HCMUS at Eyes and Ears Together 2019: Entity Localization with Guided Word Embedding and Human Pose Estimation approach

Gia-Han Diep, Duc-Tuan Luu, Son-Thanh Tran-Nguyen, Minh-Triet Tran
Faculty of Information Technology and Software Engineering Laboratory
University of Science, VNU-HCM, Vietnam
{dghan,ldtuan,tnsthanh}@apcs.vn,tmtriet@fit.hcmus.edu.vn
{dghan,ldtuan,tnsthanh,tmtriet}@selab.hcmus.edu.vn

ABSTRACT

The Eyes and Ears Together Task focuses on developing an efficient framework for analyzing and localizing entities and associated pronouns from speech transcript. We present the HCMUS Team's approach, which employs a combination of Faster R - CNN model and Word2vec architecture. We submit multiple runs with different priority orders in our combine model. Our methods show potential results and achieve up to 2x accuracy in comparison to the task organizers' approaches.

1 INTRODUCTION

Eyes and Ears Together at MediaEval 2019 challenge[10] aims to automate data collection for visual grounding by exploiting speech transcription and videos and to develop larger-scale visual grounding systems. In this challenge, we are given the collection of instruction videos called "How2" [8] and their corresponding speech transcripts. In addition, task organizers also provide list of nouns, timestamps and 2048 dimensional feature vectors extracted from Residual Neural Network 152 layers [3] for every proposal in top 20 produced by Mask-RCNN[2]. The goal of the challenge is to localize specific entities given in time-align speech transcription in videos.

Since the given region proposals may not visually contain exactly the desired object, we decided to use Faster Region-based Convolutional Neural Network (Faster R-CNN) [7] with COCO [5] and OpenImage [4] pretrained weights to re-extract the proposals of the dataset. After that, Word2vec tool and OpenPose [1] are used in different priority order so as to select the best candidate proposal. We also examine using Tesseract Optical Character Recognition (OCR) [9] for detection.

2 APPROACH

Our goal is to localize the target objects given the object concept from the frame names. We come up with our proposed approach assuming that any model used to extract features from frames (querying the target objects) has already gained some knowledge to distinguish the target objects from others. Our first approach is to exploit knowledge from the concepts using word embedding and invade regional proposal with pretrained models (section 2.2). Then we can build a dictionary to translate the target concept to known concepts on the development set and apply it to the test set. Our second approach uses Pose Estimation to localize the concepts relating to human body (section 2.3). The last approach is to use OCR for detection (section 2.4). Overall, our method's original output is bounding boxes for all detected target objects.

MediaEval'19, 29-31 October 2019, Sophia Antipolis, France

2.1 Data preprocessing

We noticed that there are synonyms, singular and plural nouns targeting the same object in the dataset; Hence, to ensure the consistency, we decided to automatically convert given labels in the frames' names into real labels.

2.2 Word Embedding with regional proposal

Our method is to create a dictionary to map the keyword of a query with an existing concept in the two datasets MS Coco and OpenImageV4. To do this, we first use word embedding (section 2.2.1) to encode the keyword in each query and each known concept to measure their contextual relationship, with L2 is used to calculate the context distance. To build the dictionary, we manually create ground truth bounding boxes for approximately 50% of the development set - section 2.2.2, and for each keyword, we choose the top 3 related concepts (lowest L2 score), which have region proposal detected with IoU $\geq$ a chosen threshold (section 2.2.2).

For the test set, with only this dictionary and region proposals, we obtain a list of correspondent proposals and select the final bounding boxes based on probability, position in comparison to the central region or just random, depending on the query concepts. Figure 2 shows how this associate dictionary is built.

2.2.1 Word Embedding. Using Word2vec tool with Google News pretrained weight and Skip-gram architecture [6], we compute a 300-dimensional vectors representing the predicted labels (from OpenImage and Coco) and the target concepts. We then calculate cosine, L1 and L2 distance between each pair of target and predicted concepts to get best distance scores and choose the best metric.

