|  |  |  |
| --- | --- | --- |
|  | **ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΜΑΚΕΔΟΝΙΑΣ**  **ΠΟΛΥΤΕΝΧΙΚΗ ΣΧΟΛΗ ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ & ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ** |  |

Αυτόματη δημιουργία λεζάντας σε εικόνες

ΕΙΔΙΚΗ ΕΡΓΑΣΙΑ



του

**ΚΩΝΣΤΑΝΤΙΝΟΥ ΠΑΛΕΓΚΑ**

**ΑΜ:1530**



**Επιβλέπων:** Νικόλαος Πλόσκας

Αναπληρωτής καθηγητής

ΚΟΖΑΝΗ/8/2023

ΑΥΤΗ Η ΣΕΛΙΔΑ ΕΙΝΑΙ ΣΚΟΠΙΜΑ ΛΕΥΚΗ

Περιεχόμενα

[Περιεχόμενα 3](#_Toc144307543)

[Κατάλογος Σχημάτων 5](#_Toc144307544)

[Κατάλογος Εικόνων 6](#_Toc144307545)

[Κατάλογος Πινάκων 7](#_Toc144307546)

[Κεφάλαιο 1: Εισαγωγή́ 8](#_Toc144307547)

[1.1 Εισαγωγή στο πρόβλημα 8](#_Toc144307548)

[1.2 Οργάνωση του τόμου 9](#_Toc144307550)

[Κεφάλαιο 2: Μετρικές και σετ δεδομένων 10](#_Toc144307551)

[2.1 Μετρικές 10](#_Toc144307552)

[2.1.1 Bleu 10](#_Toc144307553)

[2.1.2 Meteor 11](#_Toc144307554)

[2.2 Flickr8k 12](#_Toc144307555)

[Κεφάλαιο 3: Σχεδίαση του Θέματος 13](#_Toc144307557)

[3.1 Σχεδίαση και λειτουργία του λογισμικού 13](#_Toc144307558)

[3.2 Σετ Δεδομένων 13](#_Toc144307560)

[3.2.1 Προεπεξεργασία δεδομένων 13](#_Toc144307561)

[3.3 Συνελικτικό Νευρωνικό Δίκτυο 14](#_Toc144307563)

[3.4 Δημιουργία λεξικών 14](#_Toc144307564)

[3.5 Μοντέλο RNN 15](#_Toc144307568)

[3.5.1 Εκπαίδευση 16](#_Toc144307570)

[3.6 Beam Search 17](#_Toc144307572)

[Κεφάλαιο 4: Αποτελέσματα και συμπεράσματα 18](#_Toc144307573)

[4.1 Αποτελέσματα 18](#_Toc144307574)

[4.1.1 Παραδείγματα τίτλων 18](#_Toc144307576)

[4.2 Συμπεράσματα 21](#_Toc144307582)

[4.3 Μελλοντικές βελτιώσεις 21](#_Toc144307583)

[Βιβλιογραφία 22](#_Toc144307584)

Κατάλογος Σχημάτων

[Σχήμα 1:Διαδικάσια παραγωγής τίτλου 12](#_Toc144147737)

[Σχήμα 2:Επίπεδα Νευρωνικού δικτύου 15](#_Toc144147738)

Κατάλογος Εικόνων

[Εικόνα 2: Δομή σετ-δεδομένων flickr8k 11](#_Toc144321993)

[Εικόνα 3: Ο τίτλος της εικόνας πριν και μετά την επεξεργασία. 13](#_Toc144321994)

[Εικόνα 4: Αποτέλεσμα λεξικού train\_captions. 14](#_Toc144321995)

[Εικόνα 5: Αποτελέσματα λεξικού test\_features 14](#_Toc144321996)

[Εικόνα 6: Αποτέλεσμα λεξικού train\_encoded\_captions. 15](#_Toc144321997)

[Εικόνα 7: Παράδειγμα 1 19](#_Toc144321998)

[Εικόνα 8: Παράδειγμα 2 19](#_Toc144321999)

[Εικόνα 9: Παράδειγμα 3 20](#_Toc144322000)

