Contrastive learning on medical image segmentation

Χρήστος Χαρίσης
ΣΗΜΜΥ
Εθνικό Μετσόβιο Πολυτεχνείο
Αθήνα, Ελλάδα
christ.charisis@gmail.com

Μαρία Νεκταρία Μηναΐδη
ΣΗΜΜΥ
Εθνικό Μετσόβιο Πολυτεχνείο
Αθήνα, Ελλάδα
minaidimaria@gmail.com

Γεώργιος Ουζουνίδης
ΣΗΜΜΥ
Εθνικό Μετσόβιο Πολυτεχνείο
Αθήνα, Ελλάδα
giorgos_ouzounidis@hotmail.com

***Abstract*—Μια βασική προϋπόθεση για την επιτυχία της εποπτευόμενης βαθιάς μάθησης είναι ένα μεγάλο σύνολο δεδομένων, το οποίο θα περιέχει και τις ετικέτες των δεδομένων. Η συλλογή αυτού του συνόλου δεδομένων εμπεριέχει πολλές δυσκολίες και η διαδικασία είναι περίπλοκη, ειδικότερα όταν η ανάλυση περιλαμβάνει ιατρικές εικόνες. Η αυτοεπιβλεπόμενη μάθηση (Self-Supervised Learning-SSL) συμβάλλει στην επίλυση του παραπάνω προβλήματος, παρέχοντας μια στρατηγική για να προεκπαιδευτεί ένα νευρωνικό δίκτυο με δεδομένα χωρίς ετικέτα, ακολουθούμενο από fine-tuning. Το contrastive-learning, μια κατηγορία του SSL, αποτελεί τεχνική για την εκμάθηση αναπαραστάσεων σε επίπεδο εικόνας. Στην παρούσα μελέτη, προτείνονται στρατηγικές για την χρήση του SSL, στο image segmentation των ογκομετρικών ιατρικών εικόνων, σε ημι-εποπτευόμενο περιβάλλον μάθησης, με περιορισμένες ετικέτες. Ειδικότερα, προτείνουμε νέες στρατηγικές contrastive learning, οι οποίες ενισχύουν την ομοιότητα των εικόνων σε δομικό επίπεδο και εφαρμόζουμε τη μέθοδο αυτή σε διάφορα σύνολα δεδομένων, αναλύοντας σε τελικό επίπεδο, τα συμπεράσματα που προκύπτουν. Κάνουμε μια εκτεταμένη αξιολόγηση σε δύο σύνολα δεδομένων: το ACDC dataset και το Prostate Dataset. Τα δύο αυτά σύνολα δεδομένων αποτελούνται από μαγνητικές τομογραφίες της καρδιάς και του προστάτη αντίστοιχα. Η εκπαίδευση έγινε στο ACDC dataset, ενώ η αξιολόγηση και στα δύο. Η προτεινόμενη μέθοδος αποφέρει ουσιαστικές βελτιώσεις στην περίπτωση την ελλιπών δεδομένων με ετικέτα, σε σύγκριση με άλλες τεχνικές αυτοεπιβλεπόμενης και ημι-εποπτευόμενης μάθησης. Σκοπός μας είναι να συγκρίνουμε τις επιδόσεις μας με τις ήδη υπάρχουσες μεθόδους και να πετύχουμε υψηλότερες ακρίβειες με λιγότερα επισημασμένα δεδομένα.**

***Index Terms*—Contrastive learning, Image Segmentation, Deep Learning, Neural networks, Unet**

I. Εισαγωγή

Η εποπτευόμενη βαθιά μάθηση παρέχει προηγμένη τμηματοποίηση ιατρικής εικόνας, όταν είναι διαθέσιμα μεγάλα σύνολα δεδομένων. Ωστόσο, η συγκέντρωση μεγάλων σχολιασμένων συνόλων δεδομένων παρουσιάζει πολυάριθμες προκλήσεις, επομένως μέθοδοι που μπορούν να μετριάσουν αυτήν την απαίτηση είναι ιδιαίτερα επιθυμητές. Η αυτόεποπτευόμενη μάθηση (SSL) αποτελεί μια πολλά υποσχόμενη κατεύθυνση προς το σκοπό αυτό: παρέχει μια pre-training στρατηγική, που βασίζεται μόνο σε δεδομένα χωρίς ετικέτα, για την απόκτηση κατάλληλου αρχικού συνόλου δεδομένων, το οποίο μετέπειτα θα χρησιμοποιηθεί στην διαδικασία της εκπαίδευσης εργασιών, με περιορισμένους σχολιασμούς (ετικέτες). Τα τελευταία χρόνια, οι μέθοδοι αυτό-επιβλεπόμενης μάθησης είναι πολύ επιτυχημένες για την ανάλυση, όχι μόνο φυσικών, αλλά και ιατρικών εικόνων.

Στην παρούσα εργασία, θα επικεντρωθούμε στο κομμάτι της αυτό-επιβλεπόμενης μάθησης που ονομάζεται contrastive learning. Η βάση του contrastive learning στηρίζεται στο γεγονός, ότι οι διαφορετικές μορφές της ίδιας εικόνας θα πρέπει να έχουν παρόμοιες αναπαραστάσεις και αυτές οι αναπαραστάσεις θα πρέπει, φυσιολογικά, να διαφέρουν αρκετά από τις αναπαραστάσεις άλλων εικόνων. Στην πράξη αυτό περιγράφεται από το contrastive loss, το οποίο διατυπώνεται με τρόπο, ώστε να εκφράσει μαθηματικά, την διαισθητική σημασία του contrastive learning και αναλύεται σε επόμενη παράγραφο. Ένα νευρωνικό δίκτυο Unet εκπαιδεύεται με δεδομένα χωρίς ετικέτες για να ελαχιστοποιήσει το contrastive loss. Το νευρωνικό αυτό δίκτυο, εξάγει τις αναπαραστάσεις των εικόνων που είναι πολύ χρήσιμες για downstream tasks, όπως είναι η ταξινόμηση και η ανίχνευση αντικειμένων, καθώς επίσης, μας δίνει μία ικανοποιητική αρχικοποίηση ενός μοντέλου, το οποίο μπορεί στην συνέχεια να τελειοποιηθεί, ακόμα και με μικρό αριθμό επισημασμένων δεδομένων.

Οι περισσότερες παρόμοιες έρευνες επικεντρώνονται στο συνολικό contrastive loss και στην εξαγωγή συνολικών αναπαραστάσεων, αλλά δεν δίνεται η απαραίτητη σημασία στην εξαγωγή μέτρων ομοιότητας μεταξύ των εικόνων, καθώς οι περισσότερες έρευνες στηρίζονται στην αύξηση των δεδομένων. Στην παρούσα μελέτη στοχεύουμε να διερευνήσουμε την περιοχή αυτή, εξετάζοντας την τμηματοποίηση ογκομετρικών ιατρικών εικόνων, μέσω δεικτών ομοιότητας των διαφορετικών περιοχών μίας εικόνας. Χρησιμοποιούμε, όπως θα αναλυθεί και στην συνέχεια, το Unet και παραλλαγές του, προκειμένου να προεκπαιδευτεί το νευρωνικό δίκτυο, έτσι ώστε η μετέπειτα εκπαίδευση να απαιτεί μικρότερο αριθμό δειγμάτων, για την εξαγωγή της μάσκας των εικόνων. Αξιολογούμε την τεχνική αυτή σε δύο διαφορετικά σύνολα δεδομένων και συγκρίνουμε τα αποτελέσματα, που μας δίνει το μοντέλο μας, με αποτελέσματα άλλων state-of-the-art ερευνών, ενώ παραθέτουμε, εν κατακλείδι, τα συμπεράσματά μας.

A. Κλινικός Αντίκτυπος

Η αυτοματοποίηση των ιατρικών αναλύσεων της ιατρικής απεικόνισης, όπως είναι η τμηματοποίηση εικόνας (image

segmentation), αποτελεί έναν καίριο στόχο της σύγχρονης ιατρικής, καθώς ο αυτοματισμός διαδικασιών επιτρέπει ειδοποιήσεις και αποτελέσματα σε πραγματικό χρόνο στο σημείο της φροντίδας, με αποτέλεσμα οι κλινικοί γιατροί να μπορούν να λαμβάνουν έγκαιρες αποφάσεις και να βοηθούν τους ασθενείς τους πολύ πιο αποτελεσματικά. Σε πολλές χώρες είναι γνωστό ότι υπάρχει έλλειψη ακτινολόγων στα νοσοκομεία σε σύγκριση με τον αριθμό των ασθενών που απεικονίζονται, οδηγώντας έτσι σε υπερβολικό φόρτο εργασίας και επακόλουθες καθυστερήσεις στη διάγνωση, στην πρόγνωση και στις παρεμβάσεις πάνω σε διάφορες παθήσεις. Η αυτοματοποίηση χρονοβόρων αναλύσεων ιατρικής απεικόνισης, μπορεί να βοηθήσει στη μείωση του φόρτου εργασίας και στην αποτελεσματικότερη εργασία, για τους ραδιολόγους, τους ακτινολόγους, τους ογκολόγους, αλλά και κάθε άλλη ιατρική ειδικότητα και στην επιτάχυνση των διαδικασιών υγείας.

B. Δομή

Η εργασία μας αρχικά παραθέτει την σχετική βιβλιογραφία, η οποία μελετήθηκε και στην οποία στηριχθήκαμε, στην συνέχεια αναλύονται τα δύο σύνολα δεδομένων, τα οποία χρησιμοποιήθηκαν για τα πειράματά μας και παρατίθεται αναλυτικά η μεθοδολογία που ακολουθήθηκε. Στην συνέχεια αναπαράγονται και παρουσιάζονται τα αποτελέσματα του πρότυπου paper και προβαίνουμε σε περαιτέρω πειράματα, τα οποία αναλύουμε. Εν τέλει, προχωράμε στα συμπεράσματά μας και σε μελλοντικές προτάσεις.

II. Βιβλιογραφική Επισκόπηση

Πρόσφατες έρευνες έδειξαν ότι η αυτό-επιβλεπόμενη μάθηση [1]–[3] μπορεί να μάθει χρήσιμες αναπαραστάσεις από μη επισημασμένα δεδομένα, ελαχιστοποιώντας την κατάλληλη συνάρτηση κόστους, χωρίς επίβλεψη κατά τη διάρκεια της εκπαίδευσης. Το δίκτυο που προκύπτει αποτελεί μία ικανοποιητική αρχικοποίηση για τα μετέπειτα downstream tasks. Οι μέθοδοι pretext task-based αποτελούν τεχνικές που μπορούν να αποκτήσουν τις επιθυμητές ετικέτες από εικόνες χωρίς πληροφορία. Παραδείγματα τέτοιων μελετών είναι η πρόβλεψη του προσανατολισμού της εικόνας [4], η ανακατασκευή εικόνων [5] και το inpainting [2].

