MediaEval 2019: Eyes and Ears Together

Yasufumi Moriya[1], Ramon Sanabria [2], Florian Metze [2], Gareth J. F. Jones[1]
[1]Dublin City University, Dublin, Ireland
[2]Carnegie Mellon University, Pittsburgh, PA, USA
{yasufumi.moriya,gareth.jones}@adaptcentre.ie,{ramons,fmetze}@cs.cmu.edu

ABSTRACT

We describe the Eyes and Ears Together task at MediaEval 2019. This task aims to ground entities found in speech transcripts in corresponding videos. Participants are asked to develop a system that draws a bounding box around an object in a video frame for a given a query entity for a collection of instruction videos with speech transcripts. Participants must automatically label video frames with an entity. For evaluation, the dataset is manually annotated with ground truth bounding boxes of query objects.

1 INTRODUCTION

Grounding the use of natural language into physical activities and entities in the social world is a crucial human ability. Humans can associate a linguistic entity with its abstract concept and with its visual object. For example, an entity banana can be connected to a picture of a yellow fruit, a banana boat or chopped pieces of a banana fruit. Such a grounding ability can be applied to visual-question answering [3], image retrieval [8] and robotics [5].

The Eyes and Ears Together task at MediaEval 2019 focuses on visually grounding speech transcripts into videos. Although there has been previous work on visually grounding captions in images or videos [2, 6, 8, 11, 14], they do not address grounding speech in videos. The primary difference between caption grounding and speech grounding is that captions must be created through manual annotation of videos or images, whereas speech transcripts can be substituted for automatically generated ones using automatic speech recognition (ASR). This motivates us to examine the exploitation of a large archive of spoken multimedia data without manual annotation. Grounding speech into vision is also more difficult than grounding captions in that entities uttered in speech are not necessarily visible in a visual stream or entities (*e.g.,* banana) in conversational speech can refer to several objects (*e.g.,* yellow fruit, boat, chopped fruit). This differentiates our task from others using speech such as [4, 7], where associating speech segments with vision is performed on spoken captions of images. Participants in the Eyes and Ears task are asked to develop a visual grounding system on a collection of approximately 300 hours of instruction videos, How2 [13]. Using speech transcripts and videos of How2, we have automatically generated pairs of entity and video frames. The remainder of this paper describes our data collection and annotation process for construction of the evaluation set, the visual grounding task and the evaluation metrics.

Copyright 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
MediaEval'19, 27-29 October 2018, Sophia Antipolis, France

2 DATA DESCRIPTION

The How2 corpus is a collection of instruction videos developed for multimodal tasks such as multimodal automatic speech recognition, machine translation and summarisation [13]. The corpus consists of approximately 300 hours of instruction videos accompanied by their speech transcripts and crowd-sourced Portuguese translations. The corpus is partitioned into training, dev5 and val set. For the Eyes and Ears Together task, the combined dev5 and val set is referred to as the evaluation set.

2.1 Data Collection

Development of a visual grounding model requires pairs of entity and video frames. Approaches to visual grounding are often performed in a weakly-supervised manner [6, 11, 14], as annotating every video frame with a bounding box of a target entity is expensive and time-consuming. For development of the Eyes and Ears task, we extracted pairs of entity and video frames using the following steps:

- Time-align speech transcripts with audio files using an automatic speech recognition system developed on the training set of How2
- Apply the Stanford Core NLP tool to time-aligned speech transcripts to obtain part-of-speech tags of words [9]
- Retain nouns and noun phrases that are part of ImageNet labels, and assume that the object is visible in the example image [12]
- Extract video frames at the end timestamp of the nouns extracted in the previous step

The intuition of this algorithm is that when an entity is uttered in speech, it is likely to be seen in a visual stream. This approach produced 139,867 pairs of entity and video frame from the training set, and 5,267 pairs from the evaluation set. The unique number of labels was 533, which is reduced to 445 when singular and plural labels are conflated.