[Εικόνα 10: Παράδειγμα 4 20](#_Toc144322001)

[Εικόνα 11: Παράδειγμα 5 21](#_Toc144322002)

Κατάλογος Πινάκων

Πίνακας 1:Αποτελέσματα εκπαίδευσης 16

Πίνακας 2:Σύγκριση του μοντέλου με άλλες υλοποιήσεις 17

Κεφάλαιο 1: Εισαγωγή́

1.1 Εισαγωγή στο πρόβλημα

Με τη σημερινή εξέλιξη της τεχνολογίας ο υπολογιστής μπορεί να βοηθήσει τον άνθρωπο από τον τομέα της καθημερινότητας, π.χ. αυτόματη ενεργοποίηση/απενεργοποίηση οικιακών συσκευών, μέχρι και τον επαγγελματικό ή επιστημονικό τομέα, π.χ. ένα μοντέλο μπορεί να κρίνει αν ο καρκίνος είναι καλοήθης ή κακοήθης. Επιπλέον ζούμε στην εποχή που με την τεχνολογία μπορούμε να βοηθήσουμε ανθρώπους με ακρωτηριασμένα μέλη να μπορέσουν να επανέλθουν. Η παρούσα εργασία στοχεύει στη δημιουργία και αξιολόγηση ενός μοντέλου παραγωγής γλώσσας (**Natural language processing - NLP**) και ειδικεύεται στην παραγωγή τίτλου σε εικόνες (**Image Captioning**). Η χρήση του ποικίλλει, μπορεί να βοηθήσει ανθρώπους με προβλήματα οπτικά δίνοντας τίτλους σε εικόνες και διαβάζοντας τους μέσω αναγνώστες οθόνης, οργανώνει σετ εικόνων παράγοντας τίτλους, βοηθάει αρθρογράφους στην παραγωγή τίτλων για άρθρα κτλ. Τέλος, στην παρούσα εργασία θέλουμε να βελτιώσουμε την απόδοση σε σύγκριση με άλλα μοντέλα αλλά και να παρουσιάσουμε μία βάση που θα μπορεί να βελτιωθεί περαιτέρω στο μέλλον.

1.2 Οργάνωση του τόμου

Στο κεφάλαιο 2 αναφέρονται οι **μετρικές** και το **σετ δεδομένων** (**Dataset**). Στο κεφάλαιο 3 αναφέρεται η δημιουργία και η λειτουργία του κώδικα. Τέλος, στο κεφάλαιο 4 σχολιάζονται τα αποτελέσματα, γίνεται σύγκριση με αποτελέσματα δημοσιεύσεων και προτείνονται πιθανές βελτιώσεις για το μέλλον.

Κεφάλαιο 2: Μετρικές και σετ δεδομένων

2.1 Μετρικές

Στον τομέα των μοντέλων παραγωγής γλώσσας (**Natural language processing - NLP)** υπάρχουν αρκετές μετρικές. Η γενική ιδέα είναι ότι η μετρική πρέπει να συσχετίζεται άμεσα με τον ανθρώπινο τίτλο. Όμως καμία μετρική δεν το έχει επιτύχει αυτό(ακόμα), γι’ αυτό και οι περισσότερες δημοσιεύσεις αναφέρονται σε δύο ή και περισσότερες μετρικές όταν παρουσιάζουν τα αποτελέσματα τους. Στην παρούσα εργασία τα αποτελέσματα θα ερμηνευτούν με δύο μετρικές την BLEU και την Meteor.

2.1.1 Bleu

Η μετρική Βleu (**Bilingual Evaluation Understudy)** αναφέρεται στην πλειοψηφία των δημοσιεύσεων με θέμα NLP [2] και είναι εύκολα προσβάσιμη. Η μετρική βασίζεται πάνω στην επιλογή ν λέξεων και παραμετροποιείται ως bleu-1 bleu-2…bleu-n ανάλογα με το **σύνολο λέξεων (dataset)** ή αλλιώς **ν-λέξεις** που επιλέγουμε. Ένα παράδειγμα είναι: “The boys are playing football”