Επιπλέον, οι τεχνικές contrastive learning, χρησιμοποιούν ένα contrastive loss, προκειμένου να ενισχυθούν οι αναπαραστάσεις παρόμοιων περιοχών των εικόνων. Η ομοιότητα αυτή ορίζεται σε μη-επιβλεπόμενο περιβάλλον μάθησης, κυρίως μέσω διαφορετικών αναπαραστάσεων της ίδιας εικόνας ως παρεμφερή παραδείγματα, όπως προτείνουν στα [6]–[8]. Οι [9], [10] μεγιστοποιούν την κοινή πληροφορία (Mutual Information) [11], τεχνική που πλησιάζει αυτήν που έχουμε εφαρμόσει και στην μελέτη μας. Στα [9], [12], [13] μεγιστοποιείται η κοινή πληροφορία μεταξύ διαφορετικών περιοχών και επιπέδων της εικόνας, μέσω ενός encoder νευρωνικού δικτύου με ένα ή περισσότερα επίπεδα.

Σχετικές κατευθύνσεις που χρησιμοποιούν μη-επισημασμένα δεδομένα, για να αντιμετωπίσουν το πρόβλημα της έλλειψης επαρκών δεδομένων με ετικέτες, είναι η ημί-επιβλεπόμενη μάθηση και η αύξηση των δεδομένων. Οι τεχνικές ημί-επιβλεπόμενης μάθησης χρησιμοποιούν δεδομένα με, αλλά και χωρίς ετικέτες κατά τη διάρκεια της εκπαίδευσης των μοντέλων [5], [14]–[18] και για την ανάλυση ιατρικών εικόνων, κατάλληλες μέθοδοι είναι το self-training [19]–[21] και το adversarial training [22]. Όπως έχει αναφερθεί, η εκτενής αύξηση των δεδομένων έχει, επίσης, αποδειχθεί ότι είναι αποτελεσματική για την λεπτομερή ανάλυση των ιατρικών εικόνων με λίγα δεδομένα με ετικέτες. Η αύξηση αυτή μπορεί να γίνει μέσω τυχαίων συσχετίσεων [23], μετασχηματισμών αντίθεσης [24] και τέλος, μέσω GAN δικτύων [25], [26].

Οι παρόμοιες αυτές μελέτες στηρίζονται στο u-net και στα CNN, ενώ για την αξιολόγηση των αποτελεσμάτων χρησιμοποιείται ο δείκτης Dice Score. Στην εργασία μας, δεν θα στηριχτούμε σε αρχιτεκτονικές encoder, αλλά σε αρχιτεκτονικές encoder-decoder, προκειμένου να επικεντρωθούμε στην πρόβλεψη μέσω των πίξελ των εικόνων.

III. Δεδομένα

A. Σύνολα Δεδομένων

Για την αξιολόγηση της προτεινόμενης προσέγγισης, χρησιμοποιούμε δυο διαθέσιμα στο κοινό σύνολα δεδομένων, τα οποία αποτελούνται από μαγνητικές τομογραφίες (MRI). Το πρώτο είναι το σύνολο δεδομένων ACDC, το οποίο φιλοξενήθηκε στην πρόκληση MICCAI 2017 ACDC [27]. Το σύνολο δεδομένων ACDC δημιουργήθηκε από πραγματικές κλινικές εξετάσεις, που αποκτήθηκαν στο Πανεπιστημιακό Νοσοκομείο της Ντιζόν στη Γαλλία. Τα ληφθέντα δεδομένα ήταν πλήρως ανώνυμα και διεκπεραιώθηκαν σύμφωνα με τους κανονισμούς που έθεσε η τοπική επιτροπή δεοντολογίας του Νοσοκομείου της Ντιζόν. Το σύνολο δεδομένων μας, καλύπτει αρκετές, σαφώς καθορισμένες, παθολογίες με αρκετές περιπτώσεις, ώστε να εκπαιδεύονται σωστά οι μέθοδοι μηχανικής μάθησης και να αξιολογούν σαφώς τις παραλλαγές των κύριων φυσιολογικών παραμέτρων, που λαμβάνονται από το cine-MRI (συγκεκριμένα το διαστολικό όγκο και το κλάσμα εξώθησης).

Το σύνολο δεδομένων αποτελείται από 150 εξετάσεις (καθεμία από διαφορετικό ασθενή), χωρισμένες σε 5 ομοιόμορφα κατανεμημένες υποομάδες (4 παθολογικές συν 1 υγιή θεματικές ομάδες), όπως περιγράφεται παρακάτω. Τα δεδομένα εκπαίδευσης αποτελούνται από 100 3D κύβους που αναφέρονται σε μικρού άξονα καρδιακή μαγνητική τομογραφία, με εξειδικευμένους σχολιασμούς σχετικά με τις: αριστερή κοιλία, μυοκάρδιο και δεξιά κοιλία. Επιπλέον, κάθε ασθενής συνοδεύεται από τις ακόλουθες πρόσθετες πληροφορίες: βάρος, ύψος, καθώς και τις στιγμές διαστολικής και συστολικής φάσης.

Τα δεδομένα αποκτήθηκαν σε μία περίοδο 6 ετών, χρησιμοποιώντας δύο μαγνητικούς σαρωτές διαφορετικής μαγνητικής ισχύος (1,5 T (Siemens Area, Siemens Medical Solutions, Germany) και 3,0 T (Siemens Trio Tim, Siemens Medical Solutions, Germany). Οι εικόνες Cine MR αποκτήθηκαν

σε αναστολή αναπνοής με αναδρομική ή προοπτική πύλη και με ακολουθία SSFP σε προσανατολισμό μικρού άξονα. Συγκεκριμένα, μια σειρά φετών μικρού άξονα καλύπτει το LV από τη βάση έως την κορυφή, με πάχος 5 mm (ή μερικές φορές 8 mm) και μερικές φορές διάκενο μεταξύ 5 mm (τότε μία εικόνα κάθε 5 ή 10 mm, σύμφωνα με η εξέταση). Η χωρική ανάλυση κυμαίνεται από 1,37 έως 1,68 mm2 / pixel και 28 έως 40 εικόνες καλύπτουν πλήρως ή εν μέρει τον καρδιακό κύκλο (στη δεύτερη περίπτωση, με προοπτική πύλη, παραλείφθηκε μόνο το 5 έως 10 τις εκατό του τέλους του καρδιακού κύκλου), όλα ανάλογα με τον ασθενή.

Στις παρακάτω εικόνες μπορούμε να δούμε τα δεδομένα για 2 ασθενείς μαζί με της κατάλληλες επισημάνσεις για την αριστερή κοιλία, το μυοκάρδιο και την δεξιά κοιλία.

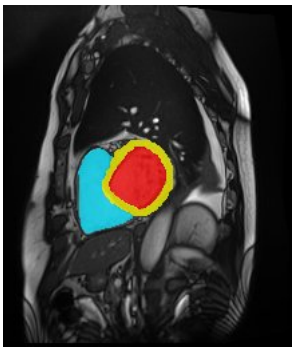


Figure 1: ACDC Dataset-patient1.

Το δεύτερο dataset αφορά τον προστάτη. Τα δεδομένα του προστάτη φιλοξενούνταν στην πρόκληση ιατρικής τμηματοποίησης MICCAI 2018 [28]. Αποτελείται από 48 τρισδιάστατα MRI σταθμικής περιοχής του προστάτη με εξειδικευμένους σχολιασμούς, για δύο δομές: περιφερειακή ζώνη και κεντρικό αδένα. Παρέχονται από το Πανεπιστήμιο Radboud (Ολλανδία), που αναφέρθηκαν σε μια προηγούμενη μελέτη τμηματοποίησης [29]. Χρησιμοποιήθηκε χειροκίνητη τμηματοποίηση ολόκληρου του προστάτη από εγκάρσιες σαρώσεις T2 με ανάλυση 0,6 x 0,6 x 4 mm και τον χάρτη φαινόμενου συντελεστή διάχυσης (ADC) (2 x 2 x 4 mm).

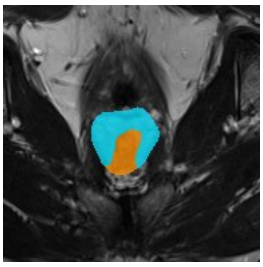


Figure 2: Prostate Dataset-patient47.

B. Προεπεξεργασία Δεδομένων

Το bias field είναι ένα σήμα χαμηλής συχνότητας που καταστρέφει τις εικόνες MRI, ειδικά αυτές που παράγονται από παλιές μηχανές μαγνητικής τομογραφίας. Οι αλγόριθμοι επεξεργασίας εικόνας, όπως τμηματοποίηση, ανάλυση υφής ή ταξινόμηση που χρησιμοποιούν τις τιμές γκρίζου επιπέδου των εικονοστοιχείων εικόνας δεν θα παράγουν ικανοποιητικά αποτελέσματα δίχως την κατάλληλη διόρθωση. Απαιτείται ένα βήμα προ-επεξεργασίας για τη διόρθωση του σήματος πεδίου προκατάληψης πριν από την υποβολή κατεστραμμένων εικόνων MRI σε τέτοιους αλγόριθμους ή οι αλγόριθμοι θα πρέπει να τροποποιηθούν. Η προσέγγιση που ακολουθήθηκε μπορεί να χαρακτηριστεί ως ένα στάδιο προεπεξεργασίας όπου η κατεστραμμένη εικόνα MRI αποκαθίσταται διαιρώντας την με ένα εκτιμώμενο σήμα πεδίου bias field χρησιμοποιώντας μια προσέγγιση επιφανειακής τοποθέτησης. Χρησιμοποιήθηκε το πρόγραμμα N4 Bias Correction Toolkit [30].

Στη συνέχεια εφαρμόζουμε τα ακόλουθα βήματα προεπεξεργασίας: (i) ομαλοποίηση έντασης κάθε τρισδιάστατου όγκου, x, χρησιμοποιώντας ομαλοποίηση min-max (ii) επαναδειγματοληψία όλων των 2D εικόνων και των αντίστοιχων ετικετών (masks), σε ένα σταθερό μέγεθος εικονοστοιχείου, χρησιμοποιώντας παρεμβολές bi-linear και nearest-neighbour αντίστοιχα, ακολουθούμενη από περικοπή ή πρόσθεση της τιμής 0 στην εικόνα ώστε όλες να προκύψουν σε ένα σταθερό μέγεθος εικόνας. Δεν χρειάστηκε να χρησιμοποιήσουμε ένα εξωτερικό εργαλείο για να ευθυγραμμίσουμε τους τόμους σε οποιοδήποτε από τα σύνολα δεδομένων, ήταν ήδη σχεδόν ευθυγραμμισμένα καθώς αποκτήθηκαν.

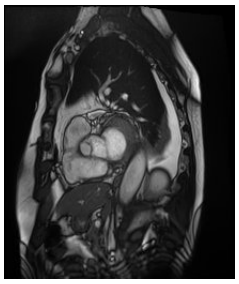


Figure 3: Original image.

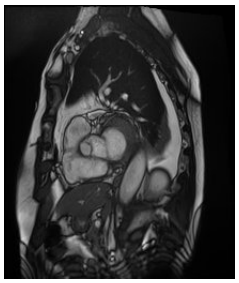


Figure 4: Noise corrected image.

