτυο εκπαιδεύεται χρησιμοποιώντας ως loss function το συνδυασμό του adversarial loss και του feature/perceptual loss. Το adversarial loss βελτιστοποιείται έτσι ώστε η κατανομή των ανακατασκευασμένων εικόνων να πλησιάζει όσο το δυνατόν περισσότερο την κατανομή των φυσικών εικόνων, ενώ το feature loss ενθαρρύνει τον Generator να παράγει εικόνες οι οποίες διαισθητικά θα ταιριάζουν με τα αρχικά οπτικά ερεθίσματα.

Στη συνέχεια, στο [15] ορίζεται ένα δίκτυο Encoder-Decoder, όπου ο Encoder αναλαμβάνει την απεικόνιση από τον χώρο των εικόνων στον χώρο των σημάτων fMRI, ενώ ο Decoder την αντίστροφη απεικόνιση, δηλαδή από τον χώρο των fMRI πίσω στον χώρο των εικόνων. Η συγκεκριμένη μέθοδος εισάγει επιπλέον και τη δυνατότητα της εκπαίδευσης του δικτύου χωρίς επίβλεψη (unsupervised) χρησιμοποιώντας unlabeled εικόνες, ενώ επίσης δανείζεται και ορισμένες τεχνικές του zero-shot learning που αφορούν την εκπαίδευση του δικτύου χρησιμοποιώντας τα unlabeled fMRI δεδομένα κατά τη διάρκεια του testing. Η τελευταία τεχνική αποτελεί μια πρωτοβουλία των συγγραφέων έτσι ώστε να αντιμετωπίσουν το πρόβλημα της έλλειψης επαρκώς μεγάλου training set με όσο το δυνατόν πιο αποτελεσματικό τρόπο.

Στο [16] ορίζεται το μοντέλο Dual-Variational Autoencoder GAN (D-VAE/GAN), το οποίο αποτελείται από τα εξής επιμέρους modules: τον Cognitive (VAE) Encoder, ο οποίος αναλαμβάνει την απεικόνιση των σημάτων fMRI σε ένα διάνυσμα χαμηλών διαστάσεων, τον Visual (VAE) Encoder, ο οποίος εκτελεί την ίδια διαδικασία με πριν χρησιμοποιώντας τις εικόνες εισόδου, τον Generator (VAE Decoder), ο οποίος αναλαμβάνει να ανακατασκευάσει την αρχική εικόνα χρησιμοποιώντας είτε την έξοδο του Visual Encoder είτε του Cognitive Encoder και τέλος τον Discriminator, ο οποίος αναλαμβάνει να ξεχωρίσει τα πραγματικά από τα ανακατασκευασμένα οπτικά ερεθίσματα. Επιπλέον, πολύ σημαντικό ρόλο παίζει η εφαρμογή του knowledge distillation, με την έννοια ότι ο Visual Encoder καθοδηγεί τον Cognitive Encoder, έτσι ώστε να προκύψει ένα είδος αντιστοίχησης μεταξύ του χώρου των εικόνων και του χώρου των fMRI. Έτσι, ο Visual Encoder λειτουργεί και ως teacher network, καθώς προσπαθεί να μεταλαμπαδεύσει τη γνώση του στο student network, δηλαδή στον Cognitive Encoder.

Τέλος, το [17] στηρίζεται στο γεγονός ότι η πολυπλοκότητα των αναπαραστάσεων στον οπτικό φλοιό αυξάνεται με ιεραρχικό τρόπο και ορίζει επιπλέον τις διαδικασίες του shape decoding και του semantic decoding χρησιμοποιώντας τα δεδομένα fMRI από το lower visual cortex (LVC) και το higher visual cortex (HVC) αντίστοιχα. Το shape decoding αφορά την εξαγωγή του περιγράμματος των αντικειμένων και πραγματοποιείται μέσω του συνδυασμού των εξόδων τριών διαφορετικών decoders, οι οποίοι εκπαιδεύονται σε τρία ξεχωριστά regions of interest (ROIs), δηλαδή στα V1, V2, V3 αντίστοιχα. Στη συνέχεια, το semantic decoding αφορά την εξαγωγή χαρακτηριστικών υψηλού επιπέδου μέσω ενός DNN. Τέλος, όλη η παραπάνω πληροφορία προωθείται στον Generator, ο οποίος πρακτικά αποτελεί ένα U-Net με skip connections.

III. Δεδομένα

Σε αυτήν τη θεματική ενότητα θα εισάγουμε για πρώτη φορά τα σύνολα δεδομένων που χρησιμοποιήθηκαν για την πειραματισμό με την προτεινόμενη μεθοδολογία. Χρησιμοποιήσαμε τρία δημόσια σύνολα δεδομένων τα οποία αρχικά είχαν συλλεχθεί για πειράματα γύρω απο την θεματική περιοχή της αναγνώρισης ερεθισμάτων ή/και ανακατασκευή αυτών. Παρακάτω θα παρουσιαστούν συνοπτικά οι διάφορες ιδιότητες των δεδομένων, ενώ περισσότερες πληροφορίες μπορούν να βρεθούν στις αρχικές δημοσιεύσεις.

