Συστήματα Κατανόησης και Παραγωγής Κειμένου (M915)

Assignment 2

Δήμητρα Κοντοέ lt1200028

Σκοπός της εργασίας ήταν το fine-tuning ενός Bert μοντέλου (https://huggingface.co/distilbert-base-uncased) για το task του extractive question-answering, δηλαδή, ουσιαστικά, η αξιοποίηση όλων των layers των encoders του distilbert-base-uncased και η επανεκπαίδευση του τελευταίου head του για την παραγωγή αποτελεσμάτων για το συγκεκριμένο task. Το task περιλαμβάνει τη διατύπωση ερωτήσεων ως προς κάποιο κείμενο που δίνεται και τον εντοπισμό, από το σύστημα, των αντίστοιχων απαντήσεων ως τμήμα (με ορισμένη αρχή και τέλος) του ίδιου κειμένου αναφοράς. Όσον αφορά τα δεδομένα, χρησιμοποιήθηκε μέσω της βιβλιοθήκης *datasets* το SQUAD 1.1 (ερωτήσεις και απαντήσεις με βάση ένα σύνολο άρθρων την Wikipedia) ως εξής:

- 10570 δείγματα για training (το validation dataset του SQUAD 1.1)
- 1000 δείγματα για validation (τα 1000 πρώτα δείγματα του train dataset του SQUAD 1.1)
- 200 δείγματα για testing (200 τυχαία επιλεγμένα δείγματα του train dataset του SQUAD 1.1)

Το κάθε δείγμα ήταν ένα dictionary αντικείμενο με τα εξής key (και τα αντίστοιχα values τους):

{'answers': {'answer_start': [ ], 'text': [ ]},
'context': ' ',
'id': ' ',
'question': ' ',
'title': ' '}

Με βάση το processing που εφαρμόσαμε στα δεδομένα, και προκειμένου να αποφύγουμε την απώλεια πληροφορίας με truncating (και πιθανότατα απώλεια κάποια απάντησης εντός του συγκεκριμένου εύρους), για κάθε input (question, context), εάν αυτό υπέρβαινε το max_length, που ορίσαμε σε 384 subtokens, το context χωριζόταν σε περισσότερα κομμάτια, με ένα επιτρεπτό όριο overlapping (doc_stride = 128), το πλήθος των τελικό πλήθος των δειγμάτων διαμορφώθηκε ως εξής: training : 10784 και validation: 1032.

Για το finetuning του μοντέλου χρησιμοποιήθηκε το αντικείμενο Trainer, της βιβλιοθήκης των transformers, με ελάχιστες διαφορές σε σχέση με τις default υπερπαραμέτρους:

```
evaluation_strategy = "epoch",
learning_rate=2e-5,
per_device_train_batch_size=batch_size,
per_device_eval_batch_size=batch_size,
num_train_epochs=3,
weight_decay=0.01
```

Με βάση το training και το evaluation loss, εκτιμάμε ότι από την 3^{η} εποχή ξεκινάει να εμφανίζεται overfitting στα training δεδομένα. Πράγματι, κάτι τέτοιο επιβεβαίωσε και ένα πείραμα που υλοποιήσαμε με 6 εποχές, καθώς το loss του training συνέχιζε να μειώνεται ενώ το loss του validation ακολουθούσε ανοδική πορεία. Θεωρούμε κάτι τέτοιο αναμενόμενο με βάση το περιορισμένο πλήθος των δεδομένων που χρησιμοποιήθηκαν για το fine-tuning (για τη την ολοκλήρωση της εκπαίδευσης εντός των επιτρεπτών χρονικών ορίων χρήσης GPU στο Google Collab).

