National Technical University of Athens
School of Electrical and Computer Engineering
MSc in Data Science and Machine Learning

Generation of Synthetic Referring Expressions for Object Segmentation in Videos

MASTER THESIS

IOANNIS V. KAZAKOS

Supervisor : Konstantinos Karantzalos
Associate Professor at NTUA

Athens, November 2020

National Technical University of Athens
School of Electrical and Computer Engineering
MSc in Data Science and Machine Learning

Generation of Synthetic Referring Expressions for Object Segmentation in Videos

MASTER THESIS

IOANNIS V. KAZAKOS

Supervisor : Konstantinos Karantzalos
Associate Professor at NTUA

Approved by the examining committee on the November 11, 2020.

.....	.....	.....
Konstantinos Karantzalos Associate Professor at NTUA	Xavier Giró-i-Nieto Associate Professor at UPC	Giorgos Stamou Associate Professor at NTUA

Athens, November 2020

.....

Ioannis V. Kazakos

MSc in Data Science and Machine Learning

Copyright © Ioannis V. Kazakos, 2020.
All rights reserved.

This work is copyright and may not be reproduced, stored nor distributed in whole or in part for commercial purposes. Permission is hereby granted to reproduce, store and distribute this work for non-propfit, educational and research purposes, provided that the source is acknowledged and the present copyright message is retained. Enquiries regarding use for profit should be directed to the author.

The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of the National Technical University of Athens.

Περίληψη

Η ενσωμάτωση της επεξεργασίας φυσικής γλώσσας στην όραση υπολογιστών έχει σημειώσει σημαντική πρόδοο τα τελευταία χρόνια χάρη στη συνεχή εξέλιξη της βαθιάς μηχανικής μάθησης. Ένα καινοτόμο πρόβλημα το οποίο συνδυάζει τη μηχανική όραση και την επεξεργασία φυσικής γλώσσας είναι αυτό της κατάτμησης αντικειμένων σε ακολουθίες εικόνων (βίντεο) με τη χρήση αναφορικών εκφράσεων, όπου μια πρόταση φυσικής γλώσσας καθορίζει ποιό αντικείμενο πρέπει να κατατμηθεί σε ένα βίντεο. Μια από τις μεγαλύτερες προκλήσεις αυτής της εφαρμογής είναι η έλλειψη συνόλων δεδομένων μεγάλης κλίμακας, εξαιτίας του υπερβολικά μεγάλου χρόνου και ανθρώπινης προσπάθειας που απαιτείται για τη συλλογή τους. Επιπλέον, τα υπάρχοντα σύνολα δεδομένων υποφέρουν από ετικέτες κακής ποιότητας καθώς, σύμφωνα με μια σχετική ανάλυση, περίπου μία στις δέκα εκφράσεις που περιέχονται σε αυτά αποτυγχάνουν να περιγράψουν μοναδικά το αντικείμενο-στόχο.

Ο σκοπός της παρούσας μεταπτυχιακής εργασίας είναι να αντιμετωπίσει αυτές τις προκλήσεις προτείνοντας μια καινοτόμο μέθοδο για την παραγωγή συνθετικών αναφορικών εκφράσεων για μια εικόνα (ενός καρέ του βίντεο). Η μέθοδος αυτή παράγει συνθετικές αναφορικές εκράσεις χρησιμοποιώντας μόνο τις ετικέτες αναφοράς των αντικειμένων μιας εικόνας ή βίντεο, καθώς και τα χαρακτηριστικά τους, τα οποία ανιχνεύονται από ένα υπερσύγχρονο βαθύ νευρωνικό δίκτυο εκπαιδευμένο για τον εντοπισμό αντικειμένων. Ένα από τα πλεονεκτήματα της προτεινόμενης μεθόδου είναι ότι ο ορισμός της επιτρέπει την εφαρμογή της σε οποιοδήποτε άλλο σύνολο δεδομένων εντοπισμού ή κατάτμησης αντικειμένων.

Χρησιμοποιώντας την προτεινόμενη μέθοδο, παρουσιάζεται το πρώτο μεγάλης κλίμακας σύνολο συνθετικών δεδομένων με αναφορικές εκφράσειες για κατάτμηση αντικειμένων σε βίντεο, βασισμένο σε ένα υπάρχον σύνολο δεδομένων κατάτμησης αντικειμένων σε βίντεο. Η παρούσα εργασία περιλαμβάνει στατιστική ανάλυση καθώς και σύγκριση του παραγόμενου συνόλου συνθετικών δεδομένων με υπάρχοντα σύνολα δεδομένων κατασκευασμένα από τον άνθρωπο.

Τα πειράματα που διεξήχθησαν σε τρία διαφορετικά σύνολα δεδομένων που έχουν χρησιμοποιηθεί για την κατάτμηση αντικειμένων σε βίντεο με τη χρήση αναφορικών εκφράσεων, αποδεικνύουν την αποτελεσματικότητα των παραγόμενων συνθετικών δεδομένων. Συγκεκριμένα, τα αποτελέσματα επιδεικνύουν ότι προ-εκπαιδεύοντας ένα βαθύ νευρωνικό δίκτυο με το προτεινόμενο σύνολο συνθετικών δεδομένων, είναι δυνατή η βελτίωση της γενίκευσης του δικτύου σε διαφορετικά σύνολα δεδομένων. Το συγκεκριμένο αποτέλεσμα έχει ακόμα μεγαλύτερη αξία αν αναλογιστεί κανείς ότι η επίτευξή του δεν συμπεριλαμβάνει κανένα επιπλέον κόστος για υποσημείωση δεδομένων από ανθρώπους.

Λέξεις κλειδιά

Όραση Υπολογιστών, Επεξεργασία Φυσικής Γλώσσας, Όραση και Γλώσσα, Αναφορικές Εκφράσεις, Κατάτμηση Αντικειμένων σε Βίντεο, Παραγωγή Συνθετικών Δεδομένων

Abstract

Integrating computer vision with natural language processing has achieved significant progress over the last years owing to the continuous evolution of deep learning. A novel vision and language task, which is tackled in the present Master thesis is referring video object segmentation, in which a language query defines which instance to segment from a video sequence. One of the biggest challenges for this task is the lack of relatively large annotated datasets since a tremendous amount of time and human effort is required for annotation. Moreover, existing datasets suffer from poor quality annotations in the sense that approximately one out of ten language expressions fails to uniquely describe the target object.

The purpose of the present Master thesis is to address these challenges by proposing a novel method for generating synthetic referring expressions for an image (video frame). This method produces synthetic referring expressions by using only the ground-truth annotations of the objects as well as their attributes, which are detected by a state-of-the-art object detection deep neural network. One of the advantages of the proposed method is that its formulation allows its application to any object detection or segmentation dataset.

By using the proposed method, the first large-scale dataset with synthetic referring expressions for video object segmentation is created, based on an existing large benchmark dataset for video instance segmentation. A statistical analysis and comparison of the created synthetic dataset with existing ones is also provided in the present Master thesis.

The conducted experiments on three different datasets used for referring video object segmentation prove the efficiency of the generated synthetic data. More specifically, the obtained results demonstrate that by pre-training a deep neural network with the proposed synthetic dataset one can improve the ability of the network to generalize across different datasets, without any additional annotation cost. This outcome is even more important taking into account that no additional annotation cost is involved.

Key words

Computer Vision, Natural Language Processing, Vision and Language, Referring Expressions, Video Object Segmentation, Synthetic Data Generation

Σύνοψη

Ο συνδυασμός της όρασης υπολογιστών και της επξεργασίας φυσικής γλώσσας έχει προσελκύσει το ενδιαφέρον της επιστημονικής κοινότητας τα τελευταία χρόνια, μιας και θεωρείται ένα σημαντικό βήμα προς τη δημιουργία αυτόνομων συστημάτων τεχνητής νοημοσύνης τα οποία θα είναι ικανά να αξιοποιούν και τα δύο είδη πληροφορίας για την επίλυση προβλημάτων του πραγματικού κόσμου [Hu16a, Yu18, Ye19]. Ένα παράδειγμα τέτοιου προβλήματος, με το οποίο καταπιάνεται η παρούσα μεταπτυχιακή εργασία, αποτελεί η κατάτμηση αντικειμένων σε εικόνες και βίντεο με τη χρήση αναφορικών εκφράσεων φυσικής γλώσσας. Ως αναφορική έκφραση ορίζεται μια πρόταση φυσικής γλώσσας αν και μόνο αν αποτελεί ακριβή περιγραφή ενός συγκεκριμένου και κανενός άλλου αντικείμενου που εμφανίζεται στην ίδια σκηνή [Reit92]. Το πρόβλημα αυτό χρησιμοποιεί σαν οδηγό μια αναφορική έκφραση που περιγράφει ένα μοναδικό αντικείμενο-στόχο προκειμένου να εντοπίσει σε επίπεδο εικονοστοιχείου το αντικείμενο διαχωρίζοντάς το από άλλα αντικείμενα του ίδιου ή άλλου τύπου.