A΄. Miyawaki08: Geometric shape and alphabetical letters

Αυτό το σύνολο δεδομένων αναφέρεται ως Visual Image Reconstruction dataset [18] ή Neuro dataset. Εικόνες με πολύ υψηλή αντίθεση (contrast) μεγέθους (10,10) παρουσιάστηκαν σε δύο υποκείμενα σε δύο διαφορετικές συνεδρίες, όπου καταγράφηκαν οι αποκρίσεις τους (fMRI) μέσω ενός 3T scanner (TR, 2 s; voxel size, 3 ×3×3 mm). Μια συνεδρία είναι η ‘συνεδρία τυχαίας εικόνας’ κατά την οποία 1320 χωρικά τυχαία πρότυπα εμφανίζονται σειριακά με ενα συγκεκριμένο σημείο προσοχής (fixation point). Κάθε ‘τυχαία εικόνα’ παρουσιάζεται για 6 δευτερόλεπτα ακολουθούμενη από 6 δευτερόλεπτα παύσης. Η άλλη συνεδρία αποκαλείται ‘figure image session’ κατά την οποία, 80 παραδείγματα από

Πίνακας I
Πληροφορίες για τα σύνολα δεδομένων. Τα ROIs υποδεικνύουν τις σχετικές περιοχές ενδιαφέροντος

Dataset	Υποκείμενα	Training/Test images	Ανάλυση	ROIs
Miyawaki08 [18]	S1	670/50	(10,10)	V1,V2,V3,V4,VP
vanGerven10 [19]	S1	90/10	(28,28)	V1,V2,V3
Schoenmakers13 [20]	S1,S2,S3	290/70	(56,56)	V1,V2

γράμματα και απλά γεωμετρικά σχήματα παρουσιάζονται σεριακά. Κάθε τέτοια εικόνα παρουσιάζεται για 12 δευτερόλεπτα και ακολουθείται από 12 δευτερόλεπτα παύσης. Στην παρούσα εργασία χρησιμοποιήθηκαν μόνο οι εικόνες της δεύτερης συνεδρίας στις οποίες επιτελέστηκε ένα ειδικό train/test split. Τα δεδομένα του υποκειμένου 1 υπάρχουν διαθέσιμα online εδώ: Neuro dataset και χρησιμοποιήθηκαν fMRI δεδομένα από τις οπτικές περιοχές V1, V2, V3, V4 και VP.

Β΄. vanGerven10: handwritten digits

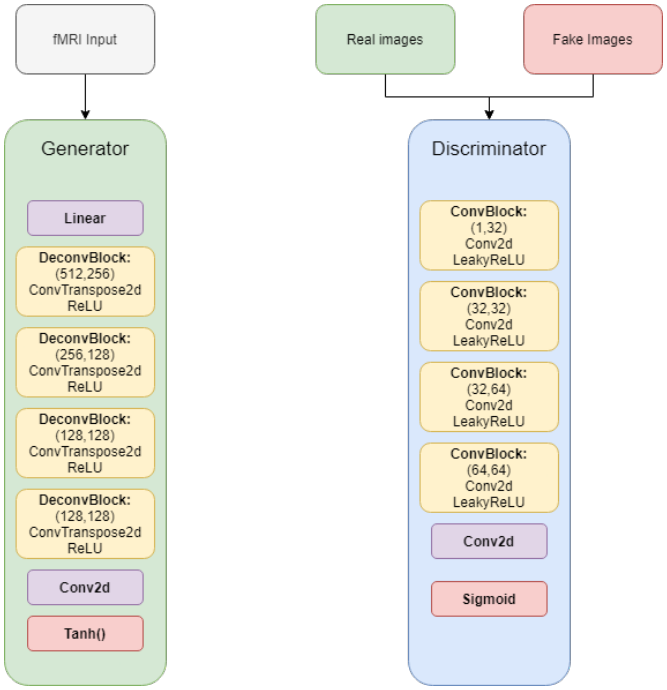
Αυτό το σύνολο δεδομένων αναφέρεται πολύ συχνά ως 69 dataset [19] το οποίο περιέχει τα δεδομένα ενός υποκειμένου όταν σε αυτό παρουσιάστηκαν 100 ασπρόμαυρες εικόνες χειρόγραφων ψηφίων (6,9) με μέγεθος (28,28). Κάθε εικόνα παρουσιαζόταν για 12.5 δευτερόλεπτα (flickering rate 6Hz) σε 3T scanner (TR, 2.5 s; voxel size, 2 ×2×2 mm). Οι εικόνες προέρχονται από το σύνολο εκπαίδευσης του MNIST και τα fMRI εξάγονται από τις περιοχές V1,V2,V3. Ακολουθώντας τα train/test split της αρχικής δημοσίευσης, 90 ζεύγη εικόνων-fMRI χρησιμοποιήθηκαν για την εκπαίδευση και τα υπόλοιπα για έλεγχο. Το σύνολο δεδομένων είναι διαθέσιμο online εδώ: 69 dataset.

Γ΄. Schoenmakers13: handwritten characters

Αυτό το σύνολο δεδομένων αναφέρεται ως BRAINS dataset και περιέχει ασπρόμαυρες εικόνες 360 παραδειγμάτων για 6 διαφορετικούς χειρόγραφους χαρακτήρες (B,R,A,I,N,S). Οι εικόνες αυτές έχουν μέγεθος (56,56) και προέρχεται από τη δημοσίευση των [21]. Τα ερεθίσματα παρουσιάστηκαν μέσω του βοθρίου του ματιού με fixation σε ένα 3T fMRI experiment (TR, 1.74 s; voxel size, 2 ×2×2 mm). Οι εικόνες παρουσιάζονταν για 1 δευτερόλεπτο με flashing rate 3Hz. Κάθε εικόνα επαναλαμβανόταν δις και οι fMRI σταθμίστηκαν στον μέσο. Χρησιμοποιήθηκαν τα fMRI δεδομένα των περιοχών V1,V2 τριών υποκειμένων [20]. Ακολουθήθηκε το train/test split της αρχικής δημοσίευσης (290/70 class-balanced χαρακτήρες), ενώ το σύνολο δεδομένων μπορεί να βρεθεί online εδώ: BRAINS dataset. Στον πίνακα I παρατίθενται συνοπτικά τα χαρακτηριστικά των δεδομένων.