IV. Μεθοδολογία

Η χρήση εποπτευόμενης βαθιάς μάθησης (supervised deep learning) μπορεί να πετύχει πολύ υψηλές αποδόσεις στο image segmentation πάνω σε σύνολα δεδομένων ιατρικής απεικόνισης. Αυτό, όμως, έχει ως προϋπόθεση την

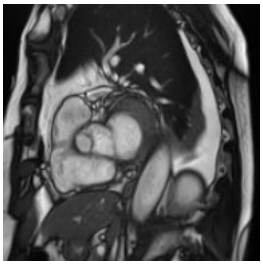


Figure 5: Noise corrected-cropped image

ύπαρξη μεγάλων επισημασμένων σετ δεδομένων. Η επισήμανση ενός μεγάλου συνόλου παραδειγμάτων από ιατρούς είναι χρονοβόρα και κοστοβόρα. Όπως αναφέρθηκε προηγουμένως, η αναμονή για τη δημιουργία επισημασμένων δεδομένων αποτελεί τροχοπέδη για τη δημιουργία και την ανάπτυξη τρεχόντων αλγορίθμων βαθιάς μάθησης. Αυτό έχει ως αποτέλεσμα να μην μπορούν να εκμεταλλευτούν στο έπακρο τις δυνατότητες των τεχνικών αυτών, τόσο ο ακαδημαϊκός χώρος όσο και ο αμιγώς ιατρικός. Έτσι, είναι ζωτικής σημασίας η ανάπτυξη αλγορίθμων που μπορούν να έχουν υψηλή απόδοση έχοντας στη διάθεσή τους μεγάλο μεν αριθμό δεδομένων αλλά μικρό δε αριθμό επισημασμένων δεδομένων.

A. Συγκριτική Μάθηση *(Contrastive Learning)*

Η απάντηση σε αυτό το πρόβλημα είναι η χρήση τεχνικών συγκριτικής μάθησης, οι οποίες βασίζονται στην σύγκριση μεταξύ των δειγμάτων ώστε να εξαχθεί πληροφορία για τα συνολικά δεδομένα που υπάρχουν στη διάθεσή μας. Η σύγκριση μεταξύ των δεδομένων δεν χρειάζεται την ύπαρξη ετικετών, καθώς συγκρίνεται η μορφή των αναπαραστάσεων των δεδομένων και εξετάζονται κατά πόσο μοιάζουν ή όχι.

Η διαδικασία αυτή λειτουργεί ως μια μορφή clustering στην οποία οι αναπαραστάσεις όμοιων δεδομένων (ίδια ετικέτα και ας μην την έχουμε στη διάθεσή μας) βρίσκονται κοντά στον χώρο αναπαράστασης ενώ ανόμοια δεδομένα αναπαρίστανται μακριά στο χώρο. Οι αναπαραστάσεις αυτές προκύπτουν με τη χρήση ενός encoder δικτύου, μετατρέποντας την αρχική εικόνα σε features αναπαράστασης, πάνω στα οποία εφαρμόζεται η σύγκριση.

Ένα διαισθητικό παράδειγμα μπορεί να δοθεί εξετάζοντας την περίπτωση διαχωρισμού μιας εικόνας αν περιέχει σκύλο ή γάτα. Χρησιμοποιώντας για όλες τις εικόνες την ίδια απεικονιστική διαδικασία μετατρέπονται οι αρχικές εικόνες σε διανύσματα χαρακτηριστικών. Είναι αναμενόμενο, πως αν και διαφορετικές μεταξύ τους, οι εικόνες των σκύλων να απεικονίζονται πιο κοντά από ότι αυτές των γατών. Έτσι, καθίσταται ως ένα βαθμό ένας διαχωρισμός μεταξύ των κλάσεων με τη λογική του clustering χωρίς να διαθέτουμε στη διάθεσή μας ετικέτες. Μπορούμε να πούμε πως έχουμε ανακαλύψει διαφορετικά υπό εξέταση αντικείμενα τα οποία διαφέρουν μεταξύ τους αλλά δεν γνωρίζουμε τι ακριβώς είναι αυτά τα αντικείμενα.

Τέλος, εφόσον έχουμε κάνει τις ομαδοποιήσεις στα δεδομένα μας, χρησιμοποιούμε τα λίγα επισημασμένα δεδομένα

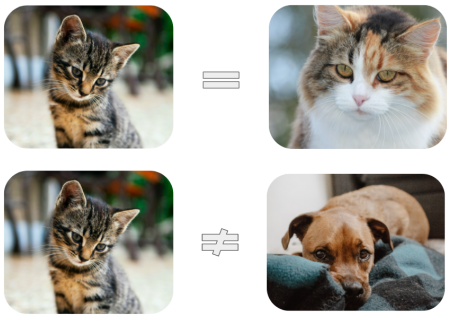


Figure 6: Παρόμοιες αναπαραστάσεις για εικόνες με παρόμοιο περιεχόμενο.

που έχουμε στη διάθεσή μας για φέρουμε το μοντέλο στην τελική του μορφή. Διαισθητικά πάλι πάνω στο παράδειγμα με τις γάτες και τους σκύλους, έχοντας στην διάθεσή μας δυο διακριτά clusters, μπορούμε να επισημάνουμε όλα τα δεδομένα του κάθε cluster με βάση τα επισημασμένα δεδομένα που πλέον ανατίθενται στο κάθε cluster. Δηλαδή, εξετάζοντας τις επισημασμένες εικόνες των γατών, μπορούμε να δούμε πως βρίσκονται εντός ή πολύ κοντά σε ένα από τα δύο αρχικά clusters. Άρα, μπορούμε να συμπεράνουμε πως όλο αυτό το cluster περιέχει απεικονίσεις γατών και αντίστοιχα το άλλο περιέχει σκύλους.

B. Περιγραφή Γενικής Μεθοδολογίας

Η προσέγγισή ακολουθεί την λογική που εξηγήθηκε προηγουμένως. Χρησιμοποιείται ένας CNN-encoder για να παραχθούν οι αρχικές αναπαραστάσεις από μια εικόνα και στη συνέχεια αυτές οι αναπαραστάσεις τροφοδοτούνται σε ένα μικρό projection δίκτυο για περαιτέρω συμπύκνωσή τους και την παραγωγή των δευτερευουσών αναπαραστάσεων. Στην εικόνα 7 φαίνεται η συνολική λογική που ακολουθείται.

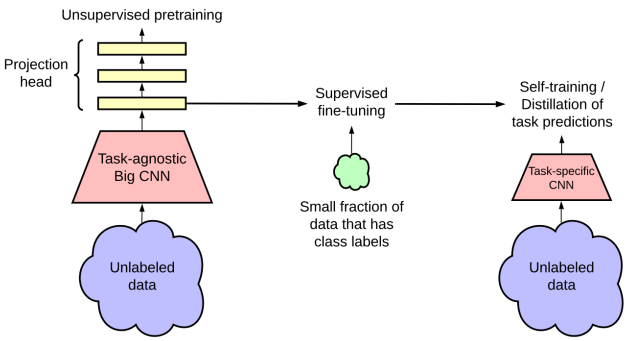


Figure 7: Δομικά στοιχεία εκπαίδευσης.

Όπως φαίνεται και από την εικόνα τα δεδομένα είναι στο μεγαλύτερο βαθμό μη επισημασμένα (μπλε χρώμα), ενώ μόνο ένα μικρό μέρος τους έχει ετικέτες (πράσινο χρώμα). Το ροζ τραπέζιο αποτελεί τον CNN-encoder ο οποίος είναι υπεύθυνος για την εξαγωγή των βασικών χαρακτηριστικών.

Τα κίτρινα στρώματα αποτελούν το projection μέρος που χρησιμοποιείται μόνο κατά το πρώτο μέρος της εκπαίδευσης και στην εφαρμογή της σύγκρισης πάνω στα μη επισημασμένα δεδομένα.

Εφόσον γίνει αυτή η αρχική εκπαίδευση, διώχνουμε το projection δευτερεύουσας αναπαράστασης και χρησιμοποιούμε κάποιο άλλο δίκτυο της επιλογής μας. Στην περίπτωσή μας, επειδή θέλουμε να παράγουμε έξοδο μια εικόνα-μάσκα με τις περιοχές ενδιαφέροντος, χρησιμοποιούμε έναν Decoder, ώστε η παραγόμενη εικόνα να είναι ίδιου μεγέθους με την αρχική. Στη συνέχεια, χρησιμοποιώντας τα επισημασμένα δεδομένα γίνεται το fine-tuning του μοντέλου. Τέλος, έχουμε στα χέρια μας ένα εκπαιδευμένο μοντέλο πάνω σε δεδομένα τα οποία στην πλειονότητά τους είναι μη επισημασμένα.

C. Περιγραφή Γενικής Αρχιτεκτονικής Δικτύου

Ως δίκτυο για την εφαρμογή της τεχνικής χρησιμοποιήθηκε το Unet [31], το οποίο έχει την γενική δομή ενός αυτοκωδικοποιητή (autoencoder) όπως φαίνεται στην εικόνα 8, όπου e είναι ο Encoder και d ο Decoder του δικτύου.

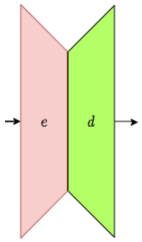


Figure 8: Γενική αρχιτεκτονική autoencoder.

Ο encoder έχει ως στόχο να εξάγει τις σημαντικότερες πληροφορίες της εισόδου σε μια συμπυκνωμένη μορφή, η οποία ονομάζεται feature map. Στη συνέχεια, ο decoder πρέπει από αυτή τη συμπυκνωμένη μορφή πληροφορίας να ανακατασκευάσει την εικόνα όσο καλύτερα μπορεί. Γενικά, η ανακατασκευή αυτή μπορεί να είναι πλήρης, δηλαδή να πρέπει η τελική παραγόμενη εικόνα να είναι ίδια με την αρχική ή μπορεί να είναι μια εικόνα-χάρτης, στην οποία να έχουν επισημανθεί διάφορες περιοχές με κάποιο label (image segmentation). Στην περίπτωσή μας θέλουμε τελικώς το δίκτυο να πραγματοποιεί τη δεύτερη λειτουργία. Χρησιμοποιώντας κατάλληλες συναρτήσεις κόστους οι οποίες θα εξηγηθούν στη συνέχεια, θα εκπαιδεύσουμε ξεχωριστά ορισμένα μέρη του δικτύου.

D. Καθολική *(Global)* και Τοπική *(Local)* Προσέγγιση

Για την παραγωγή καλύτερων αποτελεσμάτων γίνεται χρήση τόσο μιας καθολικής contrastive loss, η οποία λαμβάνει συνολικά υπόψη της όλη την εικόνα (global), όσο και μιας τοπικής (local), η οποία προσπαθεί να εξάγει τοπικά χαρακτηριστικά από την εικόνα. Διαισθητικά, η πρώτη επεξεργάζεται και βγάζει συμπεράσματα για όλη την εικόνα ενώ η δεύτερη προσπαθεί να κατηγοριοποιήσει τα διάφορα μέρη εντός της εικόνας. Στη συνέχεια θα παρουσιαστούν και οι δύο προσεγγίσεις.

1) Global Contrastive Loss: Με τη χρήση του Global Contrastive loss θέλουμε να εκπαιδεύσουμε τον encoder της αρχιτεκτονικής της εικόνας 8. Για να γίνει αυτό εξάγουμε μόνο το e μέρος και στο τέλος του ενώνουμε ένα shallow fully connected δίκτυο το οποίο αναφέρεται και ως projection head. Ο σκοπός αυτού του τελικού πλήρως συνδεδεμένου δικτύου είναι να δώσει μια παραπάνω ευελιξία στις αναπαραστάσεις που παράγει ο encoder. Η μορφή του δικτύου αυτού του βήματος φαίνεται στην εικόνα 9, όπου e είναι ο encoder και g1 το fully-connected δίκτυο.