1. **Bleu-1:**” The”, “boys”, “are”, “playing”, “football”
2. **Bleu-2:** “The boys “, “boys are “, “are playing “, “playing football “
3. **Bleu-3:** “The boys are”, “boys are playing “, “are playing football “
4. **Bleu-4:** “The boys are playing”, “boys are playing football “

Το πρώτο στάδιο είναι να βρεθεί το γινόμενο όλων των bleu-n μετρήσεων με την κάθε μέτρηση να έχει το ίδιο βάρος, το συνιστώμενο βάρος είναι 1/n. Για να βρεθεί η κάθε bleu-n μέτρηση, η αλλιώς, η κάθε **προσέγγιση** χρησιμοποιείται ο τύπος:

Σε συνέχεια από το παραπάνω παράδειγμα:

**Υποψήφια πρόταση**🡪 The boys are playing football

**Πρόταση αναφοράς**🡪 All boys are playing football

1. **P1**= 4/5
2. **P2**=3/4
3. **P3**=2/3
4. **P4**=1/2

Ο μαθηματικός τύπος είναι :

* Γεωμετρική μέση ακρίβεια προσέγγισης (**Geometric Average Precision- GAP**)

\*Εάν ένα σύνολο λέξεων δεν αντιστοιχεί σε κανένα αντίστοιχο σύνολο στην πρόταση αναφοράς δηλαδή , τότε GAP=0. Για να το αποτρέψουμε αυτό χρησιμοποιούμε μία συνάρτηση εξομάλυνσης (**Smoothing Function**), συγκεκριμένα σε όλες τις μετρήσεις της εργασίας χρησιμοποιήθηκε η 7η τεχνική εξομάλυνσης όπου και περιγράφεται στην δημοσίευση [3].

Αν το μοντέλο αντί για «The boys are playing football» προέβλεπε “the” τότε θα είχε score οπότε θα ενθαρρυνόταν να παράγει μικρές προτάσεις για να έχει μεγαλύτερη ακρίβεια. Για να αποτραπεί αυτό δημιουργήθηκε η ποινή συντομίας (**Brevity Penalty -BP**) η οποία έχει τύπο:

* **C**🡪 Σύνολο λέξεων στην υποψήφια πρόταση
* **R**🡪 Σύνολο λέξεων στην πρόταση αναφοράς

Τέλος ο τύπος για τη μετρική Bleu είναι ο εξής:

2.1.2 Meteor

Η μετρική Meteor (**Metric for Evaluation of Translation with Explicit Ordering**) όπως και η bleu αναφέρεται σε πολλές δημοσιεύσεις και δημιουργήθηκε για να αντιμετωπίσει τα προβλήματα που έχει η bleu [4] όπως:

1. Δε λαμβάνει υπόψη τη μετρική recall που βασίζεται στις σωστές απαντήσεις του μοντέλου και χρησιμοποιείται κυρίως σε ιατρικές προβλέψεις.
2. Χρήση μεγάλων συνόλων λέξεων.
3. Χρήση γεωμετρικής μέσης ακρίβειας προσέγγισης για ν-λέξεις, για το λόγο που εξηγήθηκε παραπάνω, ώστε εάν υπάρξει έστω και ένα τότε GAP=0

Ο υπολογισμός της μετρικής έχει ως εξής:

Υπολογίζουμε την **ακρίβεια (Precision),** όπου:

* **m**🡪 Το σύνολο των λέξεων της υποψήφιας πρότασης που βρίσκονται στην πρόταση αναφοράς.
* 🡪 Το σύνολο των λέξεων στην υποψήφια πρόταση.

Έπειτα υπολογίζουμε την **ανάκληση (Recall),** όπου:

* **m🡪** Το **m** ορίζεται όπως και πριν
* 🡪 To σύνολο των λέξεων στην πρόταση αναφοράς

Στη συνέχεια χρησιμοποιείται ο τύπος του αρμονικού μέσου με την ανάκληση να έχει βάρος εννιά φορές μεγαλύτερο της ακρίβειας.