Η πρόοδος αυτού του καινοτόμου ερευνητικού πεδίου έχει επωφεληθεί από την πρόσφατη πρόοδο της βαθιάς μηχανικής μάθησης η οποία για να είναι αποδοτική απαιτεί μεγάλο αριθμό δεδομένων. Ωστόσο, μια από τις κύριες προκλήσεις του προβλήματος με το οποίο ασχολείται η παρούσα εργασία είναι η έλλειψη τέτοιων μεγάλων συνόλων δεδομένων με βίντεο τα οποία να περιλαμβάνουν ταυτόχρονα ετικέτες αντικειμένων σε επίπεδο εικονοστοιχείου και εκφράσεις φυσικής γλώσσας, όπως είναι για παράδειγμα το RefCOCO [Kaze14] για στατικές εικόνες. Η δημιουργία τέτοιων συνόλων δεδομένων απαιτεί μεγάλη ποσότητα χρόνου και ανθρώπινης προσπάθειας και αυτό έχει ωθήσει την επιστημονική κοινότητα να επενδύσει σε μεθόδους όπως η ημι/αυτο-επιβλεπόμενη μάθηση και η χρήση συνθετικών δεδομένων. Τα συνθετικά δεδομένα έχουν χρησιμποποιηθεί αποτελεσματικά σε διάφορες ερευνητικές εργασίες τόσο στην όραση υπολογιστών σε προβλήματα όπως η εκτίμηση οπτικής ροής [Doso15], η ανίχνευση αντικειμένων [Peng15], η σημασιολογική κατάτμηση [Sale18] και η κατάτμηση αντικειμένων σε βίντεο [Khor19], όσο και σε προβλήματα που συνδυάζουν τη μηχανική όραση και την επεξεργασία φυσικής γλώσσας όπως η συλλογιστική μέσω εικόνων [Liu19] και η πλοήγηση μέσω όρασης και γλώσσας [Frie18].

Ακολουθώντας αυτή την κατεύθυνση, η παρούσα μεταπτυχιακή εργασία προτείνει μια καινοτόμο μέθοδο για την παραγωγή συνθετικών αναφορικών εκφράσεων για μια εικόνα (ενός καρέ του βίντεο), η οποία βασίζεται μόνο στις ετικέτες αναφοράς των αντικειμένων καθώς και στα χαρακτηριστικά τους, τα οποία ανιχνεύονται από ένα υπερσύγχρονο βαθύ νευρωνικό δίκτυο εκπαιδευμένο για τον εντοπισμό αντικειμένων [Ren15]. Η προτεινόμενη μέθοδος εφαρμόζεται σε ένα υπάρχον μεγάλης κλίμακας σύνολο δεδομένων κατάτμησης αντικειμένων σε βίντεο, το YouTube-VIS [Yang19], το οποίο χάρη σε αυτή τη μέθοδο εμπλουτίζεται με συνθετικές αναφορικές εκφράσεις, χωρίς κανένα κόστος που να αφορά ανθρώπινη εργασία. Με την προτεινόμενη μέθοδο, είναι δυνατή η δημιουργία πολλαπλών αναφορικών εκφράσεων για το ίδιο αντικείμενο σε κάθε καρέ του βίντεο συνδυάζοντας διαφορετικές ιδιότητες του αντικειμένου όπως η κλάση στην οποία ανήκει, τα χαρακτηριστικά του (π.χ. το χρώμα του), η σχετική του θέση και το σχετικό του μέγεθος ως προς άλλα αντικείμενα. Ένα παράδειγμα διαφορετικών αναφορικών εκφράσεων οι οποίες παράγονται με την προτεινόμενη μέθοδο απεικονίζεται στην Εικόνα 0.1.

Το παραγόμενο σύνολο δεδομένων, ονόματι SynthRef-YouTube-VIS, αποτελεί το πρώτο μεγάλης κλίμακας σύνολο συνθετικών αναφορικών εκφράσεων για κατάτμηση αντικειμένων σε βίντεο αποτελούμενο από 2,238 βίντεο και 15,798 διαφορετικές συνθετικές αναφορικές εκφράσεις. Η προτεινόμενη μέθοδος αλλά και το εν λόγω σύνολο δεδομένων αξιολογούνται μέσω πειραμάτων τα οποία πραγματοποιούνται χρησιμποιώντας ένα βαθύ νευρωνικό δίκτυο που ονομάζεται RefVOS [Bell20]. Το δίκτυο

Figure 0.1: Example of synthetic referring expressions automatically generated with the proposed method. Multiple referring expressions can be created for the same video or even for the same frame.

αυτό χρησιμοποιεί δύο υπερσύγχρονα μοντέλα για την κωδικοποίηση των εικόνων (καρέ του βίντεο) και των αναφορικών εκφράσεων. Συγκεκριμένα, το DeepLabv3 [Chen17b] που έχει χρησιμοποιηθεί με μεγάλη επιτυχία στη σημασιολογική κατάτμηση, κωδικοποιεί την οπτική είσοδο, και το BERT [Devl19], ένα από τα πιο επιτυχημένα μοντέλα αναπαράστασης φυσικής γλώσσας, κωδικοποιεί τις αναφορικές εκφράσεις. Ο κατάλληλος συνδυασμός των εξαγόμενων οπτικών και γλωσσικών χαρακτηριστικών παράγει την τελική κατάτμηση του αντικειμένου-στόχου.

Τα πειράματα που πραγματοποιήθηκαν στην παρούσα μεταπτυχιακή εργασία είναι δύο ειδών:

1. Το πρώτο αφορά στην προ-εκπαίδευση του μοντέλου με τη χρήση πραγματικών (παραγόμενες από ανθρώπους), συνθετικών (παραγόμενες με την προτεινόμενη μέθοδο) ή συνδυασμό και των δύο ειδών αναφορικών εκφράσεων και την αξιολόγησή του σε διαφορετικά σύνολα δεδομένων.

2. Το δεύτερο πείραμα αποσκοπεί στην απευθείας σύγκριση πραγματικών και συνθετικών αναφορικών εκφράσεων μέσω της εκπαίδευσης του μοντέλου στο ίδιο σύνολο δεδομένων βίντεο αφενός με πραγματικές εκφράσεις και αφετέρου με συνθετικές, και της αξιολόγησής του στο ίδιο σύνολο πραγματικών δεδομένων.

Τα αποτελέσματα του πρώτου πειράματος επιδεικνύουν ότι προ-εκπαιδεύοντας ένα βαθύ νευρωνικό δίκτυο με το προτεινόμενο σύνολο συνθετικών δεδομένων, είναι εφικτή η βελτίωση της ικανότητας γενίκευσης του δικτύου σε διαφορετικά σύνολα δεδομένων, ειδικά στην περίπτωση που τα συνθετικά δεδομένα χρησιμοποιούνται σε συνδυασμό με πραγματικά. Επίσης, ακόμα μεγαλύτερη βελτίωση όσον αφορά την ακρίβεια κατάτμησης εντοπίζεται όταν το προ-εκπαιδευμένο μοντέλο χρησιμοποιείται σε ένα διαφορετικό σύνολο δεδομένων από αυτό στο οποίο έχει εκπαιδευτεί. Αυτό το αποτέλεσμα είναι σημαντικό γιατί σε πολλές εφαρμογές του πραγματικού κόσμου, τα μοντέλα μηχανικής μάθησης δεν έχουν τη δυνατότητα να εκπαιδεύονται στο τελικό σύνολο δεδομένων, αλλά βασίζονται σε μεγάλο βαθμό στην προ-εκπαίδευση.

