Xception & GRU για δημιουργια περιγραφής σε εικόνες

Παπαγρηγορίου Βασίλειος Σάββας
vasilispapg@outlook.com
Π.Μ.Σ. Τεχνητή Νοημοσύνη
Α.Π.Θ.

Ι . Εισαγωγή

Αυτό που προσπάθησα αρχικά ηταν η κατανοηση των συνελεκτικων δικτυων ωστε να βρω τρόπο να παραγω κάποια στοιχεία τα οποια θα έχουν σημασία για τα αναδρομικά δίκτυα. Ξεκίνησα με απλά συνελεκτικα δίκτυα ενός ή δύο layer και στην συνέχεια προσπαθησα να χρησιμοποιησω καποιες αρχιτεκτονικές όπως ResNet-50, Xception. Για τα αναδρομικά δίκτυα ξεκίνησα με RNN και στην συνεχεια προσπαθησα να εφαμορσω μια τεχνική που διαβασα σε ενα προσφατο paper για το patch and pack με Visual Transformers (NaViT) τα οποια παιρνουν εικονες οποιαδηποτε μεγεθους. Αλλά, μας είπατε οτι αυτα απαιτουν πολλους πορους για να εκπαιδευτουν, άλλαξα τακτική και πήγα σε μια πιο απλή λύση με την χρήση των LSTM και στην συνέχεια με την χρήση των Positional Encodings. Στο τέλος, κατέληξα με το σύστημα ενός προ-εκπαιδευμένουν Xception για την εξαγωγή των χαρακτηριστικών και με την χρήση των GRU δικτύου για την δημιουργία τω περιγραφών. Για δεδομένα χρησιμοποίησα το dataset του Flickr8k απο το kaggle.

ΙΙ . Προεπεξεργασία Δεδομένων

Η προεπεξεργασία δεδομένων αποτελεί κρίσιμο στάδιο στην ανάπτυξη και εκπαίδευση των μοντέλων μηχανικής μάθησης. Στο πλαίσιο της εργασίας αυτής, η προεπεξεργασία των δεδομένων πραγματοποιήθηκε ως εξής:

Α′. Σύνολο Δεδομένων

Χρησιμοποιήθηκε το σύνολο δεδομένων Flickr8k, το οποίο περιλαμβάνει 8.000 εικόνες με περιγραφές (λεζάντες) στα Αγγλικά. Κάθε εικόνα συνοδεύεται από πέντε διαφορετικές περιγραφές που περιγράφουν την εικόνα. Αυτό το σύνολο δεδομένων είναι ευρέως χρησιμοποιούμενο για την εκπαίδευση και αξιολόγηση μοντέλων περιγραφής εικόνας.

Β′. *Tokenization*

Για την επεξεργασία των περιγραφών, χρησιμοποιήθηκε η βιβλιοθήκη `spacy` για την ανάλυση και διαχωρισμό των προτάσεων σε λέξεις (tokens). Η διαδικασία της τοκενιζατιον είναι σημαντική για τη μετατροπή των κειμένων σε μορφή που μπορεί να επεξεργαστεί το μοντέλο.

Γ′. Δημιουργία Λεξιλογίου

Η επόμενη φάση περιλάμβανε τη δημιουργία ενός λεξιλογίου από τις περιγραφές του συνόλου δεδομένων. Το λεξιλόγιο περιλαμβάνει όλες τις μοναδικές λέξεις που εμφανίζονται στις περιγραφές, με συχνότητα εμφάνισης πάνω από ένα προκαθορισμένο κατώφλι. Λέξεις με πολύ χαμηλή συχνότητα εμφάνισης εξαιρέθηκαν για να μειωθεί η πολυπλοκότητα του μοντέλου.

Δ′. Μετατροπή σε Αριθμητικές Τιμές *(Numericalization)*

Μετά την δημιουργία του λεξιλογίου, οι λέξεις στις περιγραφές μετατράπηκαν σε αριθμητικές τιμές, χρησιμοποιώντας ένα σύστημα αντιστοίχισης λέξεων σε μοναδικούς αριθμούς. Αυτή η διαδικασία είναι γνωστή ως numericalization και επιτρέπει στο μοντέλο να επεξεργαστεί τις λέξεις ως αριθμητικά δεδομένα.