Τέλος, όσο λιγότερες αντιστοιχίες ν-λέξεων υπάρχουν ανάμεσα στις προτάσεις τόσο μεγαλώνει η **ποινή** με μεγίστη τιμή 50%. Για να υπολογίσουμε την ποινή έχουμε,όπου:

* **c🡪** Ορίζεται το κομμάτι, το κομμάτι είναι το μέγιστο σύνολο λέξεων στην υποψήφια πρόταση που είναι ίδιο με το αντίστοιχο σύνολο στην πρόταση αναφοράς
* **um🡪** Ο αριθμός των ν-λέξεων που έχουν αντιστοιχιστεί

Ο τελικός τύπος είναι:

2.2 Flickr8k

Το σετ δεδομένων που χρησιμοποιήθηκε είναι το **Flickr8κ** [5], έχει συνολικά 8000 φωτογραφίες, 6000 για την εκπαίδευση (**train**), 1000 για την επιβεβαίωση (**validation**) και 1000 για τη δοκιμή (**test**). Η δομή του dataset είναι πέντε τίτλοι για την κάθε εικόνα συνολικά 40000 τίτλοι.

A text on a white background

Description automatically generated

Εικόνα 2: Δομή σετ-δεδομένων flickr8k

Κεφάλαιο 3: Σχεδίαση του Θέματος

3.1 Σχεδίαση και λειτουργία του λογισμικού

Η δημιουργία του λογισμικού για παραγωγή λεζάντας εμπεριέχει πολλά στάδια όπως φαίνεται και στο γενικό σχεδιάγραμμα (Σχήμα 1). Η διαδικασία ξεκινάει με την προ-επεξεργασία και κωδικοποίηση δεδομένων, έπειτα χρησιμοποιούμε ένα συνελικτικό νευρωνικό δίκτυο που λειτουργεί σαν κωδικοποιητής για να εξάγουμε το διάνυσμα χαρακτηριστικών της κάθε εικόνας. Τέλος, εκπαιδεύουμε ένα επαναλαμβανόμενο νευρωνικό δίκτυο που λειτουργεί σαν αποκωδικοποιητής με το διάνυσμα χαρακτηριστικών και τα κωδικοποιημένα δεδομένα και καταλήγουμε στην παραγωγή λεζάντας.

Παρακάτω θα εξηγηθούν όλα τα στάδια.

A computer screen with white text

Description automatically generated

Σχήμα : Διαδικασία παραγωγής τίτλου

3.2 Σετ Δεδομένων

Ο φάκελος με το σετ δεδομένων εμπεριέχει πέντε αρχεία:

1. Flickr8k\_Dataset 🡪 Φάκελος με όλες τις φωτογραφίες
2. Flickr\_8k.devImages.txt🡪 Αρχείο με τα ονόματα των 1000 εικόνων επιβεβαίωσης
3. Flickr\_8k.testImages.txt🡪 Αρχείο με τα ονόματα των 1000 εικόνων δοκιμής
4. Flickr\_8k.trainImages.txt🡪 Αρχείο με τα ονόματα των 6000 εικόνων εκπαίδευσης
5. Flickr8k.lemma.token.txt🡪 Αρχείο με τα ονόματα και τους ληματοποιημένους τίτλους όλων των εικόνων

3.2.1 Προεπεξεργασία δεδομένων

Τo σετ δεδομένων έχει ήδη δεχθεί **λημματοποίηση** (**lemmatization**). Η διαδικασία της ληματοποίησης είναι πολύ διαδεδομένη στον τομέα του NLP, μειώνει τις μοναδικές λέξεις του λεξιλογίου και διευκολύνει την εκπαίδευση του μοντέλου. H λογική της είναι ότι αφαιρούμε όλες τις καταλήξεις από τις λέξεις και αντικαθιστούμε τα άρθρα όπως “is, am, up, are” κτλ. με το ρήμα “be”.

Παράδειγμα:

**Αρχική πρόταση 🡪** Two people are in a small race car driving by a green hill.

**Λημματοποιημένη πρόταση🡪** Two person be in a small race car drive by a green hill.

Αφού έχουμε το λημματοποιημένο σετ δεδομένων προχωράμε σε περαιτέρω επεξεργασία.