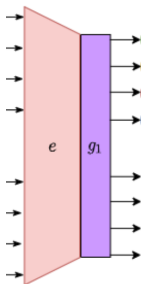


Figure 9: Αρχιτεκτονική για εκπαίδευση με global contrastive loss.

Μια άλλη πιο μαθηματική απεικόνιση του δικτύου αυτού του βήματος είναι αυτή της εικόνας 10 με τον Encoder $f(\cdot)$ και το projection $g1$.

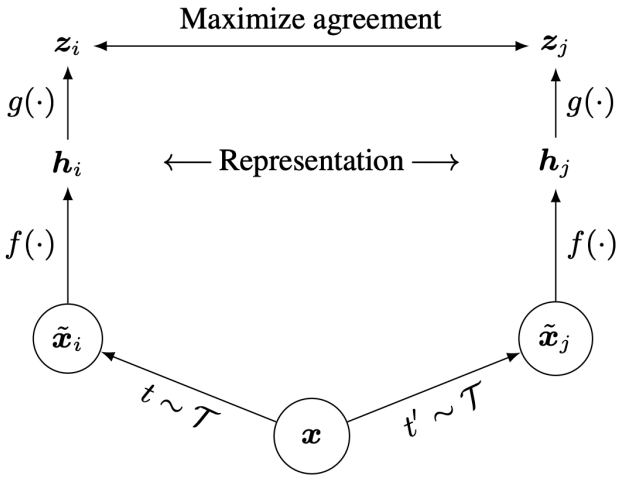


Figure 10: Διαδικασία σύγκρισης εικόνων.

Με βάση την παραπάνω εικόνα μπορούμε να αναγράψουμε πλέον και τη μαθηματική έκφραση της συνάρτησης κόστους.

$$l(\widetilde{x},\hat{x}) = -\log \frac{e^{sim(\widetilde{z},\hat{z})/\tau}}{e^{sim(\widetilde{z},\hat{z})/\tau} + \sum_{\overline{z}\in\Lambda^-} e^{sim(\widetilde{z},g(f(\overline{x}))/\tau}},$$

$$\widetilde{z} = g(f(\widetilde{x})), \hat{z} = g(f(\hat{x}))$$

Οι εικόνες $\widetilde{x}, \hat{x}$ είναι παράγωγες της x έχοντας εφαρμόσει τυχαίο αριθμό από απλούς μετασχηματισμούς t,t' και πάνω

στις οποίες γίνεται η βασική σύγκριση. Τέτοιοι μετασχηματισμοί είναι οι εξής:

- Τυχαία περικοπή (Random Cropping)
- Τυχαία περιστροφή (Random Rotation)
- Τυχαία αλλαγή χρώματος/φωτεινότητας κλπ. (Random Jittering)
- Καθρέφτισμα (Mirroring)

Συνολικά μπορούμε να δούμε πως η συνάρτηση αυτή δείχνει για μια εικόνα $\widetilde{x}$ ποια είναι η συγκριτική ομοιότητα της με την εικόνα $\hat{x}$ λαμβάνοντας υπόψη και τη συγκριτική ομοιότητα της $\widetilde{x}$ με όλες τις άλλες εικόνες που είναι αρνητικά ζεύγη της και περιέχονται στο σύνολο Λ^-. Εν δυνάμει σύνολο Λ^- αποτελείται από έναν αριθμό από όλες τις άλλες εικόνες και τις παράγωγές τους εκτός της x και των δικών της παραγώγων. Επίσης, ο παράγοντας τ αποτελεί μια σταθερά κλιμάκωσης.

Η συνάρτηση sim αποτελεί το cosine similarity δυο διανυσμάτων που δίνεται από την εξίσωση:

$$sim(a,b) = \frac{a^T b}{\|a\|\|b\|}$$

Για να είναι μικρό το κόστος θέλουμε το similarity μεταξύ των $\widetilde{z}, \hat{z}$ να είναι μεγάλο και ταυτόχρονα το similarity μεταξύ όλων των ζευγών εικόνων στο άθροισμα να είναι μικρό ώστε το κλάσμα να τείνει στη μονάδα και άρα το κόστος στο μηδέν. Διαισθητικά, αυτή η συνάρτηση κόστους αυξάνει την ομοιότητα μεταξύ των θετικών ζυγών και μειώνει την ομοιότητα των αρνητικών ζευγών. Για να βρούμε το συνολικό global contrastive loss για όλες τις εικόνες μας εφαρμόζουμε την επόμενη εξίσωση.

$$L = \frac{1}{|\ \Lambda^+\ |} \sum_{\forall(\widetilde{x},\hat{x})\in\Lambda^+} [l(\widetilde{x},\hat{x}) + l(\hat{x},\widetilde{x})]$$

όπου το Λ^+ είναι για μια εικόνα το σύνολο όλων των εικόνων που έχουν παραχθεί μετά τους μετασχηματισμούς της αρχικής εικόνας καθώς και η ίδια η αρχική εικόνα.

Δουλεύοντας με mini-batches, τα θετικά ζευγάρια κάθε εικόνας είναι η αρχική εικόνα και οι παράγωγές της από τους μετασχηματισμούς που αναφέρθηκαν, ενώ τα αρνητικά ζευγάρια είναι όλες οι υπόλοιπες εικόνες του mini-batch καθώς και οι παράγωγές τους.

Στην περίπτωσή μας, ασχολούμαστε με ιατρικές εικόνες οι οποίες είναι κύβοι δεδομένων, στον x,y άξονα είναι η διαστάσεις της εικόνας (αναφέρεται και ως τομή) ενώ στον z άξονα είναι το βάθος της εικόνας. Το γεγονός πως οι ακτινογραφίες μια ανατομικής περιοχής έχουν ληφθεί με μεγάλη προσοχή και έχουν προ-επεξεργαστεί κατάλληλα (ευθυγραμμισμός), δίνει τη δυνατότητα να θεωρήσουμε πως κάθε 3D εικόνα τόσο στους x,y άξονες (τομή ή 2D εικόνα ή απλούστερα θα αναφέρεται ως εικόνα) όσο και στο z απεικονίζουν τις ίδιες περιοχές ενδιαφέροντος. Με αυτή τη θεώρηση στη συνέχεια θα εξεταστούν διαφορετικές προσεγγίσεις στη δημιουργία των συνόλων Λ^+ και Λ^- που έχουν αναφερθεί.

Αυτή η ευθυγράμμιση μας δίνει τη δυνατότητα να θεωρήσουμε ως όμοια δεδομένα τις εικόνες που αντιστοιχούν σε ίδιες τομές (πχ πρώτη τομή στον z άξονα). Με αυτόν τον τρόπο έχουμε μια ευελιξία στο πως θα θεωρήσουμε τα σύνολα θετικών και αρνητικών ζευγών εικόνων στη συνέχεια.

Έστω πως έχουμε στη διάθεσή μας M 3D εικόνες, οι οποίες αποτελούνται από Q τομές οι οποίες είναι ευθυγραμμισμένες για όλες τις M 3D εικόνες. Ομαδοποιώντας τομές ανά έναν αριθμό $D < Q$ κάθε 3D εικόνα αποτελείται πλέον από S διαμερίσεις. Δηλαδή, αυτές οι διαμερίσεις S αποτελούνται από D συνεχόμενες τομές της αρχικής 3D εικόνας, ενώ συνδυάζοντάς τες όλες λαμβάνουμε την αρχική 3D εικόνα με τις Q τομές. Η ευθυγράμμιση που έχει αναφερθεί μας δίνει τη δυνατότητα να θεωρούμε πως όλες οι εικόνες μιας διαμέρισης S δείχνουν ίδιο ιατρικό αντικείμενο και άρα θεωρούνται θετικά δείγματα μεταξύ τους. Για διευκόλυνση των συμβολισμών, ως δύναμη γράφουμε τον αριθμό της 3D εικόνας ενώ ως βάση τον αριθμό της διαμέρισης, δηλαδή η έκφραση x_s^i αντιστοιχεί σε μια εικόνα της s διαμέρισης της i-οστής 3D εικόνας. Για περαιτέρω κατανόηση των προηγούμενων εννοιών μπορούμε να συμβουλευτούμε την εικόνα 11, όπου φαίνονται οπτικά οι συμβολισμοί(ως volume αναγράφεται μια 3D εικόνα). Επίσης, βλέπουμε πως για κάθε διαμέριση S το χρώμα είναι ίδιο ανάμεσα στις 3D εικόνες, δηλαδή απεικονίζουν ίδια περιοχή, το οποίο ακριβώς ισχύει και για όλες τις εικόνες της κάθε διαμέρισης S.

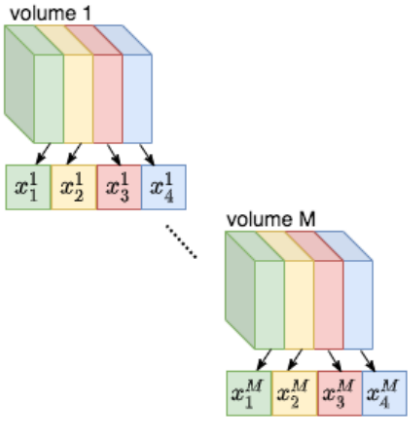


Figure 11: 3D εικόνες και διαμερίσεις τους.

Δεδομένης της ευθυγράμμισης που έχει αναφερθεί προηγουμένως, μπορούμε να πούμε πως η ίδια διαμέριση δυο 3D εικόνων δείχνουν σε ίδιο ανατομικό σημείο και άρα πως είναι όμοιες μεταξύ τους.

Επομένως, μπορούμε να διακρίνουμε 3 διαθέσιμες στρατηγικές για τη δημιουργία των Λ^+ και Λ^- σε ένα batch δεδομένων εικόνων 3D.

1) Στρατηγική G^R: Η στρατηγική αυτή είναι η απλοϊκή προσέγγιση που αναλύθηκε προηγουμένως. Δηλαδή, έχοντας ένα batch N εικόνων τυχαία επιλεγμένες από όλες τιw 3D εικόνες, εφαρμόζουμε σε κάθε εικόνα

x_s^i τυχαίους μετασχηματισμούς, όπως αυτοί που έχουν προαναφερθεί, και παράγουμε 2 καινούριες εικόνες $(\widetilde{x}_s^i, \hat{x}_s^i)$ από την αρχική. Άρα, πλέον έχουμε συνολικά $2N$ παραχθείσες εικόνες. Το σύνολο δεδομένων Λ^+ αποτελείται από τις δύο αυτές εικόνες, ενώ το Λ^- αποτελείται από τις υπόλοιπες $2N-2$.

Στις επόμενες δύο στρατηγικές για να δημιουργήσουμε επιλέγουμε τυχαία m 3D εικόνες από τις συνολικά M που έχουμε στη διάθεσή μας. Στη συνέχεια, δειγματοληπτούμε μια εικόνα ανά διαμέριση S, δηλαδή έχουμε S εικόνες ανά 3D εικόνα. Τέλος, εφαρμόζουμε τυχαίους μετασχηματισμούς σε κάθε εικόνα x_s^i παράγοντας δυο νέες εικόνες $(\widetilde{x}_s^i, \hat{x}_s^i)$, τις οποίες προσθέτουμε στο batch που θα χρησιμοποιήσουμε.