IV. Μεθοδολογία

Σε αυτήν τη θεματική ενότητα περιγράφουμε τα βασικά χαρακτηριστικά του μοντέλου μας, τόσο ως προς τις αρχιτεκτονικές του generator και του discriminator όσο και ως προς την εκπαίδευσή του, με σκοπό να επιτευχθεί η ανακατασκευή εικονικών ερεθισμάτων από δεδομένα fMRI. Οι εν λόγω αρχιτεκτονικές φαίνονται στο σχήμα 1.



Σχήμα 1: Generator-Discriminator

Α΄. Generator

Ο generator που αναπτύχθηκε είναι ένα συνελικτικό δίκτυο πρόσθιας τροφοδότησης που δέχεται δεδομένα fMRI s ως είσοδο και παράγει εικόνες x ως έξοδο. Τα δεδομένα που χρησιμοποιούμε αποτελούν ROIs που προέρχονται από fMRI και έχουν τη μορφή διανυσμάτων διάστασης χιλίων voxels. Προκειμένου, λοιπόν, να ενταχθούν στο δίκτυο, περνούν από ένα γραμμικό επίπεδο το οποίο αυξάνει τη διάσταση σε 512×7×7, ώστε στη συνέχεια να πάρουν τη μορφή δισδιάστατων δεδομένων πολύ χαμηλής ανάλυσης (7×7). Οι εικόνες που παράγει ο generator, όμως, έχουν ανάλυση 112×112, επομένως χρησιμοποιούμε 4 blocks με transposed convolutions για την αύξηση της ανάλυσης των δεδομένων εισόδου.

Σε αντίθεση με τις παραδοσιακές αρχιτεκτονικές GANs, η προτεινόμενη προσέγγιση δεν χρησιμοποιεί κάποιο διάνυσμα θορύβου ως είσοδο. Παρατηρήθηκε κατά την διάρκεια των πειραμάτων πως δεν υπάρχει σημαντικά αντιληπτή διαφορά στις ανακατασκευασμένες εικόνες όταν επιπλέον

θόρυβος εισάγεται στον generator. Αυτή η παρατήρηση συνάδει με τα αποτελέσματα των [22], [23], που δείχνουν ότι η είσοδος θορύβου δεν είναι σημαντική όταν η παραγόμενη έξοδος έχει ισχυρή εξάρτηση από τα δεδομένα εισόδου.

Επιπλέον, όπως έχει παρατηρηθεί στο [24], οι generators που χρησιμοποιήθηκαν transposed convolutions είναι επιρρεπείς στο να παράγουν ένα φαινόμενο γνώστο στην βιβλιογραφία ως 'checkerboard artifacts', δηλαδή κάποια μοτίβα πάνω στις εικόνες που έχουν τη μορφή "σκακιέρας". Όπως προτάθηκε λοιπόν στην παραπάνω έρευνα, χρησιμοποιούμε kernel πολλαπλάσιο του stride (4×4 και 2 αντίστοιχα), ώστε να αντιμετωπιστεί αυτό το πρόβλημα.

Τέλος, ακολουθώντας τις πρακτικές του [25], χρησιμοποιούμε ως συνάρτηση ενεργοποίησης τη ReLU μετά από κάθε transposed convolution, αλλά την Tanh στο επίπεδο εξόδου και αρχικοποιούμε τα βάρη του δικτύου από την Κανονική κατανομή με μέση τιμή 0 και τυπική απόκλιση 0.02.

*Β΄. **Discriminator***

Ο discriminator είναι ένα κλασικό συνελικτικό δίκτυο που πραγματοποιεί ταξινόμηση σε δύο κλάσεις. Δέχεται ως είσοδο μια εικόνα διάστασης 1×112×112 και επιστρέφει την πιθανότητα αυτή η εικόνα να είναι πραγματική ή ψεύτικη. Προσπαθεί, δηλαδή, να διακρίνει τις εικόνες σε πραγματικά οπτικά ερεθίσματα των δεδομένων εκπαίδευσης και σε ανακατασκευασμένες εικόνες οι οποίες έχουν παραχθεί από τον generator. Η εικόνα εισόδου αποτελείται από μόνο ένα κανάλι, καθώς τα οπτικά ερεθίσματα των δεδομένων που διαθέτουμε είναι μονοχρωματικά.

Κατά αντιστοιχία με τον generator, ο discriminator αποτελείται από 4 convolution blocks των οποίων οι πράξεις συνέλιξης χρησιμοποιούν kernel 4×4 και stride 2. Ο λόγος που στον discriminator χρησιμοποιήθηκαν strided συνελίξεις αντί για pooling layers είναι επειδή επιτρέπει στο δίκτυο να μάθει τη δική του pooling συνάρτηση, όπως αναφέρεται στο [25]. Επίσης, ως συνάρτηση ενεργοποίησης χρησιμοποιούμε τη LeakyReLU, η οποία βοηθάει στην καλύτερη ροή των gradients, κάτι που επιδρά καθοριστικά στη διαδικασία μάθησης του μοντέλου. Τα βάρη του discriminator αρχικοποιούνται και αυτά από την κατανομή $\mathcal{N}(0, 0.02)$.