* Μετατρέπουμε τα κεφαλαία σε μικρά
* Αφαιρούμε τα σημεία στίξης
* Αφαιρούμε όλους τους χαρακτήρες «s» που περισσεύουν από τα σημεία στίξης
* Προσθέτουμε στην αρχή κάθε πρότασης το πρόθεμα “<start>” και στο τέλος “<end>”



Εικόνα 3: Ο τίτλος της εικόνας πριν και μετά την επεξεργασία.

3.3 Συνελικτικό Νευρωνικό Δίκτυο

Στην παρούσα εργασία χρησιμοποιήθηκε το προ-εκπαιδευμένο νευρωνικό δίκτυο ResNet50 μέσω της βιβλιοθήκης Keras. Το μοντέλο εκπαιδεύτηκε στο σετ δεδομένων ImageNet [6] που αποτελείται 1000 κατηγορίες και πάνω από ένα εκατομμύριο εικόνες. Για τη χρήση του μοντέλου στην προκειμένη άσκηση καταργήθηκαν τα τελευταία πλήρως συνδεδεμένα επίπεδα (μαζί με τα βάρη τους) που συμβάλουν στην κατηγοριοποίηση, μέσω της βιβλιοθήκης Keras. Αυτό έγινε για να κρατήσουμε το διάνυσμα των 2048 χαρακτηριστικών που παράγει το μοντέλο, έτσι ώστε να το χρησιμοποιήσουμε στην εκπαίδευση του RNN.

Κώδικας:

model=ResNet50( include\_top=False, weights='imagenet', pooling='avg', input\_shape=( 224, 224, 3 ) )

3.4 Δημιουργία λεξικών

Για την οργάνωση των δεδομένων χρησιμοποιήθηκε η δομή του λεξικού στην python και δημιουργήθηκαν:

1. Τρία λεξικά **train\_captions**, **test\_captions**, **validation\_captions** με δείκτη το όνομα της εικόνας και αποτέλεσμα τους πέντε τίτλους.

**Κώδικας:** print( train\_captions[ “2764732789\_1392e962d0.jpg” ] )

**Αποτέλεσμα:**

A black background with white text

Description automatically generated

Εικόνα 4: Αποτέλεσμα λεξικού train\_captions.

1. Τρία λεξικά **train\_features**, **test\_features**, **validation\_features** με δείκτη το όνομα της εικόνας και αποτέλεσμα το διάνυσμα με 2048 χαρακτηριστικά που παράχθηκαν από το ResNet50.

**Κώδικας:** print( test\_features[ “2654514044\_a70a6e2c21.jpg” ] )

**Αποτέλεσμα:**



Εικόνα 5: Αποτελέσματα λεξικού test\_features

1. Δύο λεξικά, **words\_to\_indices** που όπως το όνομα αναφέρει έχει σαν δείκτη τη λέξη και αποτέλεσμα τον αριθμό και **indices\_to\_words** που έχει σα δείκτη τον αριθμό και αποτέλεσμα τη λέξη.

**Κώδικας:** print( words\_to\_indices[“be”] ) | print( indices\_to\_words[1711] )

**Αποτέλεσμα:** 1711 | be

1. Ένα λεξικό **train\_encoded\_captions** που έχει σα δείκτη το όνομα της εικόνας και σαν αποτέλεσμα τους πέντε τίτλους **κωδικοποιημένους** (**tokenized**), ο κάθε τίτλος είναι σε μορφή πίνακα, μεγέθους 40 και οι κενές θέσεις συμπληρώνονται με μηδενικά.

**Κώδικας:** print( train\_encoded\_captions[ “2501968935\_02f2cd8079.jpg” ] )

**Αποτέλεσμα:** A black background with white numbers

Description automatically generated

Εικόνα 6: Αποτέλεσμα λεξικού train\_encoded\_captions.