Ε′. Μετασχηματισμοί Εικόνων

Για την επεξεργασία των εικόνων, εφαρμόστηκαν διάφοροι μετασχηματισμοί ώστε να διασφαλιστεί η ομοιογένεια και η κανονικοποίηση των δεδομένων εισόδου. Αυτοί οι μετασχηματισμοί περιλαμβάνουν την αλλαγή μεγέθους των εικόνων σε συγκεκριμένες διαστάσεις και την κανονικοποίηση των τιμών των πιξελ με βάση τη μέση και τυπική απόκλιση των τιμών των πιξελ στο σύνολο δεδομένων ImageNet.

Ϛ′. Προεπεξεργασία των Περιγραφών

Οι περιγραφές προεπεξεργάστηκαν ώστε να αφαιρεθούν μη αλφαβητικοί χαρακτήρες και λέξεις με ένα μόνο χαρακτήρα. Επιπλέον, προστέθηκαν ειδικά τοκενς έναρξης και λήξης (ʿstartʾ και ʿendʾ) σε κάθε περιγραφή, καθώς και παδδινγ για να εξασφαλιστεί ότι όλες οι περιγραφές έχουν το ίδιο μήκος.

ΙΙΙ . Αρχιτεκτονική Μοντέλου

Η αρχιτεκτονική του μοντέλου αποτελείται από έναν Encoder και έναν Decoder που συνεργάζονται για να παράγουν περιγραφές εικόνας.

Α′. *Encoder* (Κωδικοποιητής)

Ο Encoder είναι υπεύθυνος για την επεξεργασία των χαρακτηριστικών των εικόνων. Η αρχιτεκτονική του Encoder περιλαμβάνει τα εξής στάδια:

- Προβολή Χαρακτηριστικών: Τα χαρακτηριστικά της εικόνας που εξάγονται από το οπτικό μοντέλο (όπως

το Xception) περνούν από ένα επίπεδο προβολής που τα μετατρέπει σε έναν νέο χώρο ενσωμάτωσης (embedding space).
- Διαμόρφωση Διάστασης: Οι διαστάσεις της εισόδου προσαρμόζονται ώστε να είναι συμβατές με τις απαιτήσεις του GRU.

Β′. *Decoder* (Αποκωδικοποιητής)

Ο Decoder είναι υπεύθυνος για την παραγωγή της τελικής περιγραφής της εικόνας. Η αρχιτεκτονική του Decoder περιλαμβάνει τα εξής στάδια:

- Ενσωμάτωση Λέξεων: Οι λέξεις της περιγραφής μετατρέπονται σε διανύσματα ενσωμάτωσης (embedding vectors) μέσω ενός επιπέδου ενσωμάτωσης.
- **GRU**: Χρησιμοποιείται μια μονάδα GRU για την επεξεργασία των ενσωματωμένων διανυσμάτων και την παραγωγή της τελικής περιγραφής.
- Τελικό Γραμμικό Επίπεδο: Ένα τελικό γραμμικό επίπεδο χρησιμοποιείται για να προβλέψει την πιθανότητα κάθε λέξης στο λεξιλόγιο.

Γ′. *Image Captioning Model*

Το τελικό μοντέλο περιγραφής εικόνας (Image Captioning Model) συνδυάζει τον Encoder και τον Decoder για να παράγει περιγραφές εικόνων. Η αρχιτεκτονική περιλαμβάνει τα εξής:

- Οπτικό Μοντέλο (**Visual Model**): Χρησιμοποιείται το Xception για την εξαγωγή χαρακτηριστικών από τις εικόνες.
- **Encoder**: Ο Encoder λαμβάνει τα χαρακτηριστικά της εικόνας και τα προβάλλει σε έναν χώρο ενσωμάτωσης.
- **Decoder**: Ο Decoder λαμβάνει τα ενσωματωμένα διανύσματα και τις εισόδους των περιγραφών για να παράγει την τελική περιγραφή της εικόνας.