2) Στρατηγική G^{D-}: Σε αυτή τη στρατηγική επεκτείνουμε το σύνολο των θετικών εικόνων ώστε να περιέχει εκτός των 2 παραγόμενων $(\widetilde{x}_s^i, \hat{x}_s^i)$ μέσω των μετασχηματισμών και την αρχική εικόνα x. Επομένως, το σύνολο θετικών ζευγών Λ^+ περιέχει τα τρία ζεύγη $(x_s^i, \hat{x}_s^i)$, $(x_s^i, \widetilde{x}_s^i)$ ενώ επιβάλουμε το αρνητικό σύνολο Λ^- να αποτελείται από τις εικόνες $\{x_k^l, \widetilde{x}_k^l, \hat{x}_k^l | \forall k \neq s, \forall l\}$, οι οποίες δηλαδή είναι οι εικόνες και οι παραγόμενές τους από οποιαδήποτε διαμέριση εκτός του s από κάθε 3D εικόνα.
3) Στρατηγική G^D: Σε αυτή τη στρατηγική επεκτείνουμε την προηγούμενη καθώς θεωρούμε πως οι εικόνες που προέρχονται από την ίδια διαμέριση αλλά από διαφορετικές 3D εικόνες αποτελούν θετικά ζεύγη, καθώς περιέχουν ίδια συνολική πληροφορία. Άρα, το θετικό σύνολο ζευγών αποτελείται από τις εικόνες της προηγούμενης στρατηγικής και επιπλέον με τα ζεύγη (x_s^i, x_s^j) αλλά και από τις παραγόμενες εικόνες τους $(\hat{x}_s^i, \widetilde{x}_s^j)$.

Η λογική της τελευταίας στρατηγικής (Στρατηγική G^D) φαίνεται στην ακόλουθη εικόνα.

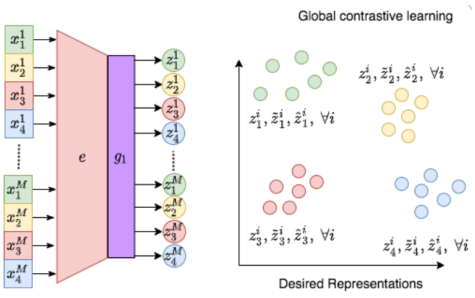


Figure 12: Αναπαραστάσεις στρατηγικής G^D) με global contrastive loss.

Βλέπουμε πως εικόνες από τις ίδιες διαμερίσεις έχουν κοντινές αναπραστάσεις σε αντίθεση με αυτές από διαφορετικές διαμερίσεις.

2) Local Contrastive Loss: Το global contrastive loss που αναπτύχθηκε στην προηγούμενη παράγραφο συγκρίνει τις αναπαραστάσεις μιας εικόνας συνολικά ώστε να βρει αν είναι όμοιες ή ανόμοιες ως ολόκληρη εικόνα. Αυτό είναι χρήσιμο όταν θέλουμε να αναγνωρίσουμε ποιο είναι το βασικό αντικείμενο που περιέχεται σε μια εικόνα ή να την κατηγοριοποιήσουμε. Για να μπορέσουμε να κατηγοριοποιήσουμε σε επίπεδο εικονοστοιχείων και να πραγματοποιήσουμε image segmentation είναι απαραίτητο να εφαρμοστεί και μια πιο τοπική προσέγγιση στα δεδομένα εικόνας που έχουμε στην διάθεσή μας. Επιπλέον, ένας πιθανός συνδυασμών των δύο μεθόδων να δίνει ακόμα καλύτερα αποτελέσματα.

Με αυτή την λογική, εξετάζεται προσέγγιση στην οποία ωθούμε το δίκτυο encoder-decoder να εξάγει χαρακτηριστικά και να συγκρίνει τοπικά υποπεριοχές των εικόνων. Για να συμβεί αυτό η μορφή του δικτύου που εκπαιδεύουμε με το Local Contrastive Loss φαίνεται στην εικόνα 13.

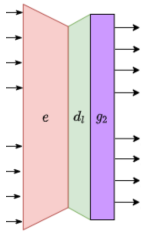


Figure 13: Αρχιτεκτονική για εκπαίδευση με local contrastive loss.

Στο σχήμα της προηγούμενης εικόνας τα πρώτα l στρώματα του decoder d_l εκπαιδεύονται με τη συνάρτηση κόστους L_l. Σε αυτή την προσέγγιση οι μετασχηματισμοί που εφαρμόζονται στις εικόνες είναι μόνο στην φωτεινότητα της και όχι μετατοπίσεις, περικοπές ή περιστροφές καθώς δεν μπορεί να διατηρηθεί η τοπικότητα των υποπεριοχών της εικόνας ώστε να γίνει σύγκριση.

Για μια εικόνα x το d_l μέρος προσπαθεί να απομακρύνει μεταξύ τους τις αναπαραστάσεις διαφορετικών περιοχών της εικόνας ενώ αυτές που αντιστοιχούν στην ίδια περιοχή να έχουν παρόμοιες αναπαραστάσεις ανεξάρτητα από τον μετασχηματισμό που τους εφαρμόζεται. Οι περιοχές αυτές που παράγει το d_l είναι $d_l(x) \in \mathbb{R}^{W_1 \times W_2 \times C}$, όπου W_1, W_2 οι διαστάσεις της υποπεριοχής και C τα κανάλια.

Η λογική της συνάρτησης κόστους είναι παρόμοια με αυτή της Global Contrastive Loss αλλά τροποποιημένη κατάλληλα ώστε να εφαρμόζεται στις υποπεριοχές της εικόνας. Θα συμβολίζουμε με Ω^+ και Ω^- τις θετικές και αρνητικές υποπεριοχές για να τις διαχωρίζουμε από τα σύνολα Λ που αναγράφονται προηγουμένως.

Η διαδικασία που ακολουθείται είναι η εξής, ένα ζεύγος όμοιων εικόνων $(\widetilde{x}, \hat{x}$ το περνάμε από τον encoder (e), τα πρώτα l στρώματα του decoder (d_l) και τέλος, από ένα shallow fully connected δίκτυο g_2. Έτσι, λαμβάνουμε τις αναπαραστάσεις (feature maps) $\widetilde{f} = g_2(d_l(e(\widetilde{x})))$ και $\hat{f} = g_2(d_l(e(\hat{x})))$. Στη συνέχεια, χωρίζουμε κάθε feature map σε A υποπεριοχές με διαστάσεις $K \times K \times K \times C$, όπου $K < min(W_1, W_2)$. Κάθε ζεύγος αντίστοιχων περιοχών των $\widetilde{f}, \hat{f}$ σχηματίζουν το θετικό σύνολο Ω^+ ενώ όλες οι υπόλοιπες υποπεριοχές σχηματίζουν το Ω^-. Όλη αυτή η διαδικασία είναι εμφανής στην εικόνα 14.

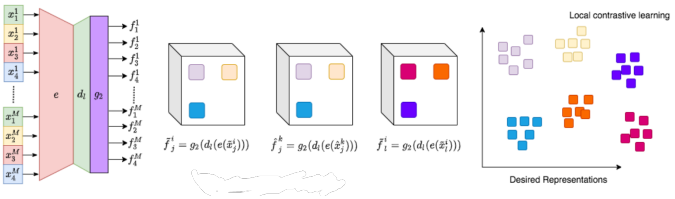


Figure 14: Μεθοδολογία local contrastive loss.

Η συνάρτηση για το Local Contrastive Loss φαίνεται στη συνέχεια.

$$l(\widetilde{x}, \hat{x}, u, v) =$$

$$-\log \frac{e^{sim(\widetilde{f}(u,v),\hat{f}(u,v))/\tau}}{e^{sim(\widetilde{f}(u,v),\hat{f}(u,v))/\tau} + \sum_{(u',v')\in\Omega^+} e^{sim(\widetilde{f}(u,v),\hat{f}(u',v'))/\tau}}$$

όπου το $sim(\cdot,\cdot)$ είναι το cosine similarity που έχει αναφερθεί προηγουμένως, και τα (u,v) είναι δείκτες για τις υποπεριοχές των feature maps και ισχύει $f(u,v) \in \mathbb{R}^C$. Για να βρούμε το συνολικό local contrastive loss για ένα σύνολο εικόνων $\mathbf{X}$ μας εφαρμόζουμε την επόμενη εξίσωση.

$$L_l = \frac{1}{|\mathbf{X}|} \sum_{x\in\mathbf{X}} \frac{1}{2A} \sum_{(u,v)\in\Omega^+} [l(\widetilde{x}, \hat{x}, u, v) + l(\hat{x}, \widetilde{x}, u, v)],$$

$$\widetilde{x} = \widetilde{t}(x), \hat{x} = \hat{t}(x), \quad \widetilde{t}, \hat{t} \sim T$$

όπου T είναι οι μετασχηματισμοί έντασης που εφαρμόζουμε σε μία εικόνα. Με αυτή την προσέγγιση έχουμε επεκτείνει την Global πορσέγγιση που αναφέρθηκε προηγουμένως και μπορεί η τεχνική να εφαρμοστεί σε pixel-prediction και image segmentation. Όπως και πριν, έχουν αναπτυχθεί κάποιες στρατηγικές για τη δημιουργία των συνόλων Ω^- και Ω^+. Στη συνέχεια, παρουσιάζονται αυτές οι δύο στρατηγικές.

1) Στρατηγική L^R: Στην στρατηγική αυτή δημιουργείται ένα mini-batch δειγματοληπτώντας N τομές (2D εικόνες) από όλες τις διαθέσιμες 3D εικόνες και τους εφαρμόζοντας τυχαίους μετασχηματισμούς έντασης. Δηλαδή, η διαδικασία δημιουργίας του θετικού συνόλου ζευγών Ω^+ γίνεται δειγματοληπτώντας εικόνες x_s^i και χρησιμοποιώντας ως θετικά ζεύγη τα $(\widetilde{f}_s^i(u,v), \hat{f}_s^i(u,v))$. Για κάθε θετικό ζεύγος το αρνητικό σύνολο δεδομένων Ω^- προκύπτει από όλες τις άλλες υποπεριοχές με δείκτες (u',v') με $u' \neq u,\ v' \neq v$, εντός των feature maps $(\widetilde{f}_s^i, \hat{f}_s^i)$.
2) Στρατηγική L^D: Όπως αναφέρθηκε και στο global contrastive loss θεωρούμε πως οι εικόνες είναι ευθυγραμμισμένες από την προεπεξεργασία που έχουν δεχθεί. Αυτή είναι και η βασική σκέψη για την στρατηγική αυτή, δηλαδή ανάμεσα σε δυο εικόνες που έχουν δεχθεί μόνο μετασχηματισμό έντασης οι υποπεριοχές τους είναι ευθυγραμμισμένες. Με αυτή τη λογική μπορούμε να δημιουργήσουμε θετικά ζεύγη υποπεριοχών παίρνοντας εικόνες που ανήκουν σε διαφορετικές 3D εικόνες, $(\widetilde{f}_s^i(u,v), \hat{f}_s^j(u,v))$, i,j διαφορετικές 3D εικόνες αλλά αναφερόμενοι στην ίδια διαμέριση s και στις 2. Όμοια, για το αρνητικό σύνολο Ω^- χρησιμοποιούμε τις εικόνες $(\widetilde{f}_s^i(u',v'), \hat{f}_s^j(u',v'))$ όπου $u' \neq u,\ v' \neq v$.

