Segmentation based on Natural Language referring expressions

Edgar A. Margffoy-Tuay
Universidad de los Andes
Carrera 1 No 18A - 12 Bogotá, Colombia
ea.margffoy10@uniandes.edu.co

Emilio Botero
Universidad de los Andes
Carrera 1 No 18A - 12 Bogotá, Colombia
e.botero10@uniandes.edu.co

Juan Camilo Pérez
Universidad de los Andes
Carrera 1 No 18A - 12 Bogotá, Colombia
jc.perez13@uniandes.edu.co

Abstract

The problem of segmenting a determined object (or objects) based on a natural language query, is tackled in this paper by means of Fully Convolutional Networks (FCN) merged with Networks based on Long Short-Term Memory (LSTM) units. The overall method is compared with State of the Art approaches in several datasets.

1. Introduction

Semantic segmentation in images is one of the fundamental tasks in Computer Vision. It consists of labeling all the pixels that belong to a certain object in an image, for all the possible predefined semantic classes. Advances in Deep Learning, together with the rise of large amounts of labeled data, have allowed for remarkable advances in this task [1, 2, 3], usually based on different architectures of Fully Convolutional Networks (FCN).

A problem that appeared recently [4] involves segmenting objects based on a natural language query regarding the image. The inputs are an image and an expression related to some elements in the image, and the goal is to segment the objects on the image specified in the sentence, that is, a segmentation mask.

In this task, in contrast to traditional segmentation, there are no predefined classes to be labeled, and therefore, the possible 'classes' to be labeled are just as *belonging to query* and *background*. Examples of the traditional segmentation problem and segmentation based on a Natural Language expressions can be seen in Figs. 1 and 2.

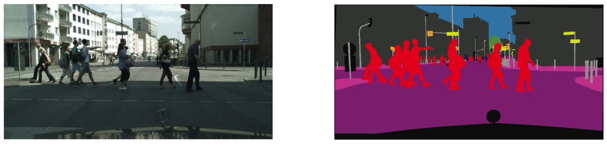
(a) Original image. (b) Segmentation

Figure 1: Example taken from [1] in the task of *traditional image segmentation*. Here, each color represents a different class from the set of all predefined classes in the dataset.

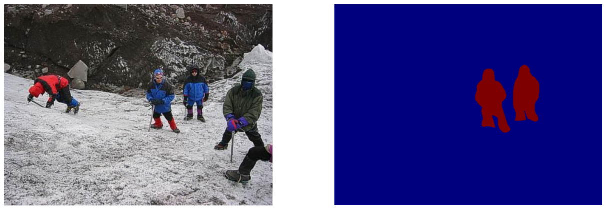
(a) Original image. (b) Segmentation based on query: *"people in blue coat"*

Figure 2: Example taken from [4] in the task of *segmentation based on natural language expressions*. Here, a single mask is the output, in which the only two labels represent *member of query* and *background*.

The facts that *(i)* this problem is much more unconstrained (label-wise), and *(ii)* there is an explicit combination that must be performed between the visual and language input, imply that the architecture of traditional networks may be useful but that the overall method must fundamentally change for it to be suitable for this task.

This task represents a very interesting challenge, since mechanisms for the combination of Computer Vision (CV) and Natural Language Processing (NLP) are yet to be explored and exploited. It is also a fundamental task for the development of algorithms and robots that can explore and process the majority of information available in our world, which involves visual and linguistic cues. Additionally, it is elementary for the interaction with humans, since it is a day-to-day task that we, as species, perform with ease and lets us interact with our surroundings.

2. Datasets

Four different datasets are used for training and evaluation of the proposed algorithm:

2.1. ReferIt

ReferIt [5] is a crowd-sourced database collected by S. Kazemzadeh *et al*. that contains images and referring expressions to objects in those images. Currently it has 130,525 expressions, referring to 96,654 distinct objects, in 19,894 photographs of natural scenes.

It has been used to perform training and evaluation in various works [4, 6], and has the characteristic of containing references to *'stuff'* like *sky*, *water* and *snow*.

2.2. RefCOCOg

The dataset was collected on Amazon Mechanical Turk and contains 85,474 referring expressions for 54,822 objects in 26,711 images. Images were selected to contain between 2 and 4 objects of the same object category [7].

2.3. RefCOCO

Collected interactively in the ReferIt game, with images that were selected to contain two or more objects of the same object category. It consists of 142,209 refer expressions for 50,000 objects in 19,994 images [7].

2.4. RefCOCO+

Similar to RefCOCO but a restriction regarding the use of location words in the image is applied, i.e. the referring expression must be based only on appearance rather than location, which depends on the perspective of the scene [7].

3. Related work

3.1. Segmentation from Natural Language Expressions

In [4], the general approach was to pass the original image through a FCN-32 to obtain a spatial map output as feature representation. Additionally, two channels, one for x coordinates and one for y coordinates, representing relative spatial coordinates , were concatenated to this output.

Meanwhile, the Natural Language expression was tokenized with a Word Embedding (WE) and each word was passed through a Long Short-Term Memory (LSTM) cell until the end of the sentence was reached. At that point, the last hidden state is taken and concatenated *at each spatial location* of the feature representation that was obtained from the FCN module.

Two convolution layers (with ReLU non-linearity between them) are then applied and, finally, a deconvolution layer for upsampling. The final segmentation mask is obtained by taking the places in which the value of the final response map are greater than 0.

The loss function is defined as the average over pixel-wise loss, and the whole network is trained with backpropagation.

3.2. Recurrent Multimodal Interaction for Referring Image Segmentation

The approach taken in [6] is to perform segmentation multiple times in the pipeline, so that the whole network does not only depend on the memorization (update of the hidden state) of the LSTM. This is to exploit the sequential property of natural language and make an analogy so that image segmentation is seen as a sequential process, too. The scheme is such that the *multimodal* information (language, image, spatial information, and their interaction) can be memorized with a multimodal-LSTM (mLSTM), where a mLSTM is a convolutional LSTM that shares weights both across spatial location and time step.

The initial processing for the image itself is similar to that in section 3.1: output of a FCN concatenated with maps representing relative spatial coordinates. After that, a word is represented by its WE and passed through a LSTM; the output is tiled with the input (the WE) and that resulting vector is concatenated with the output from the image-only initial processing at each spatial location. This tensor, that contains both visual and language information, is fed into the mLSTM. This process is recursively repeated for each word in the sentence, with the appropriate hidden states from LSTM and mLSTM being passed accordingly, until the end is reached.

Finally, a convolution is applied to the output of the last mLSTM to generate the final output with the proper dimensions. As a way of making the output less coarse, for evaluation, a DenseCRF is applied to the net's output for refinement.

References

[1] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[2] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *CoRR*, vol. abs/1411.4038, 2014.

[3] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *CoRR*, vol. abs/1412.7062, 2014.

[4] R. Hu, M. Rohrbach, and T. Darrell, “Segmentation from natural language expressions,” *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[5] S. Kazemzadeh, V. Ordonez, M. Matten, and T. L. Berg, “Referit game: Referring to objects in photographs of natural scenes,” in *EMNLP*, 2014.

[6] C. Liu, Z. Lin, X. Shen, J. Yang, X. Lu, and A. L. Yuille, “Recurrent multimodal interaction for referring image segmentation,” *CoRR*, vol. abs/1703.07939, 2017.

[7] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg, “Modeling context in referring expressions,” *CoRR*, vol. abs/1608.00272, 2016.