IV . Εκπαίδευση

Η εκπαίδευση του μοντέλου αποτελεί κρίσιμο στάδιο για την επίτευξη υψηλής απόδοσης και ακρίβειας. Η διαδικασία εκπαίδευσης περιλαμβάνει διάφορα βήματα και τεχνικές που εφαρμόστηκαν για την βελτιστοποίηση του μοντέλου.

Α′. Βρόχος Εκπαίδευσης

Ο βρόχος εκπαίδευσης (training loop) που χρησιμοποιήθηκε για την εκπαίδευση του μοντέλου περιλαμβάνει τα εξής βήματα:

- Προετοιμασία Δεδομένων: Τα δεδομένα χωρίζονται σε σύνολα εκπαίδευσης και επικύρωσης. Χρησιμοποιήθηκε η συνάρτηση SplitDataset για να δημιουργήσει τα σύνολα δεδομένων και τους αντίστοιχους φορτωτές (data loaders).
- Ορισμός Μοντέλου και Βελτιστοποιητή: Το μοντέλο ImageCaptioningModel ορίζεται και μεταφέρεται στη συσκευή (device) εκπαίδευσης. Χρησιμοποιήθηκε κυρίως ο βελτιστοποιητής Adam με διάφορες υπερπαραμέτρους, αλλά επίσης δοκιμάστηκαν και οι Adagrad και SGD.
- Ορισμός Κριτηρίου: Ως κριτήριο απώλειας (criterion) χρησιμοποιήθηκε η Cross-Entropy Loss με μάσκα για την αντιμετώπιση των τιμών padding.
- Αξιολόγηση: Μετά από κάθε εποχή, το μοντέλο αξιολογείται στο σύνολο δεδομένων επικύρωσης για να παρακολουθηθεί η απόδοσή του και να εντοπιστούν πιθανά προβλήματα υπερεκπαίδευσης.
- Αποθήκευση Προτύπου: Ανά τακτά διαστήματα, το πρότυπο αποθηκεύεται για να διασφαλιστεί ότι οι παράμετροι του μοντέλου διατηρούνται.

Β′. Κώδικας Εκπαίδευσης

Ο κώδικας εκπαίδευσης περιλαμβάνει τις εξής λειτουργίες:

- **SplitDataset**: Χωρίζει το σύνολο δεδομένων σε υποσύνολα εκπαίδευσης και επικύρωσης και δημιουργεί φορτωτές δεδομένων.
- **get_arguments**: Λαμβάνει τα επιχειρήματα γραμμής εντολών για να καθορίσει τις λειτουργίες που θα εκτελεστούν (εκπαίδευση, πρόβλεψη, αξιολόγηση, κ.λπ.).
- **train_model**: Εκπαιδεύει το μοντέλο για ένα καθορισμένο αριθμό εποχών.
- **main**: Η κύρια συνάρτηση που εκτελεί την εκπαίδευση, την πρόβλεψη ή την αξιολόγηση του μοντέλου, ανάλογα με τα επιχειρήματα γραμμής εντολών.

V . Προβλήματα και Λύσεις

Κατά τη διάρκεια της εκπαίδευσης του μοντέλου, αντιμετώπισα διάφορα προβλήματα. Το κυριότερο πρόβλημα που αντιμετώπισα ήταν η υπερεκπαίδευση (overfitting).

Α′. Πειραματισμοί

- Ρύθμιση Υπερπαραμέτρων: Πειραματίστηκα με διάφορες τιμές για τις υπερπαραμέτρους του αλγορίθμου Adam, καθώς και με άλλους αλγόριθμους βελτιστοποίησης όπως Adagrad και SGD, με υπερπαραμέτρους που κυμαίνονταν από 0.0000001 έως 0.02, σε τιμές όπως 0.0000002, 0.00001, κ.λπ.
- Αποφυγή Υπερπροσαρμογής: Εφάρμοσα τεχνικές όπως η πρόωρη διακοπή (early stopping) και η τακτική αποδοχής (dropout) για να μειώσω την υπερεκπαίδευση.