Από την άλλη, τα αποτελέσματα της σύγκρισης μεταξύ των πραγματικών και συνθετικών αναφορικών εκφράσεων, η οποία διεξάγεται στο δεύτερο πείραμα, οδηγεί στο συμπέρασμα ότι οι πραγματικές εκφράσεις, όντας πιο πλούσιες στην περιγραφή των αντικειμένων, οδηγούν σε καλύτερα αποτελέσματα. Ωστόσο, αν αναλογιστεί κανείς το μεγάλο κόστος για την συλλογή των πραγματικών εκφράσεων και το αντίστοιχο μηδενικό για τη δημιουργία των συνθετικών, τα αποτελέσματα είναι συγκρίσιμα. Επίσης, γίνεται μελέτη της επίδρασης της πληροφορίας που εμπεριέχεται στις συνθετικές αναφορικές εκφράσεις, όπου φαίνεται ότι όσο περισσότερες ιδιότητες του αντικειμένου συμπεριλαμβάνονται

(όπως για παράδειγμα η σχετική του θέση ή το χρώμα του), τόσο βελτιώνεται η τελική ακρίβεια κατάτμησης.

Τέλος, η παρούσα μεταπτυχιακή εργασία ενθαρύννει την περαιτέρω επέκτασή της προτείνοντας μελλοντικές κατευθύνσεις. Αυτές αφορούν πρώτον στην εφαρμογή της προτεινόμενης μεθόδου σε άλλα σύνολα δεδομένων εντοπισμού και κατάτμησης αντικειμένων σε εικόνες και βίντεο και δεύτερον στην ενίσχυση της προτεινόμενης μεθόδου για την παραγωγή πιο πλούσιων συνθετικών εκφράσεων με την εισαγωγή άλλων στοιχείων όπως οι σχέσεις μεταξύ των αντικειμένων που εμφανίζονται.

Ευχαριστίες

Ioannis V. Kazakos,

Αθήνα, 11η Νοεμβρίου 2020

Η εργασία αυτή είναι επίσης διαθέσιμη ως Τεχνική Αναφορά CSD-SW-TR-42-17, Εθνικό Μετσόβιο Πολυτεχνείο, Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών, Εργαστήριο Τεχνολογίας Λογισμικού, Νοέμβριος 2020.

URL: `http://www.softlab.ntua.gr/techrep/`
FTP: `ftp://ftp.softlab.ntua.gr/pub/techrep/`

Acknowledgements

Ioannis V. Kazakos,

Athens, November 11, 2020

This thesis is also available as Technical Report CSD-SW-TR-42-17, National Technical University of Athens, School of Electrical and Computer Engineering, Department of Computer Science, Software Engineering Laboratory, November 2020.

URL: `http://www.softlab.ntua.gr/techrep/`
FTP: `ftp://ftp.softlab.ntua.gr/pub/techrep/`

Contents

Περίληψη 5

Abstract 7

Σύνοψη 9

Ευχαριστίες 13

Acknowledgements 15

Contents 17

List of Figures 19

1. Introduction 21
- 1.1 Vision and Language Integration 21
- 1.2 Referring Video Object Segmentation 24
- 1.3 Main Challenges and Motivation 24
- 1.4 Thesis Objectives and Structure 26

2. Literature Review 29
- 2.1 Referring Image Segmentation 29
  - 2.1.1 Methods 29
  - 2.1.2 Relevant datasets 30
- 2.2 Referring Video Object Segmentation 31
  - 2.2.1 Methods 31
  - 2.2.2 Relevant datasets 33
- 2.3 Object Detection 35
- 2.4 Synthetic Data 38

3. Proposed Method and Generated Dataset 41
- 3.1 YouTube-VIS Dataset 41
- 3.2 Proposed Method 42
  - 3.2.1 Object class 42
  - 3.2.2 Relative size 42
  - 3.2.3 Relative location 43
  - 3.2.4 Attributes 44
  - 3.2.5 Synthetic Referring Expressions 44
- 3.3 SynthRef-YouTube-VIS Dataset 45

4. Experimental Results 49
4.1 Training Details 49
4.1.1 Pre-training 49
4.1.2 Fine-tuning on the Evaluation Datasets 50
4.2 Quantitative Evaluation Metrics 50
4.3 Quantitative Results 51
4.4 Qualitative Results 55
4.5 Ablation Study 55

5. Conclusions and Future Directions 57
5.1 Conclusions 57
5.2 Future Directions 57

Bibliography 59

List of Figures

0.1 Example of synthetic referring expressions automatically generated with the proposed method. Multiple referring expressions can be created for the same video or even for the same frame. 10

1.1 The task of referring video object segmentation. Top: A referring expression and a video are given as input. Bottom: A segmentation mask of the referent (highlighted in red) is produced at every frame. The provided referring expression is from the Refer-YouTube-VOS dataset [Seo20]. 21
1.2 Different tasks combining vision and language [Moga19]. 23
1.3 Example of an invalid referring expression from A2D Sentences [Gavr18] dataset. The expression on top of the video frame fails to uniquely identify a specific object. . 25
1.4 Categorization of referring expressions by their difficulty and correctness in the validation set of DAVIS-2017 and the test set of A2D Sentences [Bell20]. 26

2.1 Model used by Hu *et al.* [Hu16a] who introduced the task of referring image segmentation. 29
2.2 Illustration of atrous convolution with rates 1 (standard convolution), 6 and 24. . . . 32
2.3 Architecture of the RefVOS model [Bell20]. 33
2.4 Overview of R-CNN [Girs14] and Fast R-CNN [Girs15] two-stage frameworks for object detection. 36
2.5 Architecture of Faster R-CNN [Ren15]. 36
2.6 Examples from the PhraseCut dataset [Wu20]. Referring phrases are produced by combining object categories (brown text), attributes (blue text) and relationships (green text) with other objects. 38

3.1 Histogram of object instances per video for the YouTube-VIS [Yang19] dataset. . . . 41
3.2 Overview of the proposed method for generating synthetic referring expressions. Top: Ground truth labels (object class + bounding boxes) are used to compute a target object's relative location and size. Bottom: A Faster R-CNN object detector with attribute head outputs attributes for the detected objects, which are filtered by the ground truth annotations. The combined cues create a set of referring expressions that uniquely identify the target object. 42
3.3 Example of synthetic referring expressions automatically generated with the proposed method. Multiple referring expressions can be created for the same video or even for the same frame. 44
3.4 Comparison of human-collected referring expressions from Refer-YouTube-VOS [Seo20] with synthetic ones generated with the proposed method. 45
3.5 Histogram of object instances by each class in SynthRef-YouTube-VIS train and validation sets. 46

4.1 Visual explanation of the Intersection-over-Union (IoU) or Jaccard Index (J) 51
4.2 Venn diagram of the object classes in DAVIS-2017 training and validation sets. . . . 52

4.3 Qualitative results on DAVIS-2017. Subfigure 4.3a (left) shows results when the model is pre-trained only on RefCOCO, while Subfigure 4.3b (right) when it is also trained on the proposed synthetic dataset. 54
4.4 Qualitative results on A2D Sentences. The model in the left subfigure is pre-trained only on RefCOCO, while the model in the right subfigure is also trained with the generated synthetic referring expressions. 54

Chapter 1

Introduction

Inspired by the great success of deep learning in the fields of computer vision (CV) and natural language processing (NLP), the research community has invested on the integration of the aforementioned fields, by proposing several vision and language tasks and by trying to build models capable of combining visual and linguistic information effectively. A recently proposed vision and language task, addressed in the present Master thesis, is *referring video object segmentation* in which, given a linguistic phrase and a video, the goal is to generate a binary mask for the referred object in all the video frames where it is present. A visual description of the aforementioned task is provided in Figure 1.1.

1.1 Vision and Language Integration

Recent advancements in deep learning research has led the fields of computer vision and natural language processing see a significant progress in several tasks independently. This success has also increased the interest in solving challenges that combine visual and linguistic information, *i.e.* the integration of vision and language. Integrating vision and language is considered an important step towards the creation of powerful artificial intelligence (AI) systems that will be able to reason by processing multi-modal input.