*Γ΄. **Training objective***

Αρχικά ορίζουμε τη σημειογραφία που θα χρησιμοποιήσουμε. Έστω x τα δεδομένα που αναπαριστούν τις εικόνες και s τα δεδομένα fMRI. Ως $D(x)$ αναπαριστούμε τον discriminator και ως $G(s)$ τη συνάρτηση του generator που ανακατασκευάζει οπτικά ερεθίσματα x από δεδομένα fMRI s. Επομένως, $D(G(s))$ είναι η πιθανότητα η εικόνα που παρήγαγε ο generator να είναι πραγματική. Όπως περιγράφει ο I. Goodfellow [1] οι D και G παίζουν ένα minimax παιχνίδι όπου ο D προσπαθεί να μεγιστοποιήσει την πιθανότητα να ταξινομεί σωστά τις πραγματικές και ψεύτικες εικόνες ($log(D(x))$) και ο G προσπαθεί να ελαχιστοποιήσει την πιθανότητα ο D να αναγνωρίζει ως ψεύτικες τις εικόνες που παράγει ($log(1 - D(G(s))$). Ορίζουμε λοιπόν την adversarial συνάρτηση κόστους L_{adv} ως:

$$L_{adv}(D, G) = \mathbb{E}_x[\log D(x)] + \mathbb{E}_s[\log(1 - D(G(s)))]$$

Επίσης, θέλουμε να ελαχιστοποιήσουμε τη συνάρτηση κόστους ανακατασκευής εικόνων L_{img}. Για το σκοπό αυτό ορίζουμε την απόσταση $L2$ μεταξύ των εικόνων εισόδου και των ανακατασκευασμένων εικόνων:

$$L_{img}(G) = \mathbb{E}_{x,s}[||x - G(s)||_2^2]$$

Τελικά, το training objective παίρνει τη μορφή:

$$minmax_{G,D} L(D, G) = L_{adv}(D, G) + \lambda L_{img}(G)$$

όπου λ είναι η παράμετρος που ελέγχει το βαθμό στον οποίο συνεισφέρει το κόστος ανακατασκευής στο ολικό κόστος. Στα πειράματά μας ορίζουμε $\lambda = 150$ και ακολουθώντας τις πρακτικές του [25] χρησιμοποιούμε ως optimizer τον Adam με learning rate=0.0002 και Beta1=0.5. Για την επίτευξη του training objective πρακτικά χρησιμοποιούμε ως κριτήρια το Binary Cross Entropy και το Mean Squared Error αντίστοιχα.

V. Αποτελέσματα

Στα πλαίσια της παρούσας εργασίας εξετάστηκαν τα αποτελέσματα της ανακατασκευής των οπτικών ερεθισμάτων μέσω των εξής τριών μετρικών:

i) **(Pixel-wise) Cross Correlation**, το οποίο ορίζεται ως εξής (με την παραδοχή ότι οι εικόνες I_1 και I_2 έχουν τις ίδιες διαστάσεις, ενώ ο τελεστής $|| \cdot ||$ δηλώνει την L1 νόρμα):

$$CC = \frac{\sum_i \sum_j I_1(i, j) \cdot I_2(i, j)}{||I_1 \cdot I_2||}$$

(ii) **Pearson Correlation Coefficient** με τον εξής τύπο (όπου οι αρχικές εικόνες I_1 και I_2 έχουν μετατραπεί σε μονοδιάστατα διανύσματα):

$$\rho_{I_1, I_2} = \frac{cov(I_1, I_2)}{\sigma_{I_1} \cdot \sigma_{I_2}}$$

(iii) **Structural Similarity Index (SSIM)** [26], με σκοπό την ποιοτική ανάλυση της ομοιότητας των εικόνων βασισμένη στην ικανότητα της ανθρώπινης οπτικής αντίληψης. Το SSIM χρησιμοποιεί ως features τη φωτεινότητα (luminance), την αντίθεση (contrast) και τη δομή (structure) για τη σύγκριση των δύο εικόνων. Η σύγκριση της φωτεινότητας δίνεται από την εξίσωση:

$$l(x, y) = \frac{2\mu_{I_1}\mu_{I_2} + C_1}{\mu_{I_1}^2 + \mu_{I_2}^2 + C_1}$$

όπου $C_1 = (K_1 L)^2$ σταθερά, K_1 σταθερά, L το δυναμικό εύρος των τιμών των pixels (είτε από 0 έως 255 αν πρόκειται για 8-bit εικόνες είτε από 0 έως 1 εάν είναι κανονικοποιημένες) και τα μ αφορούν τις μέσες τιμές των pixels των δύο εικόνων. Στη συνέχεια, η σύγκριση της αντίθεσης δίνεται από τη σχέση:

$$c(x, y) = \frac{2\sigma_{I_1}\sigma_{I_2} + C_2}{\sigma_{I_1}^2 + \sigma_{I_2}^2 + C_2}$$

όπου $C_2 = (K_2 L)^2$ σταθερά και σ οι τυπικές αποκλίσεις των εικόνων. Τέλος, η σύγκριση της δομής των εικόνων ορίζεται ως:

$$s(x, y) = \frac{\sigma_{I_1 I_2} + C_3}{\sigma_{I_1} \sigma_{I_2} + C_3}$$

Το τελικό SSIM score για το συγκεκριμένο ζευγάρι εικόνων (I_1, I_2) δίνεται από τον τύπο:

$$SSIM(x, y) = [l(x, y)]^{\alpha} \cdot [c(x, y)]^{\beta} \cdot [s(x, y)]^{\gamma}$$

όπου οι παράμετροι $\alpha > 0$, $\beta > 0$ και $\gamma > 0$ ορίζονται έτσι ώστε να σταθμίσουμε τη συνεισφορά των επιμέρους όρων της παραπάνω σχέσης.