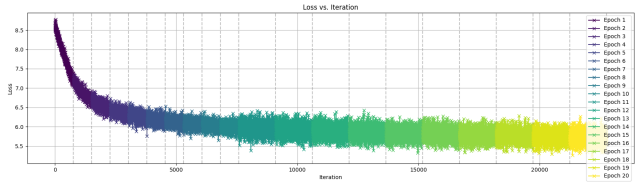
Παρά τις προσπάθειες και τους πειραματισμούς με διάφορες υπερπαραμέτρους και αλγόριθμους βελτιστοποίησης, το πρόβλημα που είχα παρέμενε. Το μοντέλο εξακολουθεί να εμφανίζει χαρακτήρες όπως 'UNK' ή 'PAD' σε ορισμένες περιπτώσεις. Αυτό δείχνει ότι υπάρχει ακόμη περιθώριο για βελτιστοποίηση και περαιτέρω πειραματισμό με τις υπερπαραμέτρους και τις τεχνικές εκπαίδευσης.

VI . Ανάλυση Απώλειας (Loss Analysis)

Κατά τη διάρκεια της εκπαίδευσης του μοντέλου, παρακολούθησα την απώλεια (loss) για να αξιολογήσω την απόδοση και την πορεία της εκπαίδευσης. Παρακάτω παρατίθενται διάφορες γραφικές παραστάσεις της απώλειας για διαφορετικές διαμορφώσεις του μοντέλου.

Α΄. Απώλεια με Κωδικοποίηση Θέσης και 8 Κεφαλές, 6 Επίπεδα Αποκωδικοποίησης

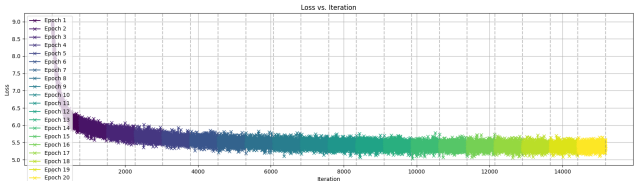
Η εικόνα παρουσιάζει την απώλεια κατά την εκπαίδευση του μοντέλου με κωδικοποίηση θέσης (positional encoding), 8 κεφαλές (heads) και 6 επίπεδα αποκωδικοποίησης (decoder layers). Η απώλεια μειώνεται σταθερά κατά την εκπαίδευση, δείχνοντας ότι το μοντέλο μαθαίνει αποτελεσματικά αλλά όχι αρκετά καλά.



Σχήμα 1. Απώλεια με Κωδικοποίηση Θέσης, 8 Κεφαλές, 6 Επίπεδα Αποκωδικοποίησης

Β΄. Απώλεια με Κωδικοποίηση Θέσης και 2 Κεφαλές, 3 Επίπεδα Αποκωδικοποίησης

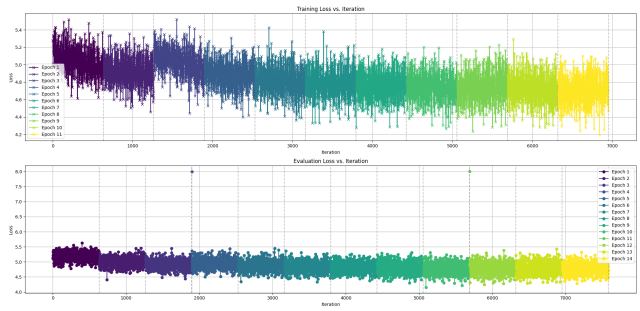
Η εικόνα δείχνει την απώλεια κατά την εκπαίδευση του μοντέλου με κωδικοποίηση θέσης, 2 κεφαλές και 3 επίπεδα αποκωδικοποίησης. Η απώλεια επίσης μειώνεται σταθερά, αλλά με λιγότερες κεφαλές και επίπεδα αποκωδικοποίησης.