Epoch	Training Loss	Validation Loss
1	3.024300	1.590448
2	1.647100	1.473043
3	1.157200	1.503966

Σε κάθε περίπτωση, αξιολογούμε ως ιδιαίτερα εντυπωσιακό, δεδομένου του μικρού πλήθος δεδομένων, το γεγονός ότι με ελάχιστες εποχές εκπαίδευσης και ελάχιστη παραμετροποίηση, το fine-tuned μοντέλο σημείωσε σχετικά αξιόλογο σκορ στα test δεδομένα (βλ. παρακάτω), τέτοιο που θα μπορούσε να το χαρακτηρίσει baseline. Κάτι που μας οδηγεί σε χρήσιμα συμπεράσματα σχετικά με τις δυνατότητες και τις επιδόσεις των μοντέλων transformers και των μεθόδων transfer learning που προσφέρονται με αυτά.

```
{'exact_match': 53.0, 'f1': 69.03276066816}
```

Σχετικά με τις μετρικές που χρησιμοποιήθηκαν:

exact match: δυαδικό μέτρο του ποσοστού των παραγόμενων απαντήσεων που αντιστοιχούν στις ground truth απαντήσεις (αναλογία των ερωτήσεων που απαντήθηκαν από το σύστημα με ακριβώς τις λέξεις που περιλαμβάνουν οι ground truth απαντήσεις).

F1: αρμονικός μέσος precision και recall. Το precision υπολογίζεται ως το πλήθος των σωστών λέξεων που προβλέφθηκαν προς το σύνολο όλων των λέξεων που προβλέφθηκαν σε μια απάντηση και το recall ως το πλήθος των σωστών λέξεων προς το πλήθος όλων των λέξεων της ground truth απάντησης.

Όσον αφορά το inference που επιχειρήσαμε με βάση ένα τυχαίο απόσπασμα κειμένου της Wikipedia, αναφέρουμε τις εξής παρατηρήσεις:

- Το μοντέλο καταφέρνει να εντοπίσει τη σωστή απάντηση στις 2 πρώτες ερωτήσεις όπου δοκιμάσαμε να αναδιατυπώσουμε με χρήση συνωνύμων την ίδια ερώτηση. Σημειώνουμε ότι το ρήμα που χρησιμοποιήθηκε στις ερωτήσεις δεν εμφανίζεται στο απόσπασμα.
- Οι επόμενες 3 ερωτήσεις απαιτούν απαντήσεις με αναφορά σε ονοματικές οντότητες. Στην πρώτη περίπτωση το μοντέλο εντοπίζει σωστά μόνο την αρχή της απάντησης, επιτυγχάνοντας να επιστρέψει μόνο την πρώτη ονοματική οντότητα στη σειρά. Εκτιμάμε ότι η παράθεση των παρενθέσεων, με πληροφορίες σχετικά με το ρόλο του κάθε μέλους του συγκροτήματος, καθώς και τα κόμματα που παρεμβάλλονται ίσως να είναι οι παράμετροι εκείνες που δυσχεραίνει τη σωστή πρόβλεψη. Η απάντηση στην επόμενη ερώτηση αναζήτησης ονοματικής οντότητας είναι μεν λανθασμένη, παρ' ολ' αυτά, θεωρούμε σημαντικό ότι το μοντέλο κατάφερε να επιστρέψει πράγματι μια ονοματική οντότητα, ακόμη κι αν δεν ήταν η σωστή. Η τρίτη ερώτηση της σειράς αυτής, αποτελεί ερώτηση για την οποία δεν υπάρχει απάντηση στο δοσμένο απόσπασμα. Πιστεύαμε περισσότερο αναμενόμενο το μοντέλο να κάνει λάθος επιστρέφοντας τους τίτλους των άλμπουμ που παρατίθενται. Ωστόσο, το μοντέλο φαίνεται να δίνει μια πιο «έξυπνη» απάντηση, στα όρια της οποίας εμφανίζεται μια φράση (*progressive rock*) που περιγράφει το είδος της μουσικής των Pink Floyd (της ονοματικής οντότητας που εμφανίζεται στην ερώτηση), καθώς και η φράση *of all times* που συντακτικά είναι μια σωστή δομή που θα μπορούσε να περιλαμβάνεται σε μια φράση-απάντηση σε ερώτηση με τη λέξη *greatest*.
- Η τελευταία ερώτηση που θέσαμε στο μοντέλο απαιτούσε μια απάντηση του τύπου "Yes or No", ωστόσο η απάντηση που λαμβάνουμε είναι και πάλι μια ονοματική οντότητα, και μάλιστα διαφορετική από αυτήν που εμφανίζεται στην ερώτηση. Ωστόσο η ονοματική οντότητα αυτή αφορά πράγματι μέλος του συγκροτήματος (το οποίο είναι μάλλον το αγαπημένο του μοντέλου μας, αν κρίνουμε από τις απαντήσεις του!!!).