Target concept	Top Predicted concepts
POLISH	Cosmetics
MOUTH	Human mouth
CALF/CALVES	Human leg, Human arm

Table 1: Example of the concept correspondent dictionary

2.2.2 Build concept dictionary from word embedding and regional proposal. We use pretrained Faster R-CNN for COCO and OpenImageV4 dataset to extract region proposals (predicting corresponding concepts - section 2.2.1) for every frames. For each image in the development set, we calculate the IoU between each outputted region proposal and each bounding box from the ground-truth, and filter out those proposals with IoU less than a chosen threshold. We then select the three-best-score predicted concepts (from the remaining proposals) for each target concept (using L2 metric) to obtain a dictionary translating the target concept into known concepts (example shown in table 1).

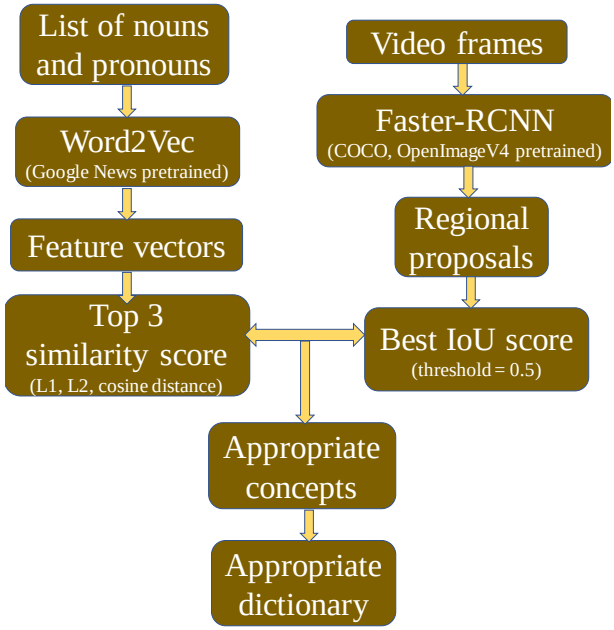

Figure 2: How associate dictionary is built using Guided Word Embedding Approach on development set

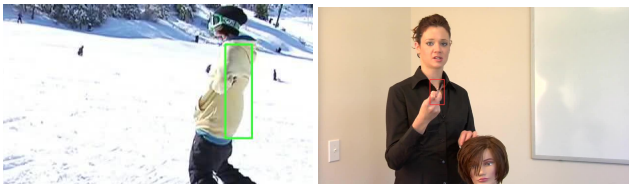

Figure 3: Example result from OpenPose approach

2.3 OpenPose Approach

We use OpenPose with concepts relating to human body, especially those neither in Coco or Openimage known concepts, such as: back, hips, toes, heels, etc. The pose estimation provide specific estimated position of human bone's keypoints. Based on those keypoints, we can expand the padding of the detected corresponding bone part to obtain the target object. We also apply human knowledge or heuristics for some concepts. For instance, we assume that human back is from hips to shoulders.

We also assume that objects hold by human hands should be somewhere either from the thumbs and index fingers to the upper hand-region or from one hand to another or from one thumb to the region expanding from that hand to the direction from elbow to hand. Figure 3 shows example result from our OpenPose approach.

2.4 OCR Approach

There are several concepts we decided to use OCR for detection. For these frames, we crop all regional proposals as input for an OCR model and take those with the output contains any words (with at least 90% letters) similar to words in the correspondent dictionary we manually built only for concepts using OCR. However, this approach is not so effective due to the image resolution.

3 EXPERIMENTS AND RESULTS

3.1 Run submissions

In Run 1, we use word embedding with regional proposal (section 2.2) with threshold 0 to get regional proposal with predicted concepts in the correspondent dictionary then use OpenPose approach (section 2.3) with only straight-forward concepts such as hands, eyes, etc. and take one-forth area in the center for the remaining, assumming target objects should be in the center of the image. We obtain Run 2 in the same way as Run 1, except for using threshold 0.5 for word embedding with regional proposal (section 2.2) approach. Run 3 and Run 4 are similar to the Run 1 and Run 2 but we take the bounding boxes with area equal nearly a half (instead of one-fourth) of the images area in the worst cases.