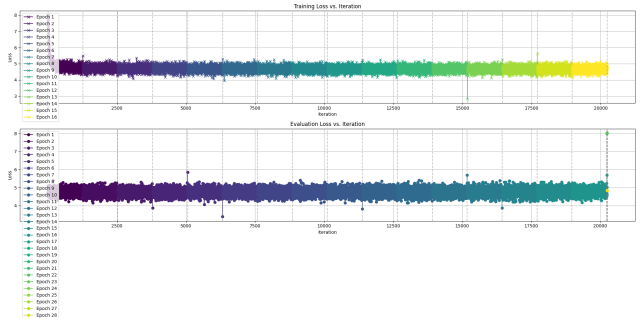
Σχήμα 2. Απώλεια με Κωδικοποίηση Θέσης, 2 Κεφαλές, 3 Επίπεδα Αποκωδικοποίησης

Γ΄. Απώλεια με *GRU*

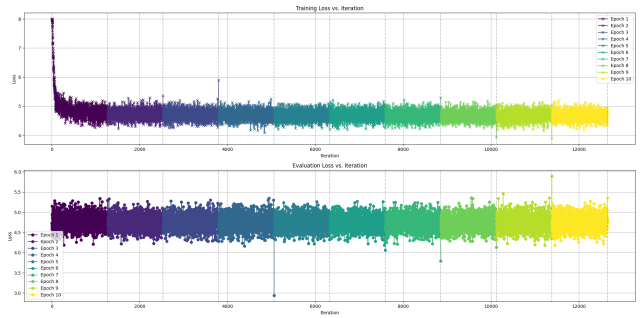
Οι παρακάτω εικόνες παρουσιάζουν την απώλεια κατά την εκπαίδευση αλλά και του validation του μοντέλου χρησιμοποιώντας GRU. Παρατηρείται οτι οσο και αυξανω το βάθος ή ανεβάζω το μέγεθος του εμβεδδινγ, δεν υπάρχει κάποια βελτίωση στην απώλεια.



Σχήμα 3. Απώλεια με GRU 8 layers 512 hidden



Σχήμα 4. Απώλεια με GRU 16 layers 4098 embedding

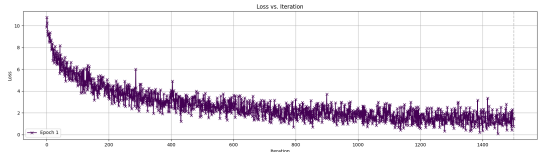


Σχήμα 5. Απώλεια με GRU 64 layers 1024 embedding

Δ΄. Υπόλοιπες αρχιτεκτονικές

Αποφάσησα να μην εμφανισω άλλες εικόνες με λοσς καθώς είχαν παρθεί από όταν δεν είχα φτιάξει το πρόβλημα που εμφανίστηκε με τον δατα λοαδερ. Παράδειγμα αυτών η παρακάτω εικόνα.

Παρακάτω η εικόνα παρουσιάζει την απώλεια κατά την εκπαίδευση του μοντέλου με κωδικοποίηση θέσης, 16 κεφαλές και 32 επίπεδα αποκωδικοποίησης. Παρατηρείται υπερεκπαίδευση (overfitting), καθώς δεν είχα ακόμει φτίαξει κατάλληλα τον data loader για να αποφύγω την υπερεκπαίδευση και υπηρχαν εικονες που εμφανιζοντουσαν και στα 2 data loaders test / train.



Σχήμα 6. Απώλεια με Κωδικοποίηση Θέσης, 16 Κεφαλές, 32 Επίπεδα Αποκωδικοποίησης (1), αποτέλεσμα οερφιττινγ

Ε΄. Συμπεράσματα

Σαν αποτέλεσμα δεν είχα κάποια ιδιαίτερη επιτυχία στην μείωση της απώλειας και την βελτίωση του μοντέλου. Το μοντέλο βγάζει αποτελέσματα αλλά δεν έχουν κάποια σημασία ή συνοχή. Παρόλα αυτά η εμπειρία αυτή με βοήθησε να καταλάβω καλύτερα την λειτουργία των συνελεκτικών και αναδρομικών δικτύων και την διαδικασία εκπαίδευσης τους.