2.2 Data Annotation

After extracting pairs of entity and video frame, we annotated the evaluation set with a bounding box for each target entity. Figure 1 shows our annotation platform using Amazon Mechanical Turk (AMT). AMT Workers were asked to draw a bounding box for each target item shown under "Labels". When there is no item visible in an image, AMT Workers can select "Nothing to Label" and submit it instead of drawing a bounding box. The authors reviewed annotated items, and out of 5,267 video frames, 2,444 video frames were kept as valid annotation, 2,331 were submitted with "Nothing to Label" and the remaining submissions discarded as invalid annotations. This demonstrates that roughly 51% of video frames extracted using the

CEUR-WS.org/Vol-2670/MediaEval_19_paper_2.pdf

procedures in Section 2.1 contain the target entity in the evaluation set.

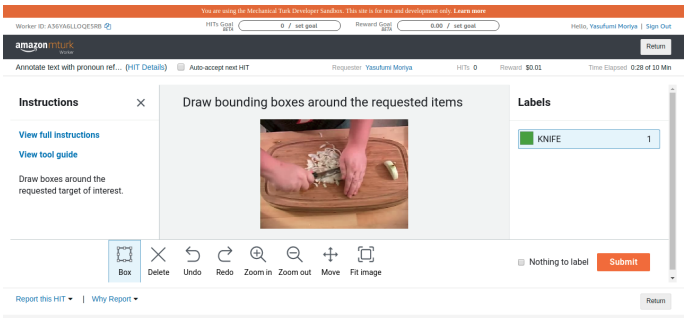

Figure 1: Our annotation platform using Amazon Mechanical Turk for the evaluation set. Workers can draw a bounding box for a target item using the interface.

3 TASK DESCRIPTION

The goal of the Eyes and Ears Together task is to identify an object in a video given a query entity. The input to the visual grounding system is a video frame and its target entity. The system is expected to produce a bounding box for the target entity contained in the the video frame selected as associated with an utterance of the target entity. Figure 2 shows an example video frame in which a bounding box capturing an onion is drawn.

Figure 2: A simple example of how a visual grounding model identifies a target entity in a video frame. A blue bounding box that captures an onion is drawn.

Approaches to weakly-supervised visual grounding can be classified into three. The most common approach is to generate multiple object region proposals (bounding boxes) using an algorithm such as a region proposal network (RPN) or selective search, and to train a visual grounding model to find a weak connection between proposed object regions and a target entity. While [6, 14, 17] base their system on multiple instance learning (MIL) using a ranking loss function, [11] aims to reconstruct a textual target entity from object region proposals to which attention weights are applied. In [10], visual grounding systems using MIL and reconstruction are developed on the How2 dataset. These systems are the baseline systems for this task. The drawback of the approaches using object region proposals is, however, that an upper-bound value for visual grounding is limited to the quality of proposed regions. In other words, when there is no overlap between any of the region proposals and a gold standard bounding box, it is impossible for a visual grounding system to draw a correct bounding box. The second type of approaches do not rely on object region proposals [15, 16]. These approaches produce a salient map of an input image given a target entity, and apply sub-window search to the map in order to discover a bounding box or segmentation of a target object. Finally, the third approach to the visual grounding system was developed on the How2 corpus [1]. This focuses on removal of video frames that are false positives of the training set (*i.e.,* removing video frames that actually do not show a target entity). However, this work is limited to grounding 11 entities, whereas the Eyes and Ears Together asks participants to ground 445 unique entities.

We provide participants with word embedding features and visual features in the form a vector representing the object of interest for each proposed bounding box. However, since the use of object region proposals limits an upper-bound score that a system can obtain, we encourage participants to explore alternative methods for the task to remove dependency on region proposals.

4 EVALUATION

Submitted visual grounding systems are evaluated through accuracy of intersection over union (IoU) of a predicted bounding box and a gold standard bounding box. IoU is a common metric to evaluate visual grounding systems [6, 14, 17]. For each video frame of the test corpus, a visual grounding model produces a bounding box of a target entity. The prediction is checked against gold standard bounding boxes. When an IoU value is higher than a threshold value with any of the gold standard bounding boxes of the video frame, this prediction is considered as positive. The final score is IoU accuracy, where the total number of positive predictions is divided by the total number of video frames in the test set.