E. Προ-εκπαίδευση συνδυάζοντας *Global* και *Local Contrastive Losses*

Κατά την προεκπαίδευση του μοντέλου γίνεται χρήση τόσο της global όσο και της local contrastive loss. Αρχικά, γίνεται η εκπαίδευση με το global contrastive loss έχοντας στη διάθεσή μας τον encoder (e) και το fully connected projection head (g1). Όταν εκπαιδευτεί το δίκτυο κρατάμε μόνο τον εκπαιδευμένο encoder. Κάνοντας freeze τον encoder (e) και τον ενώνουμε με τα l πρώτα στρώματα του decoder (d_l) και με το projection head g2. Αυτό το καινούριο δίκτυο το εκπαιδεύουμε με το local contrastive loss. Αφού τελειώσει και αυτή η εκπαίδευση απομακρύνουμε το g2 και στα χέρια μας έχουμε το e με το d_l τα οποία αντίστοιχα έχουν εκπαιδευτεί να βρίσκουν αντίστοιχα global και local χαρακτηριστικά.

Τέλος, για να παράγει το δίκτυο εικόνα μεγέθους όσο η αρχική εικόνα εισόδου ενώνουμε τα υπόλοιπα στρώματα του decoder με τυχαία αρχικοποιημένα βάρη και εκτελούμε fine-tuning σε όλο το δίκτυο με τα λίγα επισημασμένα δεδομένα, σαν να υλοποιούσαμε κανονικά μια supervised learning διαδικασία.

V. Διαδικασία Αναπαραγωγής Πειραμάτων

Στο σημείο αυτό αναπαράγονται τα κυριότερα πειράματα, που ολοκληρώθηκαν από τους [32]. Στο συγκεκριμένο πρότυπο paper, πραγματοποιούνται πολλαπλά πειράματα σε τρία διαφορετικά σύνολα δεδομένων: το ACDC Dataset, το Prostate Dataset και το MMWHS Dataset. Επειδή οι συγγραφείς έχουν πραγματοποιήσει όλα τα πιθανά πειράματα και έχουν κάνει πολύ διεξοδική έρευνα για να καταλήξουν στα συμπεράσματά τους, η εκτέλεση όλων των πειραμάτων και από τη μεριά μας είναι αδύνατη και για το λόγο αυτό έχουμε επιλέξει να εκτελέσουμε τα κυριότερα και σημαντικότερα πειράματα, τα οποία έδωσαν και στο αρχικό paper, τα βέλτιστα αποτελέσματα για τα δύο πρώτα σύνολα δεδομένων: το ACDC dataset και το Prostate Dataset. Δυστυχώς στο τρίτο σύνολο δεδομένων δεν καταφέραμε να αποκτήσουμε πρόσβαση.

Αρχικά, τα δεδομένα χωρίζονται σε δεδομένα προεκπαίδευσης και σε δεδομένα ελέγχου, κάθε ένα από αυτά αποτελεί μία ογκομετρική εικόνα με τις αντίστοιχες ετικέτες τμηματοποίησης. Προ-εκπαιδεύουμε ένα UNet δίκτυο, χρησιμοποιώντας μόνο τις προ-επεξεργασμένες μαγνητικές τομογραφίες, χωρίς τις ετικέτες τους και στην συνέχεια πραγματοποιούμε το fine-tuning στο προ-εκπαιδευμένο δίκτυο με ένα μικρό αριθμό δεδομένων με ετικέτα, από το σύνολο δεδομένων προ-εκπαίδευσης. Το σύνολο δεδομένων ελέγχου, φυσικά δεν χρησιμοποιείται σε κανένα στάδιο της προ-εκπαίδευσης ή του fine-tuning, παρά μόνο στην τελική αξιολόγηση του μοντέλου. Τα μεγέθη των συνόλων αυτών

είναι: 52 το μέγεθος του συνόλου για την προ-εκπαίδευση και 20 το σύνολο ελέγχου για το ACDC dataset, 22 το σύνολο για την προ-εκπαίδευση και 15 το σύνολο ελέγχου για το σύνολο δεδομένων του προστάτη.

Για το fine-tuning, δημιουργούμε ένα σύνολο δεδομένων εκπαίδευσης και ένα σύνολο δεδομένων επικύρωσης, τα οποία αποτελούν υποσύνολα του συνόλου προ-εκπαίδευσης. Εκτελέσαμε πειράματα στα οποία το σύνολο εκπαίδευσης διατηρήθηκε στο μέγεθος 8, ενώ το σύνολο επικύρωσης διατηρήθηκε σε μέγεθος 2. Να σημειωθεί, ότι τα μεγέθη αναφέρονται ουσιαστικά στον αριθμό των διαφορετικών ασθενών και αποτελούνται από όλες τις εικόνες που προκύπτουν από μία μαγνητική τομογραφία.

Ως μέτρο αξιολόγησης, χρησιμοποιήθηκε το Dice Similarity Coefficient. Για τα πειράματα στο fine-tuning, καταγράφονται τα αποτελέσματα στο σύνολο ελέγχου. Για κάθε πείραμα, τα σύνολα εκπαίδευσης και επικύρωσης, κατασκευάστηκαν με τυχαία δειγματοληψία από το σύνολο προ-εκπαίδευσης. Έγινε η προσπάθεια να διατηρηθούν οι ίδιες συνθήκες όπως το batch size που προοτείνεται, ο optimizer κλπ.

Αρχικά, θα θελήσουμε να αναπαράγουμε το βασικό πείραμα που αναδεικνύει την προτεινόμενη μέθοδο. Η μέθοδος αυτή προκύπτει από διάφορα πειράματα με διαφορετικές στρατηγικές στους encoder και decoder προκειμένου να επικρατήσει ο συνδυασμός εκείνος με την βέλτιστη ακρίβεια. Οπότε θα τρέξουμε για Xtr=8 ένα πείραμα με την προ-εκπαίδευση να βασίζεται στην προαναφερθείσα στρατηγική G^D και ο decoder στην, επίσης προαναφερθείσα, στρατηγική L^R local contrastive loss. Ακόμα, για το δεύτερο πείραμα ο encoder προ-εκπαιδεύτηκε με βάση την στρατηγική G^R και ο decoder με βάση την στρατηγική L^D. Οι δύο αυτοί συνδιασμοί είναι οι επικρατέστεροι με βάση τα πειράματα στο προτότυπο paper. Σε όλα τα παραπάνω πειράματα, χρησιμοποιήθηκε σύνολο δεδομένων εκπαίδευσης μεγέθους 8 (τομογραφίες 8 διαφορετικών ασθενών,Xtr).

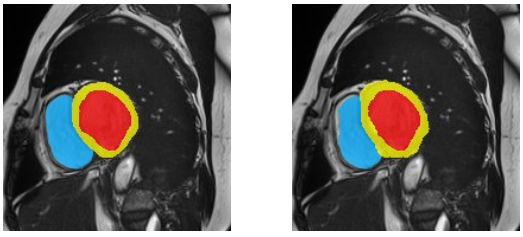
Στη συνέχεια θα επαναλάβουμε ένα συγκεκριμένο πείραμα από τα τα συγκριτικά πειράματα με διάφορες άλλες μεθόδους προεκπαίδευσης. Εκεί κατασκευάζεται μια baseline εκδοχή του δικτύου που έχει προ-εκπαιδευτεί με διάφορες data augmentation τεχνικές όπως τυχαίες περιστροφές κοψίματα κλπ. Έπειτα, αφού πραγματοποιηθεί το fine-tuning συγκρίνεται η τελική ακρίβεια του baseline με την προτεινόμενη μέθοδο και με διάφορες άλλες μεθόδους προ-εκπαίδευσης. Εμείς θα πραγματοποιήσουμε την baseline εκδοχή ώστε να την συγκρίνουμε με την "δική" μας προτεινόμενη μέθοδο.

Τέλος ένα πείραμα ακόμα θα αφορά την επιλογή του αριθμού των decoder blocks στο δίκτυο. Στο προτότυπο paper γίνεται μία εκτενής αναζήτηση για τον αριθμό των block στο decoder κομμάτι του δικτύου. Καταλήγουν στο ότι ο βέλτιστος αριθμός των μπλοκ αυτό είναι 3. Εμείς θα επαναλάβουμε το πείραμα για την προτεινόμενη μέθοδο (3 μπλοκ,G^D - L^R) με 2 και 4 μπλοκ αντίστοιχα, ώστε να επαληθεύσουμε τα αποτελέσματα. Τα αποτελέσματα όλων των πειραμάτων φαίνονται παρακάτω :

Table I

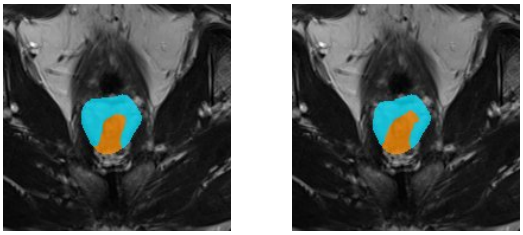
Method		**ACDC**	**Prostate**
Encoder	Decoder		
G^D	L^R	0.787	0.592
G^R	L^D	0.772	0.584
Baseline		0.742	0.569
Decoder,$l = 2$	G^D - L^R	0.765	0.581
Decoder, $l = 4$	G^D - L^R	0.770	0.578

Βλέπουμε πως τα πειράματα μας δεν κατάφεραν να παράξουν παρόμοιες ακρίβειες με την αρχική δημοσίευση. Παρόλα αυτά τα συμπεράσματα στα οποία οδηγούνται οι δημιουργοί επιβεβαιώνονται και από τα δικά μας πειράματα. Έτσι προκύπτει πως η προτεινόμενη μέθοδος είναι όντως αυτή με την καλύτερη ακρίβεια. Η στρατηγική G^D - L^R υπερτερεί των υπολοίπων. Επίσης φαίνεται πως η μέθοδος Contrastive learning έχει όντως αποτελεσματικότητα σε σχέση με μια baseline έκδοση του μοντέλου, αφού η προεκπαίδευση χαρίζει καλύτερη ακρίβεια. Τέλος το γεγονός πως ο βέλτιστος αριθμός των μπλοκ είναι 3 επιβεβαιώνεται από τα συγκρητικά πειράματα. Ακολουθούν ενδεικτικές εικόνες για τα δύο σετ δεδομένων με τις προβλέψεις και τα ground truth:



(a) Ground Truth (b) Prediction

Figure 15: Εικόνες ACDC



(a) Ground Truth (b) Prediction

Figure 16: Εικόνες Prostate

VI. Επέκταση Πειραμάτων

Το Contrastive learning αποτελεί μία επαναστατική self-supervised τεχνική. Ο τρόπος με τον οποίο λειτουργεί, καθώς και η όλη διαδικασία της προ-εκπαίδευσης των στρωμάτων του Unet δικτύου που χρησιμοποιείται στην δημοσίευση που αναφέρεται, είναι αρκετά συγκεκριμένος και χωρίς περιθώρια αλλαγών. Επίσης, στην αρχική δημοσίευση συγκρίνεται το Contrastive learning με άλλες μεθόδους όπως conditional generative models [33], την τεχνική mixup [34]

ή και adversarial learning [22]. Όλες αυτές οι τεχνικές αποτελούν self-supervised τεχνικές που εκμεταλέυονται τα δεδομένα χωρίς ετικέτα για να προ-εκπαιδεύσουν τις παραμέτρους στο εκάστοτε δίκτυο στο οποίο εφαρμόζονται. Συνεπώς κρίνεται μη δόκιμη η προσπάθεια παραλλαγής της όλης φιλοσοφίας του contrastive learning, καθώς θα αναφερόμαστε σε άλλη τεχνική και οι συγκρίσεις με τις state of the art αντίστοιχες τεχνικές έχουν ήδη πραγματοποιηθεί.

