Επεξεργασία Φωνής και Φυσικής Γλώσσας

**<u>Προπαρασκευή 3ου Εργαστηρίου:</u> Επεξεργασία και κατηγοριοποίηση με χρήση βαθιών νευρωνικών δικτύων με χρήση της βιβλιοθήκης PyTorch**


ΣΧΟΛΗ: ΗΜΜΥ


| Ονοματεπώνυμο | Αριθμός Μητρώου |
|---|---|
| Γιάννης Πιτόσκας | 03115077 |
| Αντώνης Παπαοικονόμου | 03115140 |


## 1. Προεπεξεργασία Δεδομένων

Σε αυτό το σημείο καλούμαστε να επεξεργαστούμε τα δεδομένα, ώστε να μπορούμε να εκπαιδεύσουμε στη συνέχεια το νευρωνικό μας δίκτυο.

### 1.1 <u>Κωδικοποίηση Επισημειώσεων (Labels)</u>

Παρακάτω φαίνονται τα πρώτα 10 labels από τα δεδομένα εκπαίδευσης και οι αντιστοιχίες τους σε αριθμούς:

### 1.2 <u>Λεκτική Ανάλυση (Tokenization)</u>

Για το tokenization έχουμε χρησιμοποιήσει την συνάρτηση `word_tokenize()` του πακέτου `nltk`. Στη συνέχεια φαίνονται τα πρώτα 10 παραδείγματα από τα δεδομένα εκπαίδευσης:

### 1.3 <u>Κωδικοποίηση Παραδειγμάτων (Λέξεων)</u>

Κατόπιν υλοποίησης της μεθόδου `__getitem__ της κλάσης SentenceDataset`, τυπώνουμε 5 random παραδείγματα στην αρχική tokenized μορφή τους και ύστερα όπως τα επιστρέφει η κλάση `SentenceDataset`, δηλαδή με τα στοιχεία `example, label, length`:

```
loading word embeddings...
Loaded word embeddings from cache.

EX1: First 10 train labels with encodings:

positive -> 1
positive -> 1
positive -> 1
positive -> 1
positive -> 1
positive -> 1
positive -> 1
positive -> 1
positive -> 1
positive -> 1

EX2: First 10 tokenized train data:

['the', 'rock', 'is', 'destined', 'to', 'be', 'the', '21st', 'century', "'s", 'new', '``', 'conan', '``',
'and', 'that', 'he', "'s", 'going', 'to', 'make', 'a', 'splash', 'even', 'greater', 'than', 'arnold',
'schwarzenegger', ',', 'jean-claud', 'van', 'damme', 'or', 'steven', 'segal', '.']
```

```
['the', 'gorgeously', 'elaborate', 'continuation', 'of', '``', 'the', 'lord', 'of', 'the', 'rings', '``',
'trilogy', 'is', 'so', 'huge', 'that', 'a', 'column', 'of', 'words', 'can', 'not', 'adequately', 'describe',
'co-writer/director', 'peter', 'jackson', "'s", 'expanded', 'vision', 'of', 'j', '.', 'r', '.', 'r', '.',
'tolkien', "'s", 'middle-earth', '.']
['effective', 'but', 'too-tepid', 'biopic']
['if', 'you', 'sometimes', 'like', 'to', 'go', 'to', 'the', 'movies', 'to', 'have', 'fun', ',', 'wasabi',
'is', 'a', 'good', 'place', 'to', 'start', '.']
['emerges', 'as', 'something', 'rare', ',', 'an', 'issue', 'movie', 'that', "'s", 'so', 'honest', 'and',
'keenly', 'observed', 'that', 'it', 'does', "n't", 'feel', 'like', 'one', '.']
['the', 'film', 'provides', 'some', 'great', 'insight', 'into', 'the', 'neurotic', 'mindset', 'of', 'all',
'comics', '--', 'even', 'those', 'who', 'have', 'reached', 'the', 'absolute', 'top', 'of', 'the', 'game',
'.']
['offers', 'that', 'rare', 'combination', 'of', 'entertainment', 'and', 'education', '.']
['perhaps', 'no', 'picture', 'ever', 'made', 'has', 'more', 'literally', 'showed', 'that', 'the', 'road',
'to', 'hell', 'is', 'paved', 'with', 'good', 'intentions', '.']
['steers', 'turns', 'in', 'a', 'snappy', 'screenplay', 'that', 'curls', 'at', 'the', 'edges', ';', 'it',
"'s", 'so', 'clever', 'you', 'want', 'to', 'hate', 'it', '.', 'but', 'he', 'somehow', 'pulls', 'it', 'off',
'.']
['take', 'care', 'of', 'my', 'cat', 'offers', 'a', 'refreshingly', 'different', 'slice', 'of', 'asian',
'cinema', '.']

EX3: 5 random SentenceDatasets from train set:

['watching', 'e', '.', 't', 'now', ',', 'in', 'an', 'era', 'dominated', 'by', 'cold', ',', 'loud', 'special-
effects-laden', 'extravaganzas', ',', 'one', 'is', 'struck', 'less', 'by', 'its', 'lavish', 'grandeur',
'than', 'by', 'its', 'intimacy', 'and', 'precision', '.']
[  2642   1111      3   2160    115      2      7     30   1593   2690
     22   1867      2   6663 400001  76572      2     49     15   1870
    441     22     48  12351  27141     74     22     48  22101      6
   9814      3      0      0      0      0      0      0      0      0
      0      0      0      0      0      0      0      0      0      0
      0      0      0      0      0      0      0      0      0      0
      0      0]
1
32
['although', 'the', 'subject', 'matter', 'may', 'still', 'be', 'too', 'close', 'to', 'recent', 'national',
'events', ',', 'the', 'film', 'works', '-', 'mostly', 'due', 'to', 'its', 'superior', 'cast', 'of',
'characters', '.']
[ 377    1 1699 1121  108  150   31  318  384    5  398  122  959    2    1
  320  851   12 1247  446    5   48 4741 1785    4 2154    3    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0]
1
27
['those', 'who', 'love', 'cinema', 'paradiso', 'will', 'find', 'the', 'new', 'scenes', 'interesting', ',',
'but', 'few', 'will', 'find', 'the', 'movie', 'improved', '.']
[  156    39   836  5993 63300    44   597     1    51  3469  4002     2
    35   307    44   597     1  1006  2339     3     0     0     0     0
     0     0     0     0     0     0     0     0     0     0     0     0
     0     0     0     0     0     0     0     0     0     0     0     0
     0     0     0     0     0     0     0     0     0     0     0     0
     0     0]
1
20
['the', 'director', ',', 'mark', 'pellington', ',', 'does', 'a', 'terrific', 'job', 'conjuring', 'up', 'a',
'sinister', ',', 'menacing', 'atmosphere', 'though', 'unfortunately', 'all', 'the', 'story', 'gives', 'us',
'is', 'flashing', 'red', 'lights', ',', 'a', 'rattling', 'noise', ',', 'and', 'a', 'bump', 'on', 'the',
'head', '.']
[     1    370      2    800 166469      2    261      8  11026    665
  53737     61      8  16092      2  22067   3905    414   4717     65
      1    524   1830     96     15  16711    640   4513      2      8
  23156   6057      2      6      8  15124     14      1    363      3
```

```
      0      0      0      0      0      0      0      0      0      0
      0      0      0      0      0      0      0      0      0      0
      0      0]
1
40
['to', 'say', 'that', 'this', 'vapid', 'vehicle', 'is', 'downright', 'doltish', 'and', 'uneventful', 'is',
'just', 'as', 'obvious', 'as', 'telling', 'a', 'country', 'skunk', 'that', 'he', 'has', 'severe', 'body',
'odor', '.']
[     5    204     13     38  72163   1908     15  21970 227578      6
  35996     15    121     20   4399     20   2822      8    124  40316
     13     19     32   2547    720  20547      3      0      0      0
      0      0      0      0      0      0      0      0      0      0
      0      0      0      0      0      0      0      0      0      0
      0      0      0      0      0      0      0      0      0      0
      0      0]
0
27
```

## 2. Μοντέλο

Σε αυτό το βήμα γίνεται ο σχεδιασμός του νευρωνικού μας δικτύου.