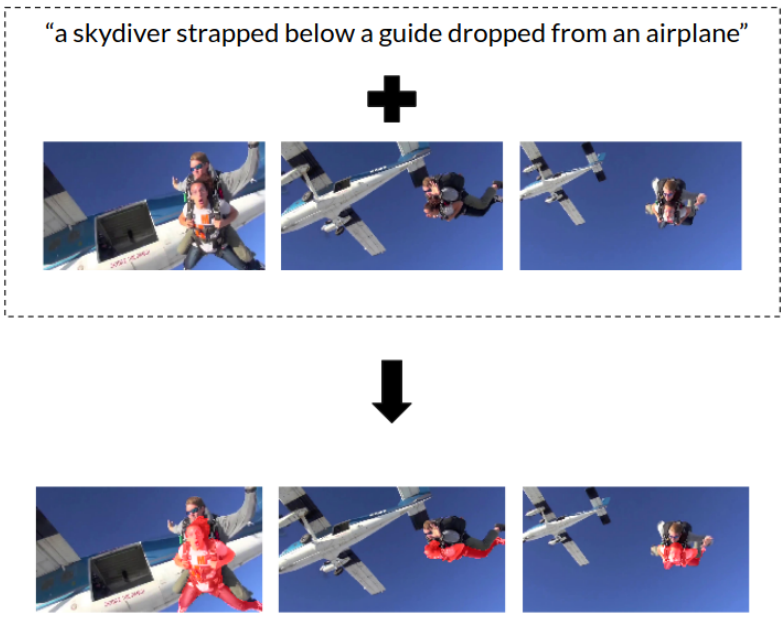

Figure 1.1: The task of referring video object segmentation. Top: A referring expression and a video are given as input. Bottom: A segmentation mask of the referent (highlighted in red) is produced at every frame. The provided referring expression is from the Refer-YouTube-VOS dataset [Seo20].

After more than half a century of research in neural networks and machine learning, deep learning has been establishing as the state-of-the-art technique of artificial intelligence since its breakthrough in 2012, when Krizhevsky et al. [Kriz12] presented a deep neural network, called AlexNet, which outperformed by a large margin all previous techniques in the Large Scale Visual Recognition Challenge (LSVRC). The release of large, high-quality, publicly available labelled datasets like ImageNet [Deng09], along with the empowerment of parallel GPU computing, which enabled the transition from CPU-based to GPU-based training, has led to the domination of deep learning in numerous AI fields, including computer vision and natural language processing.

More specifically, by using deep learning, computer vision has achieved prominent improvements in tasks such as visual content classification [Kriz12, Simo14, He16], object detection [Ren15, Redm16], semantic [Chen17a], instance [He17] and video object segmentation [Cael17, Vent19]. Convolutional neural networks (CNNs) [Fuku80, LeCu90] have become the standard approach for solving computer vision tasks. Most of the techniques rely on transferable general visual features by leveraging tasks such as image classification, detection, semantic segmentation, and action recognition. Usually, most preferred transferable global image representations are learned with deep CNN architectures like VGG [Simo14] and ResNets [He16] using large datasets such as ImageNet [Deng09]. These networks are used as the backbone of task-specific networks which transfer and enhance the obtained feature representations for solving downstream tasks.

Besides computer vision, deep learning has contributed to the significant progress in NLP research and its applications. For a long time, the majority of methods applied to NLP problems employed hand-crafted features using n-grams and bag-of-words [Joac98] models or standard machine learning techniques like Support Vector Machines (SVMs) [Cort95]. Such methods had been facing problems such as the curse of dimensionality since linguistic information was represented with high-dimensional features. However, with the recent popularity and success of word embeddings like word2vec [Miko13], which are low dimensional, distributed representations, deep neural networks have achieved superior results on various language-related tasks as compared to previously used techniques.

Similar to CNNs for computer vision, several neural network architectures and techniques have been established in NLP research such as Recurrent Neural Networks (RNNs) [Rume86], Long Short-Term Memory (LSTM) [Hoch97] and attention mechanism [Vasw17] in order to efficiently capture context in textual information. Especially in the last years, NLP has focused its efforts in solving multiple tasks at once with unsupervised pre-training of deep generalized language models like ELMo [Pete18], GPT-3 [Radf18] and BERT [Devl19], using large unlabeled corpora such as Wikipedia articles. These models have achieved incredible results in a wide variety of tasks such as machine translation, question answering and language inference.

Encouraged by the independent success of deep learning in CV and NLP fields, the research community has endeavored to build models combining vision and language. The aim of this integration is to produce systems which are able to provide complete understanding of visual and textual content at the same time. Several of the most important challenges that such systems have to tackle include:

Generation of textual descriptions about visual content and vice versa, *i.e.* generation of visual content from textual descriptions

Identification of objects and their relationships in visual content for reasoning or answering questions about them

Navigation in an environment by leveraging input from both vision and natural language instructions

Generation of short captions or longer stories about visual content

Translation of textual content from one language to another with visual content used for disambiguation

The aforementioned challenges can be associated to many practical applications of vision and language. One possible application in the biomedical domain can be the assistance of visually impaired individuals to get a holistic visual scene understanding by getting information about a scene from its textual descriptions and by answers received when asking questions about it. Other applications include automatic surveillance, autonomous driving, human-computer interaction and navigation.

Several tasks integrating language and vision have been proposed during the past years. An overview of them is depicted in Figure 1.2. These tasks include language observed in different levels such as words, phrases, sentences, paragraphs and documents while visual information is represented with images or videos. A brief description of the tasks presented in Figure 1.2 is provided below:

Referring Expression Generation and Comprehension/Segmentation: Referring expression generation focuses on the creation of referring expressions (noun phrases) that identify specific entities called targets or referents [Mao16]. The inverse task is comprehension where a target objects must be localized [Hu16b] or segmented [Hu16a] based on such expressions.

Visual Description Generation (Captioning): The goal of visual description generation or image/video captioning is to generate either global or dense descriptions to a given visual input in the form of a sentence [Elli13].

Visual Storytelling: The aim of visual storytelling is to generate stories from one or more images or a video. Visual storytelling extends visual description generation by creating several sentences forming something similar to a paragraph [Huan16].

Visual Question Answering: The goal of visual question answering (VQA) is to learn a model which comprehends the visual content at both global and local-level for finding an association with pairs of questions and answers in the natural language form [Anto15].

Visual Dialogue: The goal of the visual dialogue task is to create an AI agent which, given an image, a history about dialogues and a question about the image, is able to infer context from the history, and answer the question accurately [Das17].

Visual Reasoning: Visual reasoning targets to answer sophisticated queries by reasoning about the visual world. Efforts in this task have focused on creating diagnostic tests going beyond benchmarks such as VQA and reducing the biases of question-answer pairs by having detailed annotations describing the kind of reasoning each question requires [John17].

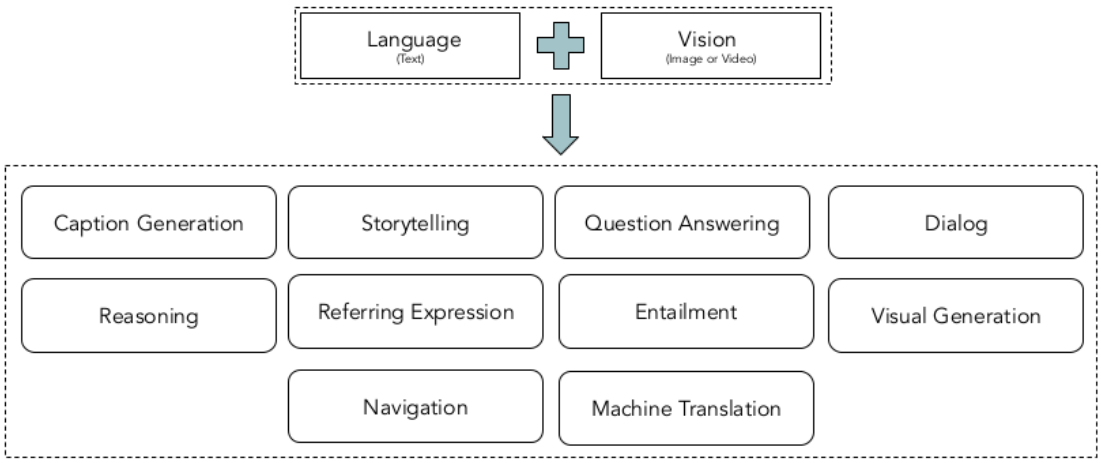

Figure 1.2: Different tasks combining vision and language [Moga19].

Image Entailment: The task of predicting whether an image semantically entails a text, given image-sentence pairs where premise is defined by an image instead of a natural language sentence [Xie19].

Language-to-Vision Generation: The aim of this task is to generate images/videos conditioned on natural language descriptions. The rapid evolution of generative adversarial networks (GANs) [Good14] has helped the growth of this task [Mans15].

Vision-and-Language Navigation: The goal of vision-and-language navigation is to enable an agent or a robot to navigate within an environment given by the photo-realistic image views by interpreting natural language instructions [Ande18].

Multi-modal Machine Translation: This task refers to the translation of a source language description into a target language using image/video information as additional context [Spec16].