3.5 Μοντέλο RNN

Η αρχιτεκτονική του μοντέλου που δημιουργήθηκε είναι η εξής:

1. **Input\_1:** Επίπεδο εισόδου που δέχεται το διάνυσμα χαρακτηριστικών
2. **Input\_2:** Επίπεδο εισόδου που δέχεται έναν πίνακα με μέγεθος 40 που αντιπροσωπεύει τις 40 λέξεις
3. **Dropout\_1** **και** **2:** Επίπεδο κανονικοποίησης που εμποδίζει την υπερ-εκπαίδευση, λειτουργεί επιλέγοντας και μηδενίζοντας ένα κομμάτι από τα δεδομένα που εισάγουμε.
4. **Embedding\_1:** Επίπεδο που μετατρέπει όλους τους αριθμούς που αντιστοιχούν σε λέξεις σε **πυκνά διανύσματα** (**dense vectors**) μεγέθους 256
5. **LTSM\_1:** Long Short-Term Memory με μέγεθος 256. Είναι το κύριο επίπεδο του RNN μοντέλου, χρησιμοποιείται όταν υπάρχει αλληλουχία δεδομένων και μπορεί να ανακαλύψει τις συσχετίσεις μεταξύ τους με μεγάλη επιτυχία.
6. **Dense\_layer:** Το πυκνό επίπεδο συνδέει κάθε νευρώνα του τωρινού επιπέδου με τους νευρώνες του προηγούμενου και υπολογίζει τα βάρη με μία αλληλουχία προσθέσεων.Τα επίπεδα dense\_1 και 2 έχουν συνάρτηση ενεργοποίησης **RELU** και το 3 έχει τη **SOFTMAX**.
7. **Add\_1:** Συνδέει τα δύο inputs

A computer screen shot of a computer

Description automatically generated

Σχήμα :Επίπεδα Νευρωνικού δικτύου

3.5.1 Εκπαίδευση

Στο κομμάτι της εκπαίδευσης δημιουργήθηκε η συνάρτηση data\_generator η οποία επιστρέφει τα δεδομένα σε κομμάτια (batch). Κάθε κομμάτι αποτελείται από ένα δυσδιάστατο πίνακα και ένα μονοδιάστατο. Ο δυσδιάστατος πίνακας περιέχει τα 2048 χαρακτηριστικά της κάθε εικόνας και τους κωδικοποιημένους τίτλους της και ο μονοδιάστατος περιέχει ένα πυκνό διάνυσμα με τη θέση της κάθε λέξης στο κωδικοποιημένο τίτλο. Η εκπαίδευση του νευρωνικού ήταν αρκετά χρονοβόρα μιας και έγιναν πολλές επαναλήψεις με πολλές εποχές, η κάθε εποχή διαρκούσε περίπου δέκα λεπτά. Στο παρακάτω πινακάκι όπου αναφέρεται «s» σημαίνει ότι στους τίτλους διαγράφηκε το “s” ενώ όπου αναφέρεται “s+a” σημαίνει ότι διαγράφηκε το “s” και το “a”. Τέλος το κ είναι η βασική μεταβλητή του αλγορίθμου “beam search” όπου περιγράφεται παρακάτω.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Μοντέλο | Εποχές | Dataset | Score | Loss | K | Bleu-4 |
| modelLemma2\_12 | 12 |  | 0.8364 | 0.6631 | 3 | 0.264 |
| modelLemma2\_12 |  |  |  |  | 5 | 0.265 |
| modelLemma12 | 12 |  | 0.8404 | 0.6574 | 3 | 0.145 |
| modeLemma12 |  |  |  |  | 5 | 0.221 |
| modelLemma7 | 8 |  | 0.8202 | 0.7944 | 3 | 0.216 |
| modeLemma7 |  |  |  |  | 5 | 0.216 |
| model7 | 8 |  | 0.8226 | 0.7518 | 5 | 0.27 |
| model49 | 50 |  | 0.8683 | 0.4753 | 5 | 0.242 |
| Model35 | 36 |  | 0.8614 | 0.5097 | 5 | 0.245 |
| model15 | 16 |  | 0.84 | 0.6314 | 5 | 0.258 |
| model5 | 6 |  | 0.8159 | 0.8115 | 5 | 0.27 |
| model4 | 5 |  | 0.8119 | 0.8545 | 5 | 0.274 |

Πίνακας : Αποτελέσματα εκπαίδευσης

**Σχολιασμός:**

Από τον πίνακα 1 παρατηρούμε ότι αν και πολλές φορές το μοντέλο έχει μεγαλύτερο score δεν αποδίδει ικανοποιητικά λόγο της υπέρ-εκπαίδευσης, για έχουμε αύξηση στην ακρίβεια (αν και ελάχιστη) ενώ ταυτόχρονα βλέπουμε ότι όταν αφαιρούμε το “a” από τους τίτλους η ακρίβεια μειώνεται.