Σχετικά με την προοπτική δοκιμής της μεθόδου σε άλλου τύπου σετ δεδομένων πέρα από τα ιατρικά, παρατηρήθηκε πως οι διαφορές στις ακρίβειες μεταξύ της τυχαίας αρχικοποίησης των βαρών του δικτύου σε σχέση με την προ-εκπαίδευση με Contrastive learning είναι αρκετά μικρές, με βάση τα πειράματα στο σετ δεδομένων 'Cityscapes'. Στο σημείο αυτό αξίζει να αναφερθεί πως η επικρατούσα τεχνική (Στρατηγικές G^D-L^R) δεν μπορεί να εφαρμοστεί σε τέτοιου είδους σετ δεδομένων. Η στρατηγική G^D λαμβάνει υπ' όψιν την συσχέτιση μεταξύ των ιατρικών εικόνων (ανάλογα την γωνία λήψης, διαμερίσεις), ενώ στο 'Cityscapes' δεν υπάρχει κάποια τέτοια συσχέτιση. Έτσι χρησιμοποιείται η G^R στρατηγική και τα αποτελέσματα δεν φανερώνουν προοπτικές γενίκευσης της μεθόδου με την παρούσα μορφή σε προβλήματα εκτός την κατάτμησης των ιατρικών εικόνων.

Στο συγκεκριμένο σημείο αξίζει να αναφερθεί η [35] και το SimCLR , μία βασική δημοσίευση περί Contrastive learning που θέτει τα θεμέλια της μεθόδου και την εισάγει στο κοινό. Σε αυτήν προκύπτουν ορισμένα συμπεράσματα για την λειτουργία της μεθόδου, την επιλογή των υπερπαραμέτρων και άλλα χαρακτηριστικά. Με βάση αυτά θα γίνει η προσπάθεια να επεκταθούν ορισμένα πειράματα που έγιναν στην αρχική δημοσίευση και να γίνει περαιτέρω εξερεύνηση της μεθόδου.

Στην αρχική δημοσίευση η παράμετρος τ που βρίσκεται μέσα στην συνάρτηση κόστους παίρνει την τιμή 0.5 χωρίς αρκετή εξερεύνηση. Η αιτιολόγηση είναι πως στην εφαρμογή του SimCLR η καλύτερες ακρίβειες προέκυψαν με τιμές 0.1 και 0.5, και από τα μετέπειτα πειράματα των δημιουργών προέκυψε πως η τιμή 0.5 δίνει καλύτερα αποτελέσματα. Στην εφαρμογή του SimCLR παρατηρείται όμως μεγάλη απόκλιση της ακρίβειας ανάμεσα σε διάφορες τιμές του τ. Δεδομένου ότι το SimCLR χρησιμοποιεί ένα διαφορετικό δίκτυο (ResNet-50) καθώς και σε διαφορετικό σετ δεδομένων (Imagenet), η αυθαίρετη επιλογή του τ με βάση τα πειράματα του SimClR κρίνεται αδόκιμη. Συνεπώς το πρώτο πείραμα που πραγματοποιήσαμε είναι να δοκιμαστούν σε όλη την διαδικασία της εκπαίδευσης οι έξτρα τιμές 1 και 0.05 για την παράμετρο τ. Τελικά, τα αποτελέσματα δικαιώνουν τους δημιουργούς της αρχικής δημοσίευσης :

Table II

T	ACDC	Prostate
0.05	0.768	0.588
1	0.776	0.572
0.5(ours)	0.787	0.592
0.5(authors)	0.872	0.684

Άλλη μια σημαντική παράμετρος που εξετάζεται στην αρχική δημοσίευση αλλά και στην εφαρμογή του SimCLR είναι το batch size. Εξετάζοντας το μεγαλύτερο σετ δεδομένων που έχουμε στην διάθεση μας, το ACDC , παρατηρούμε ότι έχουμε 100 ασθενείς x 10 γωνίες μαγνητικής τομογραφίας = 1000 2D εικόνες. Στην αρχική δημοσίευση παρατηρείται από τα πειράματα πως άσχετα με τον αριθμό των ασθενών των οποίων τα δεδομένα θα χρησιμοποιηθούν για το fine-tuning του δικτύου (Xtr), οι ακρίβεια φθίνει σε σχέση με την αύξηση του batch size. Στην περίπτωση του SimCLR δείχνεται πως το μεγαλύτερο δυνατό batch size φέρνει και καλύτερες ακρίβειες. Σε μεγάλους αριθμούς εποχών το φαινόμενο αυτό εξασθενεί σχετικά με τις διαφορές στην απόδοση, αλλά οι διαφορές συνεχίζουν να υφίστανται.

Αυτή η ασυμφωνία στην συμπεριφορά της μεθόδου ίσως να οφείλεται στην διαφορά των σετ δεδομένων, δηλαδή τις εικόνες MRI σε σχέση με τις εικόνες του ImageNet. Παρόλα αυτά υπάρχει η εντύπωση ότι δεν έγιναν τα απαραίτητα πειράματα ώστε να διασφαλισθεί το συμπέρασμα ότι στο σετ δεδομένων ACDC πρέπει να προτιμηθούν τα μικρά batch size. Οπότε πραγματοποιήθηκαν πειράματα για Xtr=8 και επιπλέον τιμές του batch size ώστε να επικυρωθεί η συμπεριφορά του μοντέλου. Τα αποτελέσματα που προέκυψαν και την επικυρώνουν είναι τα εξής :

Table III

Batch Size	80	150	500
ACDC	0.771	0.765	0.752

Σαν τελευταίο στάδιο των έξτρα πειραμάτων, θέλουμε να εξετάσουμε την σχέση αρχιτεκτονικής και προ-εκπαιδευσης. Συγκεκριμένα θα επιχειρήσουμε να αυξήσουμε το βάθος και το πλάτος του δικτύου που χρησιμοποιείται για να δούμε την συμπεριφορά του στην προ-εκπαίδευση. Η self-supervised εκδοχή της εκπαίδευσης των παραμέτρων του δικτύου με Contrastive learning, ακολουθεί τις ίδιες αρχές με τις τυπικές εκπαιδεύσεις νευρωνικών δικτύων; Η αύξηση των παραμέτρων προς εκπαίδευση, θα απαιτήσει περισσότερα δεδομένα ώστε να εκπαιδευτούν σωστά οι παράμετροι; Τα μεγάλα δίκτυα οδηγούν σε overfitting κατά την διάρκεια της προ-εκπαίδευσης με self-supervised τεχνικές;

Για να απαντηθούν τα ερωτήματα αυτά επιχειρήσαμε να μεταποιήσουμε την αρχιτεκτονική του δικτύου και να επαναλάβουμε το βασικό πείραμα. Το δίκτυο που χρησιμοποιεί η αρχική δημοσίευση είναι ένα τυπικό δίκτυο Unet. Αυτό αποτελείται από 6 συνελικτικά μπλοκ, όπου το καθένα αποτελείται από δύο 3 x 3 συνελίξεις ακολουθούμενες από ένα 2 x 2 maxpooling στρώμα με βήμα 2. Το fully connected projection head (g1) απαρτίζεται από 2 dense στρώματα 3200 και 128 αντίστοιχα. Σχετικά με τον decoder δεν θα αλλάξουμε κάτι διότι έχουν γίνει ήδη πειράματα πάνω στον αριθμό των μπλοκ. Έτσι θα αυξήσουμε των αριθμό των συνελικτικών μπλοκ από 6 σε 10 όπως και τον αριθμό των φίλτρων στα στρώματα αυτά για να αυξήσουμε το capacity του δικτύου. Το ίδιο θα κάνουμε και στο projection head, στο οποίο θα διπλασιαστούν οι νευρώνες.

Επαναλαμβάνοντας το βασικό πείραμα του segmentation του ACDC σετ δεδομένων με 8 ασθενείς για fine-tuning, το

αποτέλεσμα που προέκυψε έδωσε αντίστοιχη ακρίβεια με το αρχικό πείραμα που περιλάμβανε το αρχικό δίκτυο με τα 6 συνελικτικά μπλοκ κλπ. Η διαφορά που εμφανίστηκε ήταν ότι ο χρόνος για την εκπαίδευση ήταν μεγαλύτερος στην περίπτωση του μεγάλου δικτύου, χωρίς όμως να υπάρξει κάποια ουσιαστική συνέπεια στην τελική ακρίβεια. Ένα πρώτο συμπέρασμα που προκύπτει από το πείραμα αυτό είναι ότι και στην περίπτωση του self-supervised learning οι τυπικοί κανόνες της μάθησης παραμένουν ως έχουν. Δηλαδή τα μεγάλα δίκτυα απαιτούν και αρκετά δεδομένα για να εκπαιδευτούν σωστά. Παρόλο που το δίκτυο μεγάλωσε, δεν κατόρθωσε μεγαλύτερες ακρίβειες καθώς οι έξτρα παράμετροι που εισήχθησαν, είτε δεν εκπαιδεύτηκαν, είτε δεν ανταποκρίνονται στις ανάγκες του συγκεκριμένου προβλήματος.

VII. Μελλοντικές Κατευθύνσεις και Συμπεράσματα

Η απαίτηση μεγάλου αριθμού επισημασμένων εικόνων για την εφαρμογή μεθόδων βαθιάς μάθησης παραμένει μια επίμονη πρόκληση στην ανάλυση των ιατρικών εικόνων. Στην παρούσα εργασία, προκειμένου να αντιμετωπιστεί η ανάγκη για μεγάλα επισημασμένα σύνολα δεδομένων εκπαίδευσης, προτείνεται η μέθοδος contrastive learning σε συνδυασμό με τις τυπικές αρχιτεκτονικές που χρησιμοποιούνται, κοινώς, στο image segmentation ιατρικών εικόνων. Η μέθοδος αυτή αποδεικνύεται αρκετά υποσχόμενη, καθώς ανταποκρίνεται αποτελεσματικότερα, σε σύγκριση με άλλες state-of-the-art self-supervised μεθόδους, οι οποίες καταπιάνονται με την προ-εκαπίδευση του δικτύου Unet. Η απαίτηση για μικρότερο αριθμό δεδομένων είναι ένα σημαντικό προτέρημα της μεθόδου, καθώς φτάνουμε σε επαρκώς ανταγωνιστικά αποτελέσματα συγκριτικά με άλλες αμιγώς supervised μεθόδους, οι οποίες ασχολούνται με το ίδιο πρόβλημα της τμηματοποίησης των ιατρικών εικόνων.