1.2 Referring Video Object Segmentation

Video object segmentation is the task whose goal is to separate foreground objects from the background throughout a video sequence. This task has attracted wide attention lately due to its applicability to many practical problems including video analysis and video editing. Typically, this task has been addressed in semi-supervised or unsupervised setups. In the first case, a user manually annotates an object in a video frame and a system generates a pixel-wise binary mask for the object in the rest of the frames. The drawback of this setting is that pixel-wise annotations involve tedious and time-consuming human effort. In the unsupervised scenario [Goel18, Wang19b], estimation of object masks is performed without any guidance, by using salient features, independent motions, or known class labels. Although such approaches may be suitable for video analysis, according to Seo *et al.* [Seo20], the ambiguity and the lack of flexibility in defining foreground objects make them unsuitable for video editing which requires to segment arbitrary objects or their parts flexibly.

As an alternative approach, language referring expressions have been proposed as a different form of supervision for the task of video object segmentation. Referring expressions are linguistic expressions that allow the identification of an individual object (the *referent*) in a discourse or scene. According to computational linguistics and natural language processing community, a (noun) phrase is considered as a referring expression if it is an accurate description of the referent, but not of any other object in the current scene [Reit92]. Such linguistic expressions allow a more natural and direct human-computer interaction than interactive annotations in form of bounding boxes, masks, scribbles or points. Also, such expressions could be parsed from human speech processing systems allowing a direct human-machine communication in applications such as autonomous driving where the driver would refer to an object in the road scene and the computer would identify it.

The task of video object segmentation using referring expressions is a novel task first addressed by Khoreva *et al.* [Khor18] in 2018 and later tackled in a similar setting from Gavrilyuk *et al.* [Gavr18] and Wang *et al.* [Wang19a] as "actor-action segmentation from a sentence". The name *referring video object segmentation*, in correspondence to referring image segmentation, was introduced by Seo *et al.* [Seo20] who also released the first large-scale benchmark for the task, called *Refer-YouTube-VOS*, in a concurrent work to the Master present thesis. An example illustrating the task of referring video object segmentation is provided in Figure 1.1.

1.3 Main Challenges and Motivation

Despite the increasing interest and research in the field, referring image and video object segmentation remains an extremely challenging task which is still far from being solved. The main challenges for the task presented below, are divided into those concerning data and those concerning models. In terms of models used for referring video object segmentation, challenges include:

Temporal consistency: In contrast to static images, where the task of referring image segmentation has achieved significant progress, the video domain is more challenging as objects appearing in a video frame may disappear, reappear or be occluded from other objects. Consistency of segmentation masks across video frames is a challenging task and previous works have employed different techniques in order to achieve it. Recurrent architectures [Seo20] and 3D CNNs [Gavr18, Wang19a] are some of the ways previous works have tackled this challenge. Also, Khoreva *et al.* [Khor18] include a temporal consistency score in their objective function used for computing box proposals in each frame, based on the assumption that objects tend to move smoothly, and thus box proposals in consecutive frames should have a high overlap.

Model size: Models used for the task of referring video object segmentation suffer from excessive size in terms of parameters which leads to huge memory requirements and the need of a great amount of time to train. The combination of vision and language demands at least two branches for encoding the image (video frame) and linguistic (referring expression) data, usually a deep CNN and a LSTM. Especially in videos, where the time dimension is added, architectures can be even more complex in order to achieve temporal consistency as described above. Also, as presented in Chapter 2, many of these architectures use attention to effectively capture the dependencies between visual and linguistic features, adding extra layers and parameters to the deep architectures. This is a significant challenge considering the embedding of such models in cars or mobile phones which have restricted memory and computational resources.

The present Master thesis focuses on the challenges concerning the limitations of currently available datasets for the task of referring video object segmentation. In particular, the main challenges are:

Lack of large-scale datasets: Before the release of Refer-YouTube-VOS [Seo20], which was created concurrently with the present Master thesis, no large-scale dataset existed for the task of referring video object segmentation. As it is explained in Subsection 2.2.2, existing datasets used for language-guided video object segmentation were either small in the number of videos [Khor18] or object classes [Gavr18]. Especially in deep learning and computer vision research, large datasets and benchmarks have proven their fundamental importance, enabling targeted progress and objective comparisons, thus their absence can impend the evolution of a scientific field. The annotation cost in terms of money and/or time is one of the main reasons for the absence of large-scale datasets. The present work addresses this challenge by proposing a method to automatically generate synthetic referring expression, thus eliminating human labour-intensive annotations.

Figure 1.3: Example of an invalid referring expression from A2D Sentences [Gavr18] dataset. The expression on top of the video frame fails to uniquely identify a specific object.

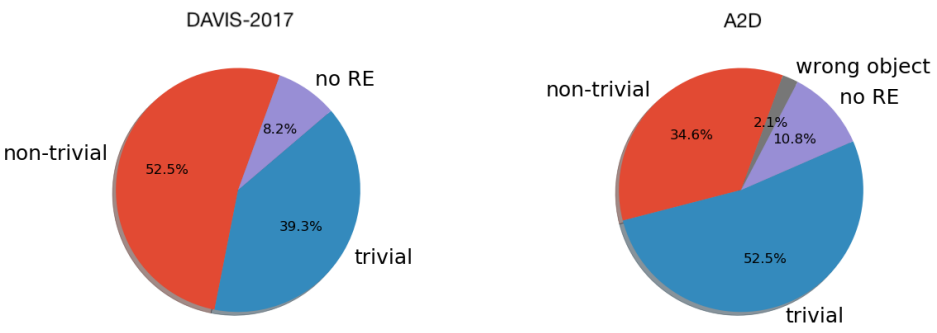

Figure 1.4: Categorization of referring expressions by their difficulty and correctness in the validation set of DAVIS-2017 and the test set of A2D Sentences [Bell20].

Referring expressions quality: A previous work [Bell20] has argued that existing benchmarks used for video object segmentation with referring expressions are annotated with expressions that sometimes fail to address the objective of the task, namely to unambiguously refer to a specific object. Actually, because of the huge annotation time of large-scale dataset, language expressions in existing benchmarks for referring image and video object segmentation are mainly collected through crowdsourcing platforms like Amazon Mechanical Turk[1]. Although the crowdsourced annotations are usually validated from other experts, cases of bad annotations are still observed, like in the example of A2D Sentences [Gavr18], appearing in Figure 1.3. In this example, the provided referring expression (*ñthe man in red is running in a raceò*) fails to unambiguously refer to a specific object, as it could be referring to three different instances present in the appearing video frame. An extensive analysis of the aforementioned work, illustrated in Figure 1.4, shows that approximately 10% of the referring expressions of two existing datasets used for the addressed task are invalid.

As it is also observed in Figure 1.4, a significant proportion of the videos in existing datasets concerns trivial cases in which the target object could be identified with simple phrases. For example, in a video including one *dog* and one *ball*, each of them can be referred unambiguously, using just the class or supercategory in which they belong *i.e.* simply saying *ña dog/animalò* and *ña ballò*.

Inspired by this simple scenario and the fact that existing datasets for object detection/segmentation are labeled in terms of the objects class, the idea of the present work is to create high quality synthetic referring expressions, starting from the referent's class and then enhancing them with other cues, without any human annotation cost. The proposed synthetic referring expressions are created on top of the YouTube-VIS [Yang19] dataset, which is described in detail in Section 3.1. The main advantage of this dataset is that all instances of a specified set of classes are annotated, allowing thus the creation of valid referring expressions.

1.4 Thesis Objectives and Structure

The present Master's thesis has the following objectives:

1. Study current methods in referring video object segmentation by reviewing related literature.
2. Underline the main challenges encountered on this task.
3. Propose a novel method for generating synthetic referring expressions.

[1] https://www.mturk.com/

4. Present and disseminate the first large-scale synthetic dataset for referring video object segmentation.

5. Evaluate the effectiveness of the proposed synthetic data in pre-training a deep neural network for the current task.

6. Compare the obtained performance using synthetic data with previous works.

7. Suggest future research directions regarding the use of synthetic referring expressions for video object segmentation.

Related literature is reviewed in *Chapter 2* where previous research works in the task of referring image and video object segmentation are explored by analyzing different techniques employed for solving the task. In the last section of this chapter, examples of scientific works using synthetic data in computer vision, natural language processing and their combination are presented.

Following, *Chapter 3* describes in detail the proposed method for generating synthetic referring expressions by explaining how different cues are combined for their creation. Moreover, Chapter 3 introduces the synthetic dataset created using the aforementioned method and includes an analysis of its statistics as well as some examples of synthetic referring expressions with their corresponding video frames.