Πίνακας II
Απόδοση του μοντέλου με διαφορετικό πλήθος voxels στα δεδομένα Brains

Dataset	Cross Correlation	PearsonR	SSIM
Brains V1+V2	**0.803**	**0.397**	**0.482**
Brains V1	0.795	0.375	0.48
Brains V2	0.795	0.368	0.479

Πίνακας III
Απόδοση του μοντέλου με διαφορετικό πλήθος voxels στα δεδομένα fMRI MNIST 69

Dataset	Voxels	Cross Correlation	PearsonR	SSIM
69	3000	**0.942**	**0.8**	**0.723**
69	2000	0.929	0.789	0.7
69	1000	0.89	0.68	0.659

Πίνακας IV
Απόδοση του μοντέλου με διαφορετικό πλήθος voxels στα δεδομένα Neuron

Dataset	Voxels	Cross Correlation	PearsonR	SSIM
Neuron	All	0.992	0.97	0.858
Neuron	4000	0.992	0.971	0.855
Neuron	3000	**0.995**	**0.978**	**0.873**
Neuron	2000	0.978	0.943	0.818
Neuron	1000	0.89	0.811	0.617

Τα αποτελέσματα που προέκυψαν παρατίθενται στους πίνακες II, III και IV για τα datasets BRAINS, Digits 69 και Neuro αντίστοιχα, ενώ στα σχήματα 2, 3 και 4 παρουσιάζονται ορισμένα ζευγάρια από πραγματικές και ανακατασκευασμένες εικόνες για τα προαναφερθέντα datasets. Τέλος, στον πίνακα V συνοψίζονται τα σχετικά αποτελέσματα όλων των αρχιτεκτονικών, συμπεριλαμβανομένης και της δικής μας προσέγγισης, χρησιμοποιώντας τις μετρικές Linear (Pearson) correlation και SSIM για τα datasets BRAINS και Digits 69 αντίστοιχα.

(α′) Ερεθίσματα (β′) Ανακατασκευή

(γ′) Ερεθίσματα (δ′) Ανακατασκευή

(ε′) Ερεθίσματα (ς′) Ανακατασκευή

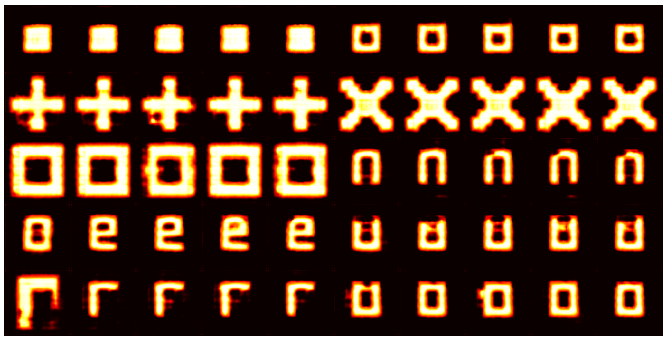
Σχήμα 2: Brains data
Training GIF

(α′) Ερεθίσματα

(β′) Ανακατασκευή

Σχήμα 3: Digits 69 data
Training GIF

(α′) Ερεθίσματα



(β′) Ανακατασκευή

Σχήμα 4: Neuro data
Training GIF

Μιας και η ανακατασκευή εικόνας είναι ενα ποιοτικό πρόβλημα που ποσοτικοποιείται δύσκολα, αξίζει να παρατηρηθούν τα βασικά της σημεία. Συγκεκριμένα, είναι εμφανές ότι στο σύνολο δεδομένων των Digits 69 η ανακατασκευή μπορεί να χαρακτηριστεί ως επιτυχημένη, καθώς είναι διακριτή τόσο η κλάση της κάθε εικόνας,

Πίνακας V
Αξιολόγηση ανακατασκευής εικόνων βασισμένη σε συνήθεις μετρικές (Linear Correlation, SSIM) σε δύο σύνολα δεδομένων για διαφορετικές προσεγγίσεις. Τα αποτελέσματα αποτελούν το μέσο όρο 20 επανεκπαιδεύσεων για κάθε μοντέλο (mean ± std)

Method	vanGerven10 (Digits 69)		Schoenmakers13 (BRAINS)	
	Linear Correlation	SSIM	Linear Correlation	SSIM
SMLR	0.767 ± 0.033	0.466 ± 0.030	0.481 ± 0.096	0.191 ± 0.043
BCCA	0.411 ± 0.157	0.192 ± 0.035	0.348 ± 0.138	0.058 ± 0.042
DCCAE-A	0.548 ± 0.044	0.358 ± 0.097	0.354 ± 0.167	0.186 ± 0.234
DCCAE-S	0.511 ± 0.057	0.552 ± 0.088	0.351 ± 0.153	0.179 ± 0.117
De-CNN	0.799 ± 0.062	0.613 ± 0.043	0.470 ± 0.149	0.322 ± 0.118
DGMM	0.803 ± 0.063	0.645 ± 0.054	0.498 ± 0.193	0.340 ± 0.051
Denoiser GAN	0.531 ± 0.049	0.529 ± 0.043	0.319 ± 0.032	0.465 ± 0.031
DGMM +	0.813 ± 0.053	0.651 ± 0.044	0.502 ± 0.193	0.360 ± 0.050
D-VAE/GAN	**0.837 ± 0.014**	0.714 ± 0.014	**0.740 ± 0.020**	**0.587 ± 0.019**
Ours	0.807 ± 0.005	**0.718 ± 0.004**	0.395 ± 0.006	0.478 ± 0.007

αλλά και οι λεπτομέρειές της, όπως το σχήμα ή το πάχος των γραμμών. Στη συγκεκριμένη περίπτωση, οι αποκλίσεις μεταξύ πραγματικών και ανακατασκευασμένων εικόνων οφείλονται στις τιμές του brightness για ορισμένα pixels των ανακατασκευασμένων εικόνων, τα οποία εμφανίζονται με έντονο κόκκινο χρώμα (δηλαδή παίρνουν χαμηλή τιμή). Τέλος, τα συμπεράσματά μας αντικατοπτρίζονται και στις τιμές των μετρικών που χρησιμοποιήθηκαν, οι οποίες είναι ιδιαίτερα υψηλές.