For Run 5, we also use word embedding with regional proposal approach (section 2.2) with threshold 0.5 and OpenPose approach with more hypothesis (section 2.3). We also use OCR approach (section 2.4) for this run for some concepts and lastly, one-fourth center region of the frame in worst cases.

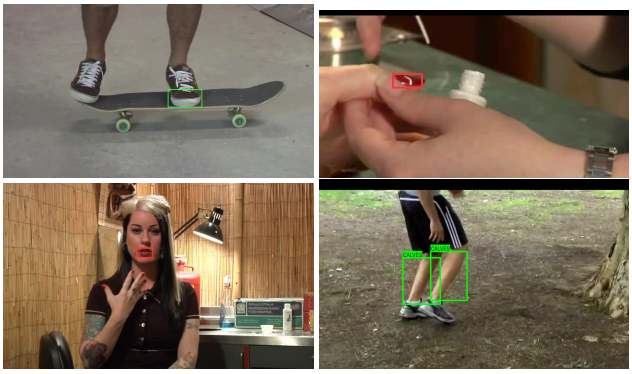

Figure 1: Example result of Word Embedding and Regional Proposal approach

3.2 Result

Table 2 shows that Run 2 is better than Run 1 and Run 4 is better than Run 3. Run 5 is our best model overall, which suggests that applying human knowledge for specific scenarios may get even better result.

Threshold	0.5	0.3	0.1
Run 1	0.208	0.35	0.545
Run 2	0.209	**0.354**	**0.551**
Run 3	0.213	0.346	0.54
Run 4	0.215	0.35	0.547
Run 5	**0.216**	0.348	0.542

Table 2: Eyes and Ears Together challenge 2019's result

4 CONCLUSION AND FUTURE WORKS

Eyes and Ears Together challenge is a novel problem trying to map knowledge gained from natural language to vision. Our current approach only focus on extracting proposals using Faster R-CNN, calculating distance between pair of target and known concepts, as well as using OpenPose keypoints as a guide for our hypothesis with human body parts. With the current approach, we gained a humble accuracy using only pretrained models, which can be increase with better detectors.

ACKNOWLEDGMENTS

Research is supported by Vingroup Innovation Foundation (VINIF) in project code VINIF.2019.DA19. We would like to thank AIOZ Pte Ltd for supporting our team with computing infrastructure.

REFERENCES

[1] Zhe Cao, Gines Hidalgo, Tomá imon, Shih-En Wei, and Yaser Sheikh. 2016. Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), 1302–1310.

[2] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. 2017. Mask R-CNN. *2017 IEEE International Conference on Computer Vision (ICCV)* (2017), 2980–2988.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Jun 2016). https://doi.org/10.1109/cvpr.2016.90

[4] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper R. R. Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, and Vittorio Ferrari. 2018. The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale. *ArXiv* abs/1811.00982 (2018).

[5] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *ECCV*.

[6] Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *CoRR* abs/1301.3781 (2013).

[7] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (Eds.). Curran Associates, Inc., 91–99.

[8] Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: A Large-scale Dataset for Multimodal Language Understanding. *CoRR* abs/1811.00347 (2018). arXiv:1811.00347 http://arxiv.org/abs/1811.00347

[9] Ray Smith and Google Inc. 2007. An overview of the Tesseract OCR Engine. In *Proc. 9th IEEE Intl. Conf. on Document Analysis and Recognition (ICDAR*. 629–633.

[10] Florian Metze Yasufumi Moriya, Ramon Sanabria and Gareth J. F. Jones. 2019. Eyes and Ears Together Task at MediaEval 2019. *Media Eval' 2019* (2019).