5 RUN DESCRIPTION

Every team is allowed to submit up to 5 system runs. We will report IoU accuracy with a threshold 0.5, 0.3 and 0.1. The final score with the threshold 0.5 will be used to rank the submitted systems.

6 CONCLUSION

This overview paper describes the Eyes and Ears Together task at the MediaEval 2019 benchmark. This is the first step towards large-scale visual grounding for speech transcripts. Unlike caption grounding into vision, in speech grounding, a target entity cannot be shown or an uttered entity can refer to different objects depending on the context, as naturally happens in conversational speech. Since the dataset was constructed through the automatic approach described in the paper, it can contain false positive labels. Three different types of approaches to weakly-supervised visual grounding were reviewed in Section 3. For evaluation, we will employ IoU of predicted bounding boxes and gold standard.

7 ACKNOWLEDGEMENT

This work was partially supported by Science Foundation Ireland as part of the ADAPT Centre (Grant 13/RC/2106) at Dublin City University.

REFERENCES

[1] Elad Amrani, Rami Ben-Ari, Tal Hakim, and Alex Bronstein. 2019. Toward self-supervised object detection in unlabeled videos. *arXiv preprint arXiv:1905.11137* (2019).

[2] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. 2018. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *arXiv preprint arXiv:1804.03619* (2018).

[3] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 6904–6913.

[4] David Harwath, Galen Chuang, and James Glass. 2018. Vision as an interlingua: Learning multilingual semantic embeddings of untranscribed speech. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4969–4973.

[5] Jun Hatori, Yuta Kikuchi, Sosuke Kobayashi, Kuniyuki Takahashi, Yuta Tsuboi, Yuya Unno, Wilson Ko, and Jethro Tan. 2018. Interactively Picking Real-World Objects with Unconstrained Spoken Language Instructions. In *International Conference on Robotics and Automation*.

[6] D. A. Huang, S. Buch, L. Dery, A. Garg, L. Fei-Fei, and J. C. Niebles. 2018. Finding "It": Weakly-Supervised, Reference-Aware Visual Grounding in Instructional Videos. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*. 5948–5957.

[7] H. Kamper, S. Settle, G. Shakhnarovich, and K. Livescu. 2017. Visually Grounded Learning of Keyword Prediction from Untranscribed Speech. In *Interspeech*. 3677–3681.

[8] A. Karpathy and L. Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Computer Vision and Pattern Recognition (CVPR)*. 3128–3137.

[9] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*. 55–60.

[10] Y. Moriya, R. Sanabria, F. Metze, and G. J. F. Jones. 2019. Grounding Object Detections With Transcriptions. In *Workshop on New Tasks for Vison and Language*.

[11] Anna Rohrback, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. 2016. Grounding of Textual Phrases in Images by Reconstruction. In *European Conference on Computer Vision (ECCV)*. 817–834.

[12] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* 115, 3 (2015), 211–252.

[13] R. Sanabria, O. Caglayan, S. Palaskar, D. Elliott, L. Barrault, L. Specia, and F. Metze. 2018. How2: A Large-scale Dataset For Multimodal Language Understanding. In *Workshop on Visually Grounded Interaction and Language (ViGIL)*. NeurIPS.

[14] Jing Shi, Jia Xu, Boqing Gong, and Chenliang Xu. 2019. Not All Frames are Equal: Weakly-supervised video grounding with contextual similarity and visual clustering losses. In *Computer Vision and Pattern Recognition (CVPR)*.

[15] Fanyi Xiao, Leonid Sigal, and Yong Jae Lee. 2017. Weakly-Supervised Visual Grounding of Phrases With Linguistic Structures. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[16] Raymond A. Yeh, Minh N. Do, and Alexander G. Schwing. 2018. Unsupervised Textual Grounding: Linking Words to Image Concepts. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[17] Luowei Zhou, Nathan Louis, and Jason J Corso. 2018. Weakly-Supervised Video Object Grounding from Text by Loss Weighting and Object Interaction. In *British Machine Vision Conference*.