3.6 Beam Search

Ο αλγόριθμος Beam Search είναι ένας ευριστικός αλγόριθμος ο οποίος δοκιμάζει πιθανούς συνδυασμούς λέξεων ώστε να βρει το καλύτερο αποτέλεσμα.

Λειτουργεί ως εξής:

1. Δημιουργεί ένα δένδρο, με κορυφή τη λέξη “<start>” το οποίο στο πρώτο επίπεδο έχει όλες τις λέξεις ταξινομημένες με αύξουσα σειρά πιθανότητας. Να σημειωθεί ότι κάθε λέξη έχει μία πιθανότητα λόγο του ότι το νευρωνικό επιστρέφει ένα πίνακα πιθανοτήτων.
2. Έπειτα αναπτύσσει τις κ-λέξεις που έχουν τη μεγαλύτερη πιθανότητα και επιλέγει τη λέξη με τη μεγαλύτερη **conditional probability.**

Ο τύπος για την conditional probability είναι ο:

1. Τέλος, επαναλαμβάνει το δεύτερο βήμα μέχρι να συναντήσει τη λέξη <end> ή το σύνολο των λέξεων να φτάσει το μέγιστο μήκος(40).

Κεφάλαιο 4: Αποτελέσματα και συμπεράσματα

4.1 Αποτελέσματα

Οι μετρήσεις έγιναν σε bleu -1 bleu-3 bleu-4 και meteor score στο dataset Flickr8k.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Μοντέλα | Bleu-1 | Bleu-3 | Bleu4 | Meteor |
| m-RNN [7] | 57.78 | 23.07 | --------------------- | --------------------- |
| NIC[8] | 63 | --------------------- | --------------------- | --------------------- |
| [9] | 69.9 | 36.8 | 26.5 | 23 |
| Το δικό μας | 79.4 | 38 | 27.4 | 37.4 |

Πίνακας 2:Σύγκριση του μοντέλου με άλλες υλοποιήσεις

Τα παραπάνω αποτελέσματα δείχνουν ότι το μοντέλο μας έχει βελτίωση στο bleu-1 και στο bleu-4 η αλήθεια είναι πως στις δημοσιεύσεις που αναφέρονται δεν είναι ακριβής η διαδικασία προ-επεξεργασίας των τίτλων, όπως και ότι δεν αναφέρουν την τεχνική εξομάλυνσης που χρησιμοποιήθηκε, γι’ αυτό και η σύγκριση των αποτελεσμάτων δεν είναι έμπιστη. Από την άλλη, το μοντέλο μας υπερισχύει στη μετρική Meteor με μεγάλη διαφορά που μας δείχνει πως είναι πιο έμπιστο στην παραγωγή τίτλων.

4.1.1 Παραδείγματα τίτλων

Παρακάτω παραθέτουμε μερικά παραδείγματα εικόνων και τους αντίστοιχους τίτλους που παρήγαγε το μοντέλο.