Chapter 4 consists of an extensive analysis of the conducted experiments and the obtained results. More specifically, the datasets and metrics used in the experiments are described in detail and tables and figures of results comparing results with previous works are illustrated.

Finally, in Chapter 5, the conclusions of the present thesis are summarized and future research directions for the topic under study are suggested.

Chapter 2

Literature Review

2.1 Referring Image Segmentation

2.1.1 Methods

Referring image segmentation, the task of segmenting objects or regions in images given a linguistic expression, was introduced by Hu *et al.* [Hu16a]. The authors distinguish this task from previous ones that were restricted to a fixed set of classes, like semantic segmentation [Long15, Chen17a], the task of predicting pixel-wise labels for a predefined set of object or stuff categories, or instance segmentation [He17], which additionally distinguishes different instances of an object class. Previous works about grounding natural language expressions were limited to only resolving a bounding box in an image [Hu16b, Mao16], therefore this was the first attempt of grounding language at pixel level.

The model they employ for solving this novel tasks consists of four main components which are depicted in Figure 2.1. The first is a language encoder based on a LSTM network. The input language expression is first converted into a sequence of fixed-length vectors using an embedding matrix. Then each of the word embeddings of the sequence $S = (w_1; :::; w)$ is processed by the LSTM network at each time step t. At the final time step $t = T$, when the the whole text sequence is processed by the LSTM, the hidden state h of dimension $D_{text} = 1000$ is used as the encoded vector representation of the language expression. The second and third components of the model are two fully convolutional neural networks where the first of them is used as the image encoder and the second as a pixel classification network. The image encoder is a fully convolutional network as the one proposed by Long *et al.* [Long15] for semantic segmentation which, given an image of input $W \quad X$, outputs a spatial feature map of dimension $w \quad x \quad D_{im}$. This means that the final spatial feature map includes $D_{im} = 1000$ local descriptors for each pixel of the pooled $w \quad x$ image where

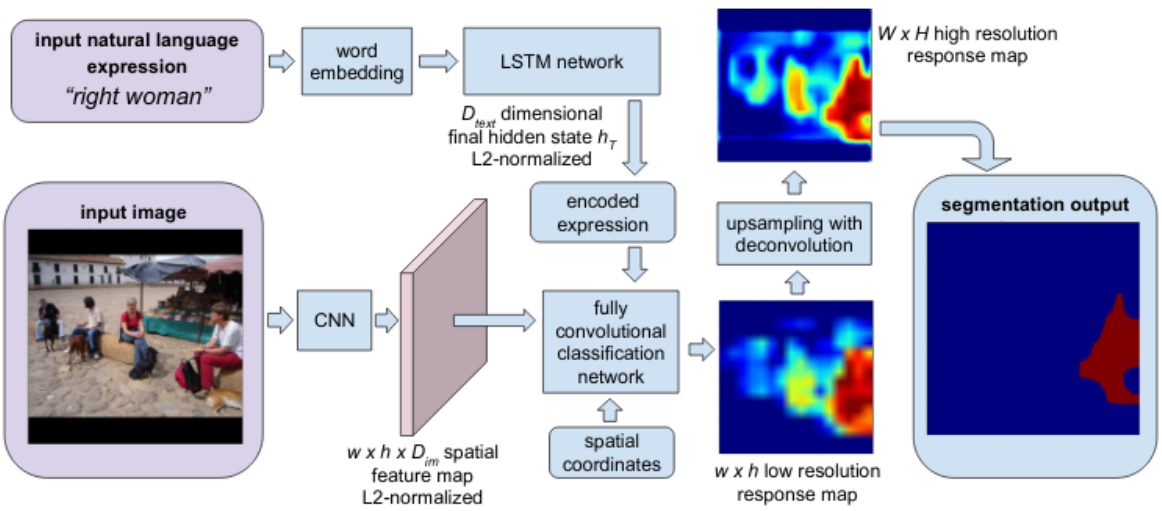

Figure 2.1: Model used by Hu *et al.* [Hu16a] who introduced the task of referring image segmentation.

$w = W/s$ and $h = H/s$ with $s = 32$. Also, two extra channels are added to each local descriptor, representing the relative coordinates of each pixel location from the upper left corner and the lower right corner of the feature map (represented as $(-1; -1)$ and $(+1; +1)$ respectively), so that the model can reason about spatial relationships found in expressions *e.g. ñcat on the leftò*.

After having extracted the visual and language features, h is first tiled and concatenated to the local descriptor at each spatial location of the spatial feature map to obtain a $w \quad h \quad D$ (where $D = D_{im} + D_{text} + 2$) multi-modal spatial map containing both visual and linguistic features. The combined features are then passed through the third component of the model which is a two-layer fully convolutional classifier consisting of two $1 \quad 1$ convolutions, that outputs a $w \quad h$ low resolution segmentation map of the image. In order to recover the original image size, an upsampling operation using deconvolution (or transpose convolution) [Zeil11] is performed producing a $W \quad H$ high resolution response map, whose values represent the confidence of whether a pixel belongs to the referred object. Since their work was the first to directly predict segmentation based on natural language expressions, the authors compare their model performance with strong baselines they created such as segmentation from bounding boxes or classification proposals and combination of per-word semantic segmentation, indicating that their method outperforms all these baselines.

Instead of modeling the image and langauge features independently and then combining them to produce the segmentation map, subsequent works tried to jointly model the two modalities, in order to better exploit the correlations between words and image regions. Liu *et al.* [Liu17] combine visual and word features with an LSTM to recurrently refine the segmentation masks. Dynamic filters were used in [Marg18] and [Chen19b] to capture the recursive nature of language and the spatial information of the target object respectively. Li *et al.* [Li18] presented a recurrent refinement network (RRN) which refines the segmentation result by utilizing the feature pyramid structures in order to take advantage of multi-scale semantics.

Posterior works in referring image segmentation leverage attention to model the visual information of each word. MAttNet [Yu18] decomposes referring expressions using three modules related to the object's appearance, location and relationships with other objects and then uses both language and visual attention to direct each module to focus on the desired part of the expression and the image. Shi *et al.* [Shi18] use attention to extract keywords from a referring expression which are important for identifying the target object. Cross-modal self-attention is used in CMSA [Ye19] to better capture the long-range dependencies between linguistic and visual features. While STEP [Chen19a] works in the same direction, it also uses a convRNN [Xing15] to refine the textual representation and improve the segmentation. A recent work by Hu *et al.* [Hu20] proposes a bi-directional cross-modal attention module to learn the relationship between multi-modal features. Finally, Huang *et al.* [Huan20] use multi-modal graph reasoning to identify the correct object as well as suppress other irrelevant ones.

2.1.2 Relevant datasets

RefCOCO

RefCOCO is a large-scale dataset and benchmark for referring image segmentation. It is collected on top of the Microsoft COCO (Common Objects in Context) image collection [Lin14], which includes images of complex everyday scenes containing common objects in their natural context.

It is one of the three most frequently used benchmarks for referring image segmentation: RefCOCO, RefCOCO+ and RefCOCOg [Yu16]. RefCOCOg was collected using Amazon Mechanical Turk in a non-interactive setup, while RefCOCO and RefCOCO+ were collected using the Refer-it Game [Kaze14]. In this two-player game, the first player is shown an image with a segmented target object and asked to write a natural language expression referring to the target object. The second player is shown only the image and the referring expression and asked to click on the corresponding object. If the target object is correctly identified, the players receive points and swap roles. Otherwise, they are presented with a new image and target object. Images in these collections were selected with the requirement to contain two or more objects of the same object category.

RefCOCO consists of 142,209 referring expressions for 50,000 objects in 19,994 images. The average number of words in its sentences is 3.61. Unlike RefCOCO+, where annotators are disallowed to use location words in their referring expressions, RefCOCO does not have any restrictions on its expressions. Moreover RefCOCO's referring expressions tend to be more concise than the ones from RefCOCOg which have an average length of 8.43. Another advantage of RefCOCO over RefCOCOg is that is contains more instances of same-category objects, having an average of 3.9 over 1.6 respectively.

Besides images, RefCOCO has been used for pre-training frame-based models on the task of referring video object segmentation like in the works of Khoreva *et al.* [Khor18] and Bellver *et al.* [Bell20], since a similar large-scale dataset for videos was not available. In the experiments of the present work, RefCOCO is also used along with the proposed synthetic dataset in order to assess how the synthetic data can contribute to a better pre-training of a deep neural network.