### 2.1 Embedding Layer

- Γιατί αρχικοποιούμε το embedding layer με τα προεκπαιδευμένα word embeddings;

  Διότι άμα δεν αρχικοποιήσουμε το embedding layer με τα προεκπαιδευμένα word embeddings τότε τα weights παίρνουν random τιμές το οποίο φαίνεται να έχει χαμηλότερες επιδόσεις. Χρησιμοποιώντας την προσέγγιση των προεκπαιδευμένων word embeddings μπορούμε να τα χρησιμοποιήσουμε και να τα συντονίσουμε ιδανικά για τα δεδομένα.

- Γιατί κρατάμε παγωμένα τα βάρη του embedding layer κατά την εκπαίδευση;

  Δίοτι βλέπουμε ότι το embedding layer επωφελείται από αυτά τα προεκπαιδευμένα word embeddings. Τα προεκπαιδευμένα αυτά μέρη δε θα έπρεπε να ενημερώνονται κατά τη διάρκεια της εκπαίδευσης, για να αποφευχθεί το να «ξεχνιούνται» αυτά που ήδη «γνωρίζουν». Τα μεγάλα gradient updates που ενεργοποιούνται από layers που είναι τυχαία αρχικοποιημένα μπορούν να διαταράξουν τα ήδη γνωστά χαρακτηριστικά.

### 2.2 Output Layer(s)

Για την κατηγοριοποίηση θα πρέπει τώρα να προβάλλουμε τις αναπαραστάσεις των κειμένων στον χώρο των κλάσεων.

- Γιατί βάζουμε μια μη γραμμική συνάρτηση ενεργοποίησης στο προτελευταίο layer; Τι διαφορά θα είχε αν είχαμε 2 ή περισσότερους γραμμικούς μετασχηματισμούς στη σειρά;

  Ο σκοπός της συνάρτησης ενεργοποίησης είναι η εισαγωγή μη γραμμικότητας στο δίκτυο. Μη γραμμικότητα σημαίνει κιόλας ότι η έξοδος δεν μπορεί να αναπαραχθεί από έναν γραμμικό συνδυασμό των εισόδων. Πρέπει να εφαρμόσουμε μια συνάρτηση ενεργοποίησης τέτοια ώστε το δίκτυο να γίνει πιο ισχυρό και να προσθέσει την ικανότητα να μάθει κάτι πολύπλοκο από τα δεδομένα και να αντιπροσωπεύει μη γραμμικές σύνθετες αυθαίρετες αντιστοιχίες μεταξύ εισόδου και εξόδου. Ως εκ τούτου, χρησιμοποιώντας μη γραμμική ενεργοποίηση, είμαστε σε θέση να παράγουμε μη γραμμικές απεικονίσεις από την είσοδο στην έξοδο.

Στην περίπτωση, τώρα, γραμμικής συνάρτησης θα είχαμε σταθερό gradient ανεξάρτητο της εισόδου και επομένως το descent θα γίνει σε σταθερό gradient. Αν υπάρχει λοιπόν κάποιο σφάλμα στην πρόβλεψη, οι αλλαγές που έγιναν με το back propagation είναι σταθερές. Τώρα όσον αφορά τα connected layers. Κάθε layer ενεργοποιείται από γραμμική συνάρτηση. Αυτή η ενεργοποίηση με τη σειρά της πηγαίνει στο επόμενο layer ως είσοδος και το δεύτερο layer υπολογίζει το σταθμισμένο άθροισμα σε αυτή την είσοδο και με τη σειρά της, πυροδοτεί με βάση μια άλλη γραμμική συνάρτηση ενεργοποίησης. Ανεξάρτητα από το layers έχουμε, αν όλα είναι γραμμικά, η τελική συνάρτηση ενεργοποίησης του τελευταίου στρώματος δεν είναι τίποτε άλλο παρά μια γραμμική συνάρτηση της εισόδου του πρώτου layer. Αυτό σημαίνει ότι αυτά τα δύο layers (ή N layers) μπορούν να αντικατασταθούν από ένα μόνο layer. Όλο το δίκτυο λοιπόν εξακολουθεί να είναι ισοδύναμο με ένα μόνο layer με γραμμική συνάρτηση ενεργοποίησης (ένας συνδυασμός γραμμικών συναρτήσεων σε γραμμικό τρόπο είναι ακόμα μία γραμμική συνάρτηση).