Τα πράγματα δεν είναι τόσο απλοϊκά στο B.R.A.I.N.S dataset όπου παρατηρούνται τρεις διαφορετικές περιπτώσεις ανακατασκευής. Πρώτον, παρατηρούμε τις επιτυχημένες όπως αυτές φαίνονται στο 2(β′) όπου τόσο η συγκεκριμένη κλάση, αλλά και χαρακτηριστικά του γράμματος ανακατασκευάζονται. Έπειτα, βλέπουμε περιπτώσεις γραμμάτων που οι κλάσεις των στοιχείων είναι αισθητές αλλά οχι ευδιάκριτες, ενώ τέλος υπάρχει και η τρίτη περίπτωση κατά την οποία η ανακατασκευή αποτυγχάνει εντελώς και τα αποτελέσματα θυμίζουν άλλη κλάση (που πολύ πιθανόν να οφείλεται σε ένα χαρακτηριστικό πρόβλημα της εκπαίδευσης των GANs με ονομασία “mode collapse"). Στο ανθρώπινο μάτι, τα ανακατασκευασμένα γράμματα στις σειρές 2(β′) και (δ′) μπορούν να γίνουν αντιληπτά (έστω και δύσκολα), ωστόσο αυτό δεν ισχύει και για τα αποτελέσματα της σειράς 2 (ζ′). Τέλος, παρατηρούμε ότι η τιμή του SSIM είναι υψηλότερη από το Pearson correlation coefficient, το οποίο φαίνεται ότι επιβάλλει μεγαλύτερη ποινή λόγω αυτών των αναντιστοιχιών μεταξύ των πραγματικών και των ανακατασκευασμένων ερεθισμάτων.

Τέλος και για την περίπτωση του Neuro dataset λαμβάνουμε άκρως ικανοποιητικά αποτελέσματα, αφού στην πλειονότητα των εικόνων η κατηγορία είναι εμφανής. Παρ′ όλα αυτά, βλέπουμε πως θα μπορούσαν να είχαν προκύψει προβλήματα όπως συγχύσεις μεταξύ ’O’ και ’U’. Κλείνοντας, στη συγκεκριμένη περίπτωση ο συντελεστής PearsonR παίρνει υψηλότερη τιμή από το SSIM, καθώς οι μικρές ατέλειες που εμφανίζονται κατά την ανακατασκευή περιορίζουν έως ένα βαθμό την πιστότητά της βάσει της προσωπικής μας οπτικής αντίληψης.

VI. Συμπεράσματα και μελλοντικές επεκτάσεις

Φαίνεται από τις παραπάνω μετρικές πως ορισμένες αρχιτεκτονικές παρήγαγαν καλύτερα αποτελέσματα από τα δικά μας πειράματα. Το φαινόμενο αυτό μπορεί πιθανώς να αιτιολογηθεί από δύο αρκετά ενδιαφέρουσες παρατηρήσεις. Από τη μία, όπως είναι λογικό σε πειράματα τέτοιου είδους, η ακρίβεια των αποτελεσμάτων είναι ανάλογη της ποιότητας των αντίστοιχων δεδομένων. Είναι άξιο σχολιασμού λοιπόν το σημαντικό πρόβλημα έλλειψης ποιοτικών δεδομένων που φαίνεται να υπάρχει στους ερευνητικούς κύκλους του fMRI reconstruction.

Παραδείγματος χάριν, στο B.R.A.I.N.S dataset μπορούμε να σημειώσουμε δύο αστοχίες. Αφενός θα έπρεπε να έχουν ληφθεί περισσότερα δείγματα, κάτι που πάντα βοηθά σε ένα πρόβλημα τόσο υψηλής δυσκολίας και, αφετέρου, ίσως θα έπρεπε οι ασθενείς να είχαν εκτεθεί για περισσότερο χρόνο στα διαθέσιμα οπτικά ερεθίσματα. Στην παρούσα περίπτωση εκτέθηκαν σε αυτά για μόλις ένα δευτερόλεπτο, κάτι που μπορεί να επηρεάσει την ποιότητα των αποτελεσμάτων αν αναλογιστούμε πως αρκετοί ασθενείς μπορεί να μην προλάβουν να συγκεντρωθούν στο ερέθισμα και να το παρατηρήσουν μέσα σε ένα δευτερόλεπτο.

Βέβαια, πολλές από αυτές τις ελλείψεις στα πειραματικά δεδομένα είναι λογικές αν αναλογιστούμε τη δυσκολία που συνοδεύει ένα πείραμα fMRI. Τις περισσότερες φορές τα αποτελέσματα των συγκεκριμένων εξετάσεων δεν είναι δυνατόν να διατεθούν ανοιχτά στο ευρύ κοινό μιας και συσχετίζονται άμεσα με πραγματικούς ασθενείς. Συνεπώς, προστατεύονται από πολλαπλούς νόμους ιατρικού απορρήτου, κάτι που έχει ως αποτέλεσμα πολλές μελέτες fMRI

να κρατάνε τα δεδομένα τους κρυφά. Επιπλέον, σε αυτό το φαινόμενο συνεισφέρει και το γεγονός ότι το κόστος μιας ανοικτής μελέτης fMRI είναι αρκετά υψηλό. Σε κάθε περίπτωση, θα έδινε μια πολύ σημαντική ώθηση στον κλάδο η δημοσίευση μεγαλύτερου και ποιοτικότερου όγκου δεδομένων fMRI.