2.2 Referring Video Object Segmentation

2.2.1 Methods

Despite the increasing interest in referring image segmentation, only a few works have explored the segmentation of objects using referring expressions in the video domain *i.e.* referring video object segmentation. Khoreva *et al.* [Khor18] were the first to transfer the referring expression segmentation task from images to videos by collecting referring expressions for the DAVIS-2017 dataset [Pont17]. They use the image-based MAttNet [Yu18] model, pretrained on RefCOCO [Kaze14], to localize the target object, and then train a segmentation network with DAVIS-2017 to produce the pixel-wise prediction, using also temporal coherency across frames. Gavrilyuk *et al.* [Gavr18], in a relevant work, provide natural language sentences for Actor-Action Dataset (A2D) [Xu15] and J-HMDB [Jhua13] which are datasets used for action and human pose recognition and segmentation. They employ a 3D fully-convolutional model with dynamic filters in order to segment an actor in each frame of a video as specified by a language query. Although the task is similar to referring video object segmentation, the referring expressions they provide are intended to describe an actor and its action. The first large-scale dataset for referring video object segmentation, called Refer-YouTube-VOS, has been created concurrently to the present Master thesis by Seo *et al.* [Seo20] on top of YouTube-VOS [Xu18], a popular benchmark for video object segmentation. Besides the dataset, the authors propose a model called URVOS, which performs language-based object segmentation and mask propagation jointly using a single deep neural network. The network combines a cross-modal attention module, inspired by CMSA [Ye19] and a memory attention module to encourage temporal consistency across frames.

RefVOS

In another recent work, RefVOS [Bell20] has been the first model to leverage BERT [Devl19] for encoding the referring expressions. They have shown that using BERT instead of a bidirectional LSTM fed with GloVe embeddings [Penn14], which is a common practice in related works, brings significant improvements to the final segmentation. In the present Master thesis, RefVOS is the model used for the conducted experiments which aim at evaluating the proposed method and the generated synthetic dataset. A visual description of the architecture of RefVOS is depicted in Figure 2.3.

RefVOS is a frame-based model which uses DeepLabv3 [Chen17b] as its visual encoder. Convolutional Neural Networks deployed in fully convolutional fashion have shown to be effective for the task of semantic segmentation. However, the repeated combination of max-pooling and striding at consecutive layers of these networks significantly reduces the spatial resolution of the resulting feature maps. In order to recover the spatial resolution, deconvolutional (or transposed convolution) layers [Zeil11] have been employed in previous works using also skip connections to combine high resolution features from the contracting path to the upsampled output [Long15, Noh15, Ronn15].

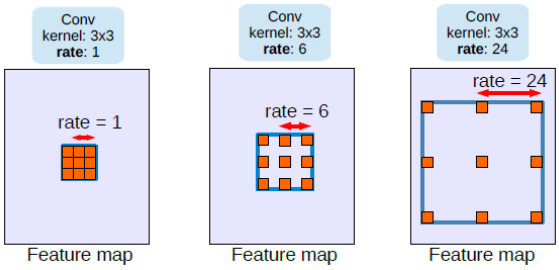

Figure 2.2: Illustration of atrous convolution with rates 1 (standard convolution), 6 and 24.

Instead of decreasing and then increasing the feature maps spatial resolution, DeepLabv3 uses "atrous convolution", which was originally developed for the efficient computation of the undecimated wavelet transform in the "algorithme à trous" scheme [Hols90] and then used in the convolutional neural networks context [Gius13, Serm13, Papa15]. Considering two-dimensional signals, for each location i on the output y and a filter w, atrous convolution is applied over the input feature map x as:

$$y[i] = \sum_{k} x[i + r \quad k]w[k] \tag{2.1}$$

where atrous rate r corresponds to the stride with which the input signal is sampled. This is equivalent to convolving the input x with upsampled filters produced by inserting $r \quad 1$ zeros between two consecutive filter values along each spatial dimension (hence the name "atrous convolution" where the French word "trous" means holes in English). Typical convolution is a special case of Equation 2.1 where $r = 1$. A visualization of atrous convolution with different atrous rates can be seen in Figure 2.2. By increasing the atrous rate r, one is able to use a wider field-of-view without the need to apply multiple convolutions or use larger kernels, *i.e.* without increasing the computational cost.

As seen in Figure 2.3 (top branch), DeepLabv3 applies four parallel atrous convolutions with different atrous rates, an architecture called Atrous Spatial Pyramid Pooling (ASPP) initially proposed in the first version of DeepLab [Chen17a], which is used in order to effectively capture multi-scale information. Besides the three $3 \quad 3$ atrous convolutions, a $1 \quad 1$ convolution and a global average pooling layer are involved. The features extracted from the five different operations are further processed in separate branches and fused to generate the final result. RefVOS model applies the ASPP architecture with atrous rates of 12, 24 and 36, as depicted in Figure 2.3.

The authors of DeepLabv3 also introduce the term of *output stride* to denote the ratio of input image resolution to the final feature map output resolution. Typical CNN architectures used for classification have an output stride of 32, meaning that the dimension of final feature responses, before fully connected layers, is 32 times smaller than the respective of the input image. Atrous convolutions, from the other side, allow to extract dense features without significantly decreasing the spatial resolution. RefVOS [Bell20] uses the architecture of DeepLabv3 with an output stride of eight.

Finally, in order to recover feature maps to the original image resolution for efficient segmentation, DeepLabv3 uses bilinear interpolation, which is sufficient in this setting because the feature maps produced with atrous convolutions are quite smooth. This way there is no need for extra deconvolutional (transpose convolution) layers which would increase the number of parameters and consequently memory requirements and total training time.

In order to obtain a linguistic embedding for the referring expression, RefVOS uses BERT, which stands for Bidirectional Encoder Representations from Transformers, and is a state-of-the-art language representation model presented by Devlin *et al.* [Devl19] (Google AI). Language model pre-training has been shown to be effective for improving performance in several natural language processing tasks. Before the publication of BERT, two typical strategies for applying pre-trained language representations to downstream tasks were used: (i) *feature-based*, such as ELMo [Pete18] and (ii) *fine-tuning*, such as the Generative Pre-trained Transformer (OpenAI GPT) [Radf18].

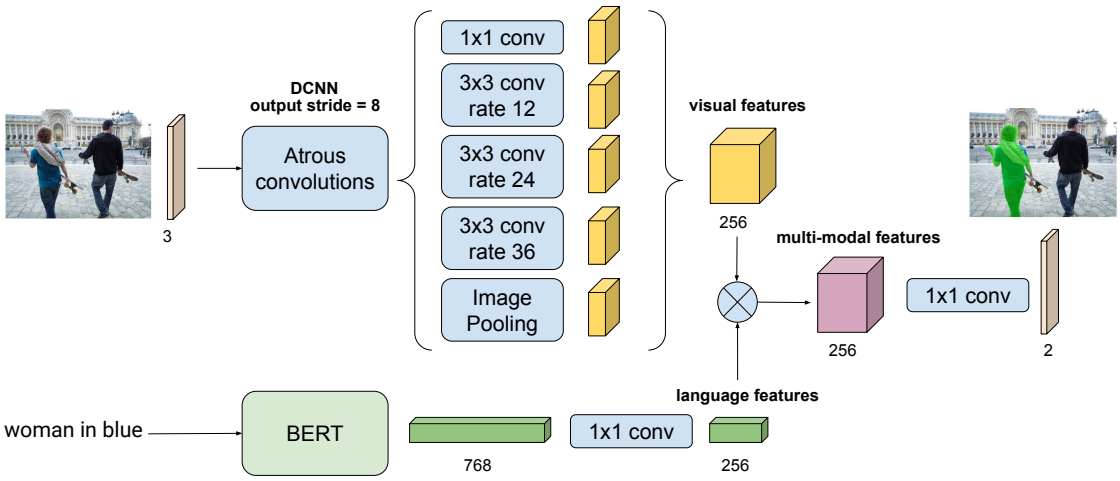

Figure 2.3: Architecture of the RefVOS model [Bell20].

The aforementioned approaches share the same objective function during pre-training, where they use unidirectional language models to learn general language representations. This means that they look at a text sequence either from left to right or combined left-to-right and right-to-left while training. In contrast, BERT is applying a bidirectional training and its performance in several downstream tasks shows that a language model which is bidirectionally trained can have a deeper sense of language context and flow than single-direction language model.