### 2.3 Forward pass

Σε αυτό το σημείο θα σχεδιάστηκε ο τρόπος με τον οποίο θα μετασχηματίσει το δίκτυο τα δεδομένα εισόδου στις αντίστοιχες εξόδους.

- Αν θεωρήσουμε ότι κάθε διάσταση του embedding χώρου αντιστοιχεί σε μια αφηρημένη έννοια, μπορείτε να δώσετε μια διαισθητική ερμηνεία για το τι περιγράφει η αναπαράσταση που φτιάξατε (κέντρο-βάρους);

  Διαισθητικά είναι σαν να κρατάμε με το κέντρο βάρους την «επικρατέστερη» τιμή για να μπορέσει να γίνει αντιληπτό σε ποια κατηγορία ανήκει ένα τμήμα δεδομένων (πχ θεματολογία).

- Αναφέρατε πιθανές αδυναμίες της συγκεκριμένης προσέγγισης για να αναπαραστήσουμε κείμενα.

  Στην περίπτωση των κειμένων μια αδυναμία αποτελεί το γεγονός ότι υπάρχουν ίδιες λέξεις με διαφορετική σημασιολογία. Ακόμη, σε μια πρόταση μπορεί να υπάρχουν λέξεις που κάθε μία απο αυτές εμφανίζεται συνηθέστερα σε διαφορετικές θεματολογίες. Γενικότερα, μπορεί η αναπαράσταση με το κέντρο βάρους να εμφανίσει αδυναμίες ορισμένες φορές, καθώς δεν είναι πάντα ενδεικτική μια τέτοιου είδους επιλογή σε κάποια ειδικά context.

## 3 Διαδικασία Εκπαίδευσης

### 3.1 Φόρτωση Παραδειγμάτων (DataLoaders)

- Τι συνέπειες έχουν τα μικρά και μεγάλα mini-batches στην εκπαίδευση των μοντέλων;

  Τα δύο βασικά πράγματα που πρέπει να ληφθούν υπόψη κατά τη βελτιστοποίηση του μεγέθους mini-batches είναι η χρονική πολυπλοκότητα της εκπαίδευσης και η θορυβώδης εκτίμηση του gradient. Συγκεκριμένα, ο υπολογισμός του gradient είναι κατά προσέγγιση γραμμικός στο μέγεθος του batch. Επομένως, θα χρειαστεί περίπου 100 φορές περισσότερο για να υπολογίσετε το gradient ενός batch μεγέθους 10.000 από ένα μεγέθους 100. Επιπλέον, το gradient ενός μοναδικού σημείου δεδομένων θα είναι πολύ πιο θορυβώδες από το gradient ενός batch μεγέθους 100. Ακόμη, τα μεγάλα batch ρίχνουν την ποιότητα του μοντέλου, με κριτήριο την ικανότητα γενίκευσης, καθώς αυτή μειώνεται σε πολύ μεγάλο βαθμό. Στην πράξη χρησιμοποιούνται μικρού προς μεσαίου μεγέθους mini-batches περίπου 10-500. (εμείς επιλέξαμε 100) (βλ. ON LARGE-BATCH TRAINING FOR DEEP LEARNING: GENERALIZATION GAP AND SHARP MINIMA, Published as a conference paper at ICLR 2017)

- Συνήθως ανακατεύουμε την σειρά των mini-batches στα δεδομένα εκπαίδευσης σε κάθε εποχή. Μπορείτε να εξηγήσετε γιατί;

Εάν η σειρά των δεδομένων σε κάθε εποχή είναι ίδια, τότε το μοντέλο μπορεί να το χρησιμοποιήσει ως τρόπο μείωσης του σφάλματος του training, κάτι που είναι ένα είδος overfitting.

### 3.2 Βελτιστοποίηση

- Κριτήριο: Χρησιμοποιήθηκε το `BCEWithLogitsLoss()` αφού έχουμε 2 κλάσεις

- Παράμετροι: Βελτιστοποίηση των παραμέτρων για τις οποίες `p.requires_grad==True`

- Optimizers: Έγινε χρήση του `RMSProp()`

### 3.3 Εκπαίδευση

Για την αξιολόγηση κάθε batch ανάλογα με τα epochs που έχουμε επιλέξει (50 στην δική μας περίπτωση) γίνεται χρήση των συναρτήσεων `train_dataset()` και `eval_dataset()`.