Two dogs playing in the snow

Description automatically generated

Εικόνα 7: Παράδειγμα 1

A person in a swimming pool with a dog

Description automatically generated

Εικόνα 8: Παράδειγμα 2

A person and person dancing in a parade

Description automatically generated

Εικόνα 9: Παράδειγμα 3

A football player wearing a red uniform

Description automatically generated

Εικόνα 10: Παράδειγμα 4

A group of women in a crowd

Description automatically generated

Εικόνα 11: Παράδειγμα 5

4.2 Συμπεράσματα

Το μοντέλο που παρουσιάστηκε στην παρούσα ειδική εργασία έχει καλύτερα αποτελέσματα όσον αφορά τις μετρικές bleu-1, bleu-4 και Meteor από τις αναφερόμενες δημοσιεύσεις [7]–[9], αυτό όμως δε σημαίνει ότι είναι καλύτερο. Όπως προαναφέρθηκε ο τρόπος προ-επεξεργασίας είναι διαφορετικός και όχι μόνο αυτό αλλά φάνηκε στη δημοσίευση [8] όπου έβαλαν ανθρώπους να θέσουν τίτλους σε εικόνες και το τελικό bleu-1 score ήταν 70 το οποίο μας δείχνει ότι το μεγαλύτερο score του μοντέλου μας (79.4) δεν είναι απαραίτητα καλύτερο. Επιπλέον το μοντέλο παρατηρήθηκε ότι πολλές φορές βάζει μεγάλους τίτλους χωρίς να έχουν κάποιο νόημα. Γενικά θεωρούμε πως η παρούσα εργασία είναι μία καλή βάση ώστε να αναπτυχθεί μελλοντικά με περισσότερες βελτιώσεις αλλά και καλύτερη σύγκριση με Sota (State of the art) μοντέλα παραγωγής γλώσσας.

4.3 Μελλοντικές βελτιώσεις

Η παρούσα εργασία μπορεί να δεχτεί αρκετές βελτιώσεις. Μια ιδέα είναι να αλλάξει η δομή δεδομένων, από **λεξικά** να γίνουν **tensors** ώστε να τρέχει το μεγαλύτερο κομμάτι του προγράμματος στην κάρτα γραφικών. Να δοκιμαστούν μεγαλύτερα datasets όπως **Flickr30k** το οποίο περιέχει συνολικά 30,000 εικόνες ή το **MSCOCO** με 328,000 εικόνες. Τέλος μία σημαντική βελτίωση είναι να χρησιμοποιηθούν κι άλλες μετρικές όπως το **Sacrebleu** [10] που αντιμετωπίζει προβλήματα παραδείγματος χάρη την προ-επεξεργασία δεδομένων.

Βιβλιογραφία

[1] W. Williams, “The Psyonic Ability Hand,” *Bionics For Everyone*, Nov. 20, 2020. https://bionicsforeveryone.com/the-psyonic-ability-hand/ (accessed Aug. 30, 2023).

[2] “Papineni et al. - 2001 - BLEU a method for automatic evaluation of machine.pdf.” Accessed: Aug. 23, 2023. [Online]. Available: https://aclanthology.org/P02-1040.pdf

[3] “Chen and Cherry - 2014 - A Systematic Comparison of Smoothing Techniques fo.pdf.” Accessed: Aug. 23, 2023. [Online]. Available: https://aclanthology.org/W14-3346.pdf

[4] “Proceedings of the....pdf.” Accessed: Aug. 23, 2023. [Online]. Available: https://aclanthology.org/W05-0909.pdf

[5] M. Hodosh, P. Young, and J. Hockenmaier, “Framing Image Description as a Ranking Task Data, Models and Evaluation Metrics Extended Abstract”.

[6] “ImageNet,” *Wikipedia*. Jul. 22, 2023. Accessed: Aug. 25, 2023. [Online]. Available: https://en.wikipedia.org/w/index.php?title=ImageNet&oldid=1166513261

[7] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille, “Explain Images with Multimodal Recurrent Neural Networks.” arXiv, Oct. 04, 2014. Accessed: Aug. 27, 2023. [Online]. Available: http://arxiv.org/abs/1410.1090

[8] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and Tell: A Neural Image Caption Generator.” arXiv, Apr. 20, 2015. Accessed: Aug. 27, 2023. [Online]. Available: http://arxiv.org/abs/1411.4555

[9] S. He and Y. Lu, “A Modularized Architecture of Multi-Branch Convolutional Neural Network for Image Captioning,” *Electronics*, vol. 8, p. 1417, Nov. 2019, doi: 10.3390/electronics8121417.

[10] M. Post, “A Call for Clarity in Reporting BLEU Scores,” in *Proceedings of the Third Conference on Machine Translation: Research Papers*, Belgium, Brussels: Association for Computational Linguistics, 2018, pp. 186–191. doi: 10.18653/v1/W18-6319.