BERT makes use of Transformer [Vasw17], an attention mechanism that learns contextual relations between words (or sub-words) in a text. As opposed to directional models, which read the text input sequentially (left-to-right or right-to-left), the Transformer encoder reads the entire sequence of words at once, therefore it is considered bidirectional. This characteristic allows the model to learn the context of a word based on all of its surroundings (left and right of the word). In order to achieve that, the authors of BERT, use a "masked language model" (MLM) pre-training objective, inspired by the Cloze task [Tayl53]. The masked language model randomly masks some of the tokens from the input, and the objective is to predict the original vocabulary id of the masked word based only on its context. Unlike left-to-right language model pre-training, the MLM objective enables the representation to fuse both left and right context, which allows a bidirectional pre-training.

Sentences given as input to BERT are transformed to token sequences. The first token of every sequence is always a special classification token ([CLS]). The final hidden state corresponding to this token is used as the aggregate sequence representation for classification tasks. Since in the task of referring image/video object segmentation the whole sentence is important for the identification of the referred object, RefVOS uses the learned embedding corresponding to the [CLS] token as its linguistic representation, which is subsequently combined with the visual features.

The output of BERT is a 768-dimensional vector for each token of the sequence. To obtain a multi-modal embedding, the encoded linguistic phrase is first converted to a 256-dimensional embedding through a linear projection, in order to match the number of extracted visual features from DeepLabv3, *i.e.* the feature maps. Then, the linguistic embedding is element-wise multiplied with the visual features at every pixel position, producing a multi-modal embedding. Finally, a 1 1 convolutional layer predicts two maps, one for the foreground class, *i.e.* the referent, and another for the background.

2.2.2 Relevant datasets

DAVIS-2017

The first dataset combining video object segmentation and referring expressions was DAVIS-2017 [Pont17], where the name DAVIS stands for "Densely Annotated Video Segmentation". The first

DAVIS dataset was presented in 2016 by Perazzi *et al.* [Pera16], being the first benchmark specifically designed for the task of video object segmentation, including 50 videos with one pixel-wise annotated object in each. In 2017 a new version of DAVIS was released by Pont-Tuset *et al.* [Pont17], which, besides having a bigger number of sequences (100 additional videos), was modified to include multiple annotated objects in its videos, in contrast with the previous version. Of course, for the task of referring video object segmentation, which is tackled in this work, DAVIS-2017 is a more suitable benchmark, since disambiguation between different classes of objects is one of the main challenges of this task.

DAVIS-2017 was annotated with referring expressions by Khoreva *et al.* [Khor18], who were the first to propose the replacement of the first frame mask supervision with a referring expression for the task of video object segmentation. They collected two different types of annotations from two annotators: (i) first frame annotations which are the ones produced by only looking at the first frame of the video and (ii) full video annotations which are produced after seeing the whole video sequence. The annotation procedure involved a non-interactive referential two-player game setup. A first annotator was asked to provide a language description of the object, which has a mask annotation by looking either at the first frame or at the full video (according to the type of annotation previously described). Then another annotator is given the first frame or full video and the corresponding description, and is asked to identify the referred object. If the second annotator correctly identifies the target object the expression is accepted, otherwise, it is corrected to remove ambiguity and to specify the object uniquely.

The augmented with referring expressions DAVIS-2017 contains 1,544 referring expressions for 386 unique objects appearing in 150 videos. The average length for the first frame and full video expressions is 5.5 and 6.3 words respectively. Although the videos of DAVIS-2017 consist of a large number of annotated frames (69.7 on average) in comparison to other relevant datasets, its validation set is much smaller than the respective of other datasets including only 30 videos. For this reason the experiments of the present work include also an evaluation on both the training and validation sets of DAVIS-2017 (90 videos in total), for the models which are not fine-tuned on this dataset.

A2D Sentences

Another dataset used in language-guided video object segmentation is A2D Sentences, created by Gavrilyuk *et al.* [Gavr18]. This dataset is based on the Actor-Action Dataset (A2D) [Xu15], which is a benchmark for action understanding consisting of 3,782 videos from YouTube. It includes seven annotated actor classes considered to perform a set of eight possible actions. A2D Sentences is the augmented version of A2D with natural language descriptions, stating what each actor is doing in each video.

The creators of the dataset, following the guidelines of RefCOCO dataset [Kaze14], ask the annotators for a discriminative referring expression of each actor instance if multiple objects are present in a video. A2D Sentences is finally composed of 6,656 sentences for 3,782 videos and 4,825 objects. Its sentences contain on average more words than the extended with referring exprssions DAVIS-2017 [Khor18] (7.3 versus 5.9). Since it is a dataset targeted for action description, its sentences emphasize on verbs having a total of 225 different verbs.

Refer-YouTube-VOS

The last video dataset augmented with referring expressions is Refer-YouTube-VOS, created by Seo *et al.* [Seo20] who employed Amazon Mechanical Turk to collect referring expressions for YouTube-VOS [Xu18]. YouTube-VOS is the largest existing benchmark for video object segmentation, including more than four thousand high-resolution videos collected from YouTube with a small duration of three to six seconds each. It includes pixel-level mask annotations for 94 different object categories at every five frames, while its videos have a frame rate of 30 frames per second.

In order to collect crowd-sourced referring expressions, the authors of Refer-YouTube-VOS firstly selected around 50 annotators after performing a validation test. Each annotator was given a pair of

videos, the original video and the mask-overlaid one with the target object highlighted, and was asked to provide a discriminative sentence within 20 words that describes the target object accurately. Similar to Khoreva *et al.* [Khor18], two types of annotations were collected, one based on the first-frame and one on the full video. After the initial annotation, a verification and cleaning step was conducted for all annotations, and objects which could not be localized using just the produced language expression, were excluded from the dataset. In the end, Refer-YouTube-VOS consists of 27,899 expressions, referring to 7,451 objects in 3,975 videos, being the largest dataset with referring expressions in the video domain. Finally, Refer-YouTube-VOS has the largest average number of words per referring expression which is 7.5 for the first-frame annotations and 10.0 for the full-video ones.

Since YouTube-VOS, the basis of Refer-YouTube-VOS, and YouTube-VIS, the basis of the present work's proposed synthetic dataset, have a high overlap in their videos, the subset of Refer-YouTube-VOS that corresponds to YouTube-VIS has served as a benchmark for a direct comparison of the human-produced referring expressions with the respective synthetic ones proposed in the present Master thesis.

2.3 Object Detection

Object detection is the task of locating and classifying existing objects of a certain semantic class, as well as labeling them with rectangular bounding boxes which show the confidence of their existence. Being a classic computer vision problem, before the deep learning revolution in 2010s object detection has been approached with other machine learning-based methods. These methods first extract hand-crafted features like Haar [Viol01] or HOG [Dala05] features and SIFT keypoints [Lowe99] and then use machine learning techniques such as SVMs [Cort95] to do the classification.

However, after the recent advancements in deep learning, CNN-based methods have pushed the state-of-the-art in object detection as these techniques are able to detect objects in an end-to-end fashion without specifically defining features, outperforming classic computer vision methods in terms of detection accuracy. The frameworks of deep learning-based object detection methods can be mainly categorized into two types. The first one follows the traditional two-stage object detection pipeline, generating region proposals at first and then classifying each proposal into different object classes. The second considers object detection as a regression or classification problem, adopting a unified framework to acquire final object classes and locations in one step (single-stage detectors).

Two-Stage Detectors

Regarding two-stage detectors, the first stage is called a Region Proposal Network (RPN). A RPN takes an image (of any size) as input and outputs a set of rectangular object proposals, each with an objectness score, which measures the proposal's membership to a known set of object classes versus the background. Two-stage object detectors were introduced in the Selective Search work [Uijl13], while R-CNN [Girs14] was the first work to upgrade the second-stage classifier to a convolutional neural network yielding large gains in accuracy and introducing the deep learning era of object detection. The RPN of R-CNN extracts nearly 2000 region proposals, warps them into a square and feeds them to a convolutional neural network that produces a 4096-dimensional feature vector as output. Finally, a SVM takes as input these features acting as a classifier which decides on the presence of the object within that candidate region proposal. Besides predicting the presence of an object within the region proposal, the algorithm also predicts four values which are offset values to increase the precision of the bounding box. The pipeline of R-CNN is illustrated in Figure 2.4a.

Fast R-CNN [Girs15] instead of inputting 2000 region proposals to the CNN, uses the CNN to generate a feature map from the input image. From the convolutional feature map, the region proposals are identified and warped into squares. Then, a region of interest (RoI) pooling layer is used to reshape them into a fixed size so that they can be fed into a fully connected layer. The RoI pooling operation significantly speeds up the process as the same input feature map can be used for a big number of