Κλείνοντας, μελλοντικά οι κατευθύνσεις που πρέπει να δοθούν αποτυπώνονται ως εξής: Αρχικά θα μπορούσε να εξερευνηθεί (1) η εφαρμογή του contrastive learning σε συνδυασμό και με άλλες αρχιτεκτονικές εκτός του Unet. Οι αρχιτεκτονικές αυτές μπορούν να είναι του τύπου encoder-decoder ή άλλες προσαρμοσμένες στην contrastive τεχνική που αναφέρονται σε Πλήρως Συνελικτικά Δίκτυα (FCN) ή και άλλης μορφής. Άλλη μια κατεύθυνση μπορεί να είναι (2) ένα εκτενέστερο benchmark διάφορων αντίστοιχων self-supervised μεθόδων, προκειμένου να εξερευνηθεί η δυνατότητα του συνδυασμού τους σε ενιαίες αποτελεσματικότερες μεθόδους. Ακόμα, (3) η πιθανότητα να εξελιχθεί μία μέθοδος transfer learning στο κομμάτι της προεκπαίδευσης του εκάστοτε δικτύου ώστε να λύθεί και το πρόβλημα των λιγοστών δεδομένων, επισημασμένων ή μη. Επιπροσθέτως (4), μία πολλά υποσχόμενη τεχνική που χρησιμοποιείται όλο και περισσότερο στην περιοχή της επεξεργασίας φυσικής γλώσσας, όρασης υπολογιστών και όχι μόνο, είναι η τεχνική της προσοχής (attention). Ο μηχανισμός αυτός δεν έχει αυστηρό μαθηματικό ορισμό και η βασική του λειτουργία είναι η ανάθεση διαφορετικών βαρών (σημασίας) σε διαφορετικά στοιχεία του δικτύου. Τα attention layers στο UNet ενσωματώνονται είτε στα skip connections, είτε στα channels, είτε σαν extra spatial attention layers. Με τα στρώματα προσοσχής, το δίκτυο αποδίδει περισσότερο βάρος στα πιο σχετιζόμενα χαρακτηριστικά, καταφέρνοντας έτσι να αποδώσει πλήρως την χωρική συσχέτιση. Τέλος, (5) η συστηματική ποσοτικοποίηση των οφελών του self-supervised learning θα δώσει την απαραίτητη ώθηση στην επιστημονική κοινότητα να ερευνήσει τον τομέα αυτόν και να τον αναπτύξει.

VIII. ARIS

Αντιμετωπίστηκαν αρκετά προβλήματα κατά την διάρκεια των πειραμάτων στον ARIS. Προέκυψαν διάφορα θέματα με την Tensorflow, τα οποία προκάλεσαν διάφορα errors. Το debugging ήταν αδύνατο στον ARIS, καθώς το PENDING status διαρκούσε αρκετό χρόνο, γεγονός που δυσκόλευε την διαδικασία. Επίσης η χρήση της GPU ήταν δυνατή μόνο μέσω του συστήματος slurm, οπότε δεν μπορούσαμε να κάνουμε debugging απο command line. Τα προβλήματα αυτά δεν συναντήθηκαν στον τοπικό υπολογιστή εξ αρχής, και έτσι καταφύγαμε συμπληρωματικά σε αυτόν προκειμένου να φέρουμε τα πειράματα και την συγκεκριμένη εργασία εις πέρας και εντός προθεσμίας.

References

[1] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," *CoRR*, vol. abs/1505.05192, 2015. [Online]. Available: http://arxiv.org/abs/1505.05192

[2] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," *CoRR*, vol. abs/1604.07379, 2016. [Online]. Available: http://arxiv.org/abs/1604.07379

[3] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," *CoRR*, vol. abs/1603.09246, 2016. [Online]. Available: http://arxiv.org/abs/1603.09246

[4] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," *CoRR*, vol. abs/1803.07728, 2018. [Online]. Available: http://arxiv.org/abs/1803.07728

[5] L. Chen, P. Bentley, K. Mori, K. Misawa, M. Fujiwara, and D. Rueckert, "Self-supervised learning for medical image analysis using image context restoration," *Medical Image Analysis*, vol. 58, p. 101539, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841518304699

[6] A. Dosovitskiy, J. T. Springenberg, M. Riedmiller, and T. Brox, "Discriminative unsupervised feature learning with convolutional neural networks," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, Eds., vol. 27. Curran Associates, Inc., 2014. [Online]. Available: https://proceedings.neurips.cc/paper/2014/file/07563a3fe3bbe7e3ba84431ad9d055af-Paper.pdf

[7] I. Misra and L. van der Maaten, "Self-supervised learning of pretext-invariant representations," *CoRR*, vol. abs/1912.01991, 2019. [Online]. Available: http://arxiv.org/abs/1912.01991

[8] K. He, H. Fan, Y. Wu, S. Xie, and R. B. Girshick, "Momentum contrast for unsupervised visual representation learning," *CoRR*, vol. abs/1911.05722, 2019. [Online]. Available: http://arxiv.org/abs/1911.05722

[9] O. J. Hénaff, A. Srinivas, J. D. Fauw, A. Razavi, C. Doersch, S. M. A. Eslami, and A. van den Oord, "Data-efficient image recognition with contrastive predictive coding," *CoRR*, vol. abs/1905.09272, 2019. [Online]. Available: http://arxiv.org/abs/1905.09272

[10] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *CoRR*, vol. abs/1807.03748, 2018. [Online]. Available: http://arxiv.org/abs/1807.03748

[11] M. Tschannen, J. Djolonga, P. K. Rubenstein, S. Gelly, and M. Lucic, "On mutual information maximization for representation learning," *CoRR*, vol. abs/1907.13625, 2019. [Online]. Available: http://arxiv.org/abs/1907.13625

[12] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio, "Learning deep representations by mutual information estimation and maximization," 2019.

[13] P. Bachman, R. D. Hjelm, and W. Buchwalter, "Learning representations by maximizing mutual information across views," 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/ddf354219aac374f1d40b7e760ee5bb7-Paper.pdf

[14] O. Chapelle, B. Scholkopf, and A. Zien, Eds., "Semi-supervised learning (chapelle, o. et al., eds.; 2006) [book reviews]," *IEEE Transactions on Neural Networks*, vol. 20, no. 3, pp. 542–542, 2009.

[15] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization." *CAP*, pp. 281–296, 2005.

[16] D. Lee, "Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks," 2013.

[17] D. P. Kingma, D. J. Rezende, S. Mohamed, and M. Welling, "Semi-supervised learning with deep generative models," *CoRR*, vol. abs/1406.5298, 2014. [Online]. Available: http://arxiv.org/abs/1406.5298

[18] A. Rasmus, M. Berglund, M. Honkala, H. Valpola, and T. Raiko, "Semi-supervised learning with ladder networks," in *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds., vol. 28. Curran Associates, Inc., 2015. [Online]. Available: https://proceedings.neurips.cc/paper/2015/file/378a063b8fdb1db941e34f4bde584c7d-Paper.pdf

[19] D. McClosky, E. Charniak, and M. Johnson, "Effective self-training for parsing," in *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, ser. HLT-NAACL '06. USA: Association for Computational Linguistics, 2006, p. 152–159. [Online]. Available: https://doi.org/10.3115/1220835.1220855

[20] D. Yarowsky, "Unsupervised word sense disambiguation rivaling supervised methods," in *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics*, ser. ACL '95. USA: Association for Computational Linguistics, 1995, p. 189–196. [Online]. Available: https://doi.org/10.3115/981658.981684

[21] D. McClosky, E. Charniak, and M. Johnson, "Effective self-training for parsing," in *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*. New York City, USA: Association for Computational Linguistics, Jun. 2006, pp. 152–159. [Online]. Available: https://www.aclweb.org/anthology/N06-1020

[22] Y. Zhang, L. Yang, J. Chen, M. Fredericksen, D. P. Hughes, and D. Z. Chen, "Deep adversarial networks for biomedical image segmentation utilizing unannotated images," in *Medical Image Computing and Computer Assisted Intervention MICCAI 2017*, M. Descoteaux, L. Maier-Hein, A. Franz, P. Jannin, D. L. Collins, and S. Duchesne, Eds. Cham: Springer International Publishing, 2017, pp. 408–416.

[23] D. C. Ciresan, U. Meier, J. Masci, L. M. Gambardella, and J. Schmidhuber, "High-performance neural networks for visual object classification," *CoRR*, vol. abs/1102.0183, 2011. [Online]. Available: http://arxiv.org/abs/1102.0183

[24] J. Hong, B.-Y. Park, and H. Park, "Convolutional neural network classifier for distinguishing barrett's esophagus and neoplasia endomicroscopy images," *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, vol. 2017, p. 2892—2895, July 2017. [Online]. Available: https://doi.org/10.1109/EMBC.2017.8037461

[25] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, Eds., vol. 27. Curran Associates, Inc., 2014. [Online]. Available: https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf

[26] P. Costa, A. Galdran, M. I. Meyer, M. D. Abràmoff, M. Niemeijer, A. M. Mendonça, and A. Campilho, "Towards adversarial retinal image synthesis," *CoRR*, vol. abs/1701.08974, 2017. [Online]. Available: http://arxiv.org/abs/1701.08974

[27] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. G. Ballester, G. Sanroma, S. Napel, S. Petersen, G. Tziritas, E. Grinias, M. Khened, V. A. Kollerathu, G. Krishnamurthi, M.-M. Rohé, X. Pennec, M. Sermesant, F. Isensee, P. Jager, K. H. Maier-Hein, P. M. Full, I. Wolf, S. Engelhardt, C. Baumgartner, L. Koch, J. Wolterink, I. Isgum, Y. Jang, Y. Hong, J. Patravali, S. Jain, O. Humbert, and P.-M. Jodoin, "Deep Learning Techniques for Automatic MRI Cardiac Multi-structures Segmentation and Diagnosis: Is the Problem Solved?" *IEEE Transactions on Medical Imaging*, vol. 37, no. 11, pp. 2514–2525, May 2018. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01803621

[28] "Medical segmentation decathlon chalenge." [Online]. Available: http://medicaldecathlon.com/index.html

[29] "https://link.springer.com/chapter/10.1007/978-3-642-33418-4$_5$1."

[30] N. J. Tustison, B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich, and J. C. Gee, "N4itk: Improved n3 bias correction," *IEEE Transactions on Medical Imaging*, vol. 29, no. 6, pp. 1310–1320, 2010.

[31] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *CoRR*, vol. abs/1505.04597, 2015. [Online]. Available: http://arxiv.org/abs/1505.04597

[32] K. Chaitanya, E. Erdil, N. Karani, and E. Konukoglu, "Contrastive learning of global and local features for medical image segmentation with limited annotations," *CoRR*, vol. abs/2006.10511, 2020. [Online]. Available: https://arxiv.org/abs/2006.10511

[33] K. Chaitanya, N. Karani, C. F. Baumgartner, E. Erdil, A. Becker, O. Donati, and E. Konukoglu, "Semi-supervised task-driven data augmentation for medical image segmentation," *Medical Image Analysis*, vol. 68, p. 101934, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S136184152030298X

[34] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=r1Ddp1-Rb

[35] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," 2020.