Από την άλλη, εντοπίζουμε πιθανές περιπτώσεις διαρροής δεδομένων (data leakage) σε πολλές από τις αρχιτεκτονικές που μελετήσαμε. Η διαρροή δεδομένων είναι ένα φαινόμενο που εμφανίζεται είτε από άγνοια του σχεδιαστή, είτε σκόπιμα με στόχο τη δημιουργία τεχνητά βελτιωμένων αποτελεσμάτων. Ο πιο απλός τρόπος με τον οποίο λαμβάνει χώρα αυτό το φαινόμενο είναι η συμπερίληψη των δεδομένων test στο σύνολο δεδομένων με το οποίο εκπαιδεύονται τα μοντέλα (training set). Όταν συμβαίνει αυτό, το μοντέλο μπορεί να χρησιμοποιήσει δεδομένα testing, στα οποία κανονικά δε θα είχε πρόσβαση, προκειμένου να εκπαιδευτεί. Φυσικά, ο στόχος της εκπαίδευσης είναι η ανάπτυξη της ιδιότητας του μοντέλου να προβλέπει όσο τον δυνατόν καλύτερα τα δεδομένα του testing dataset. Είναι προφανές πως η μέθοδος αυτή οδηγεί σε παραπλανητικά αποτελέσματα και προκαλεί βλάβη στην ικανότητα γενίκευσης του μοντέλου.

Στη δημοσίευση Voxels-to-Pixels and Back [15], οι συγγραφείς εισάγουν μια διαδικασία unsupervised εκπαίδευσης σε δεδομένα του testing set. Ενώ αυτό αποτελεί κλασικό δείγμα διαρροής δεδομένων, οι συγγραφείς υποστηρίζουν πως πρόκειται για μια έγκυρη προσέγγιση μιας και δε χρησιμοποιούν τα labels των συγκεκριμένων testing εικόνων. Μπορούμε όμως να δούμε ξεκάθαρα από τα αποτελέσματα τους πως, όταν στην εκπαίδευση της αρχιτεκτονικής δεν χρησιμοποιούνται testing δεδομένα, η απόδοση πέφτει κατακόρυφα. Αντίστοιχα, στον κώδικα των υλοποιήσεων ShapeGAN [17] και D-VAE/GAN [16], είναι πρόδηλο, πως οι εικόνες εκπαίδευσης και οι εικόνες ελέγχου συμπτύσσονται στο ίδιο διάνυσμα πριν δοθούν στο δίκτυο. Προφανώς, έχουμε άλλη μια περίπτωση διαρροής δεδομένων και, συνεπώς, παραπλανητικών αποτελεσμάτων - κάτι το οποίο επισημαίνουν και ορισμένοι reviewers των εν λόγω δημοσιεύσεων.

Σε ότι αφορά τις μελλοντικές επεκτάσεις που μπορούν να γίνουν στις εν λόγω τεχνικές, μπορούμε να αναγνωρίσουμε δύο κατευθύνσεις προς τις οποίες κινείται η έρευνα. Η πρώτη [27] αφορά τα δεδομένα που μπορούμε να ανακατασκευάσουμε από τα σήματα fMRI. Αντί να προσπαθήσουμε να ανακατασκευάσουμε την αρχική εικόνα-ερέθισμα, μπορούμε να επικεντρωθούμε στην ανάσυρση τρισδιάστατων χαρτών βάθους (3D depth maps) από τα ερεθίσματα. Οι εν λόγω χάρτες μας επιτρέπουν να χαρακτηρίσουμε κάθε voxel με βάση την πληροφορία βάθους που δίνει και, με αυτό τον τρόπο, να πάρουμε σημαντικές πληροφορίες για τα features των ερεθισμάτων. Αυτή η τεχνική προσθέτει ένα ακόμα βήμα στα end-to-end μοντέλα που μοιάζουν με το Voxels to Pixels και βοηθά το δίκτυο να παράγει ακόμα πιο πιστές ανακατασκευές εικόνων.

Από την άλλη, η μεθοδολογία μας μπορεί να επεκταθεί με τη χρήση του transfer learning, μιας τεχνικής που μας επιτρέπει να μεταφέρουμε γνώση από ένα GAN το οποίο έχει καλή απόδοση σε κάποιο πρόβλημα, σε ένα νέο GAN που επιδιώκει να λύσει ένα διαφορετικό πρόβλημα [28]. Η συγκεκριμένη τεχνική χρησιμοποιείται ευρέως σε όλους τους τομείς του deep learning και μπορεί να μας βοηθήσει να επιλύσουμε το πρόβλημα των περιορισμένων δεδομένων μιας και δε θα βασιζόμαστε πλέον εξολοκλήρου στο μικρό όγκο δεδομένων fMRI που έχουμε στην κατοχή μας.

VII. Ευχαριστίες

Η παρούσα εργασία εκπονήθηκε στα πλαίσια του μεταπτυχιακού μαθήματος της 'Ανάλυσης Βιο-δεδομένων' στο πρόγραμμα σπουδών ε.δε.μ^2 του τμήματος των Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών.

Ιδιαίτερες ευχαριστίες θα θέλαμε να απονείμουμε στην διδάσκουσα του μαθήματος κα. Ζαρκογιάννη, καθώς μας εισήγαγε στο Deep Learning στον κλάδο των βιοδεδομένων, ενώ ταυτοχρόνα μας οδήγησε στην καλλιέργεια των soft skills μας αναφορικά με την συνεργασία, την συνεννόηση και την οργάνωση χρόνου και πόρων.

VIII. Βιβλιογραφία

[1] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014.

[2] R. Poldrack and M. Farah, "Progress and challenges in probing the human brain," *Nature*, vol. 526, pp. 371–379, 2015.

[3] T. Yoshida and K. Ohki, "Natural images are reliably represented by sparse and variable populations of neurons in visual cortex," *Nature*, vol. 872, 2020.

[4] Y. Miyawaki, H. Uchida, O. Yamashita, M.-a. Sato, Y. Morito, H. C. Tanabe, N. Sadato, and Y. Kamitani, "Visual image reconstruction from human brain activity using a combination of multiscale local image decoders," *Neuron*, vol. 60, no. 5, p. 915—929, December 2008. [Online]. Available: https://doi.org/10.1016/j.neuron.2008.11.004

[5] K. Kay, T. Naselaris, R. Prenger, and J. Gallant, "Identifying natural images from human brain activity," *Nature*, vol. 452, pp. 352–5, 04 2008.

[6] S. Nishimoto, A. T. Vu, T. Naselaris, Y. Benjamini, B. Yu, and J. L. Gallant, "Reconstructing visual experiences from brain activity evoked by natural movies," *Current Biology*, vol. 21, no. 19, pp. 1641–1646, 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0960982211009377

[7] T. Naselaris, R. J. Prenger, K. N. Kay, M. Oliver, and J. L. Gallant, "Bayesian reconstruction of natural images from human brain activity," *Neuron*, vol. 63, no. 6, pp. 902–915, 2009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0896627309006850

[8] U. Güçlü and M. A. J. van Gerven, "Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream," *Journal of Neuroscience*, vol. 35, no. 27, pp. 10 005–10 014, 2015. [Online]. Available: https://www.jneurosci.org/content/35/27/10005

[9] S. Zhou, C. R. Cox, and H. Lu, "Improving whole-brain neural decoding of fmri with domain adaptation," *bioRxiv*, 2018. [Online]. Available: https://www.biorxiv.org/content/early/2018/08/07/375030

[10] P. Ghaemmaghami, M. Nabi, Y. Yan, G. Riccardi, and N. Sebe, "A cross-modal adaptation approach for brain decoding," *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 969–973, 2017.

[11] T. Horikawa and Y. Kamitani, "Hierarchical neural representation of dreamed objects revealed by brain decoding with deep neural network features," 2017.

[12] ——, "Generic decoding of seen and imagined objects using hierarchical visual features," 2016.

[13] G. St-Yves and T. Naselaris, "Generative adversarial networks conditioned on brain activity reconstruct seen images," in *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2018, pp. 1054–1061.

[14] G. Shen, K. Dwivedi, K. Majima, T. Horikawa, and Y. Kamitani, "End-to-end deep image reconstruction from human brain activity," *Frontiers in Computational Neuroscience*, vol. 13, p. 21, 2019. [Online]. Available: https://www.frontiersin.org/article/10.3389/fncom.2019.00021

[15] R. Beliy, G. Gaziv, A. Hoogi, F. Strappini, T. Golan, and M. Irani, "From voxels to pixels and back: Self-supervision in natural-image reconstruction from fmri," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper/2019/file/7d2be41b1bde6ff8fe45150c37488ebb-Paper.pdf

[16] Z. Ren, J. Li, X. Xue, X. Li, F. Yang, Z. Jiao, and X. Gao, "Reconstructing perceived images from brain activity by visually-guided cognitive representation and adversarial learning," *CoRR*, vol. abs/1906.12181, 2019. [Online]. Available: http://arxiv.org/abs/1906.12181

[17] T. Fang, Y. Qi, and G. Pan, "Reconstructing perceptive images from brain activity by shape-semantic gan," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 13 038–13 048. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/9813b270ed0288e7c0388f0fd4ec68f5-Paper.pdf

[18] Y. Miyawaki, H. Uchida, O. Yamashita, M.-a. Sato, Y. Morito, H. C. Tanabe, N. Sadato, and Y. Kamitani, "Visual image reconstruction from human brain activity using a combination of multiscale local image decoders," *Neuron*, vol. 60, no. 5, pp. 915–929, 2008.

[19] M. A. Van Gerven, F. P. De Lange, and T. Heskes, "Neural decoding with hierarchical generative models," *Neural computation*, vol. 22, no. 12, pp. 3127–3142, 2010.

[20] S. Schoenmakers, M. Barth, T. Heskes, and M. Van Gerven, "Linear reconstruction of perceived images from human brain activity," *NeuroImage*, vol. 83, pp. 951–961, 2013.

[21] L. Van der Maaten, "A new benchmark dataset for handwritten character recognition," *Tilburg University*, pp. 2–5, 2009.

[22] M. Mathieu, C. Couprie, and Y. LeCun, "Deep multi-scale video prediction beyond mean square error," 2016.

[23] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jul. 2017. [Online]. Available: https://doi.org/10.1109/cvpr.2017.632

[24] A. Odena, V. Dumoulin, and C. Olah, "Deconvolution and checkerboard artifacts," *Distill*, 2016. [Online]. Available: http://distill.pub/2016/deconv-checkerboard

[25] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2016.

[26] 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.

[27] G. Gaziv and M. Irani, "More than meets the eye: Self-supervised depth reconstruction from brain activity," *arXiv preprint arXiv:2106.05113*, 2021.

[28] Y. Wang, C. Wu, L. Herranz, J. van de Weijer, A. Gonzalez-Garcia, and B. Raducanu, "Transferring gans: generating images from limited data," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 218–234.