DCU-ADAPT at MediaEval 2019: Eyes and Ears Together

Yasufumi Moriya, Gareth J. F. Jones
ADAPT Centre, School of Computing, Dublin City University, Dublin 9, Ireland
yasufumi.moriya@adaptcentre,gareth.jones@dcu.ie

ABSTRACT

We describe the DCU-ADAPT participation in the Eyes and Ears Together task at MediaEval 2019. Our submitted systems were developed to choose object bounding boxes from automatically generated proposals given query entities. The first system finds relevance between object proposals and queries using multiple instance learning. The second system employs an attention mechanism to find object proposals which are most likely correspond to the given queries. The last system is a baseline system which chooses region proposals at random. We observed that the first two systems produced higher accuracy than the random baseline. The best approach was to use multiple instance learning which resulted in accuracy of 9% when the threshold of intersection over union was 0.5.

1 INTRODUCTION

The nature of human communication is often a multimodal process, where textual, visual and audio information are simultaneously processed. The Eyes and Ears Together task at MediaEval 2019 aims to ground speech transcripts into videos [7]. Visual grounding tasks are conducted on images or videos and manually created captions [4, 5, 8], but rarely on vision and speech. Speech grounding is interesting, in that this replicates human communication, where listening to speech and seeing objects happen simultaneously. A practical advantage of grounding speech into vision is that, unlike caption grounding, speech transcripts can be obtained easily from user generated content (e.g., YouTube) or using automatic speech recognition.

As a task organiser, we generated pairs of video frames and entities from the How2 dataset [7, 9]. The challenge of this task is that systems need to discover relationships between objects and entities without explicit annotation of objects, since pairs of video frames and entities are automatically aligned.

In this paper, we describe our investigation into whether two existing approaches employed for caption grounding could be applied to speech grounding. The common characteristics of these approaches are that they both use pre-computed candidate region proposals of objects. The first approach is to find relationships between object proposals and queries using contrastive loss [4]. This employs an established approach referred to as *multiple instance learning (MIL)* which is often applied to other computer vision tasks [3]. The second approach is to use the attention mechanism [1], with an object bounding box which has the highest attention weight taken as a prediction given a query entity [8]. To compare these approaches to the most basic system, the final system randomly chooses object bounding boxes from candidate region proposals.

MediaEval'19, 27-29 October 2019, Sophia Antipolis, France

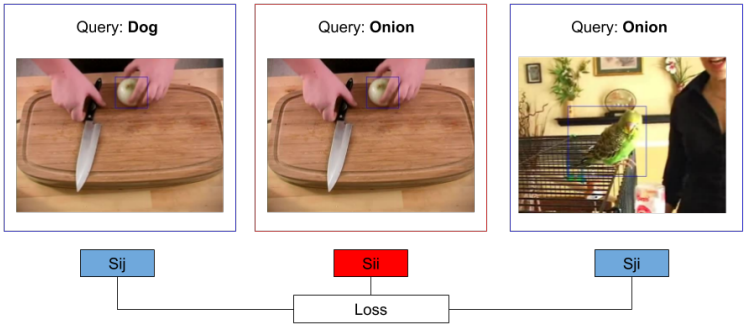

Figure 1: Computation of loss function using contrastive loss.

2 OUR APPROACH

We use machine learning approaches to visual grounding using automatically generated object proposals. For each video frame, there are n object proposals. We extract n fixed-length feature vectors by cropping a video according to object proposals and applying a convolutional neural network (CNN) to each cropped image. Each query entity associated with a video frame is also transformed into a fixed-length vector using a word embedding model.

2.1 Multiple Instance Learning

Given region proposals transformed into fixed-length vectors, and a query entity also represented as a vector, a neural network model can find the region proposal which is the most strongly associated with the query entity [4]. This can be expressed in the following equations.

$$\phi(r_{ijk}) = W_r(f_{CNN}(r_{ijk})) \quad (1)$$

$$\psi(e_i) = W_e(f_{EMB}(e_i)) \quad (2)$$

$$\bar{k} = \arg\max_k(sigmoid(\phi(r_{ijk})^T \cdot \psi(e_i))) \quad (3)$$

where i denotes the ith entity, j - jth the video frame of an entity and k - kth a region proposal, $\phi(r_{ijk})$ is a CNN feature of r_{ijk}, $\psi(w_i)$ is a word embedding of query entity e_i, and $\bar{k}$ is an index of the region proposal which is the most strongly associated with e_i. While f_{CNN} and f_{EMB} are fixed during training, in the neural network model W_r and W_e are updated at training time.

At training time, given region proposals and a query entity, a neural network model is trained to find relationships between video frames and query entities, as shown in Figure 1. For each pair of a video frame and a query entity, there are two additional pairs which create a mis-match between a video frame and a query entity. The loss function penalises a model when it gives a higher score to

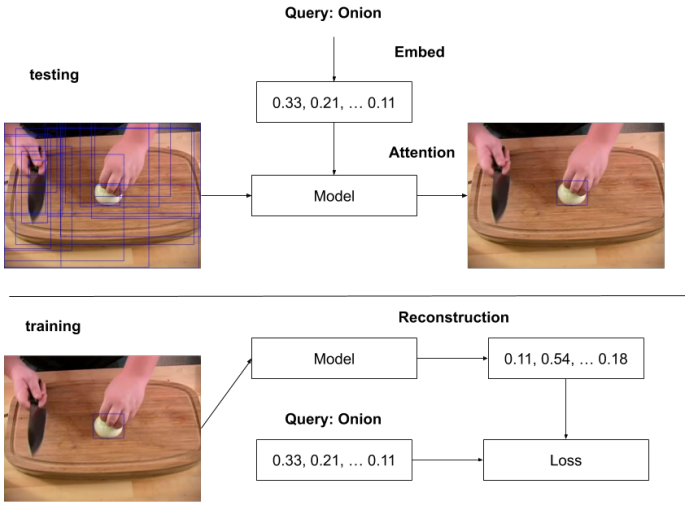

Figure 2: Computation of loss function using reconstruction.

a mis-matched pair. This is expressed in Equation 5.

$$S_{ii} = \sum_{j} \max_{k}(\phi(r_{ijk})^T \cdot \psi(e_i)) \quad (4)$$

$$L = \sum_{i}^{I}(\max(0, S_{il} - S_{ii} + \delta) + \max(0, S_{li} - S_{ii} + \delta)) \quad (5)$$

where S_{ii} is a correctly matched image-entity pair, S_{il} is the current image and a random query entity and S_{li} a random image and the current query entity.

2.2 Reconstruction

A neural network can find a region proposal that is the most strongly associated with a query entity using attention mechanism [1].

$$\bar{k} = \arg\max_{k}(f_{ATTN}([\phi(r_{ijk}); \psi(e_i)])) \quad (6)$$

This is applied in Equation 6, where f_{ATTN} is an attention function which computes attention weights over k region proposals given concatenation of visual features $\phi(r_{ijk})$ and an embedded query entity $\psi(e_i)$.

At training time, a model can learn a relationship between a visual object and a query entity by reconstructing an embedded query entity from a region proposal which has the highest attention weight [8]. Figure 2 shows how an object bounding box is found at testing time, and how a model is trained to reconstruct a query entity from a region proposal at training time. Formally, the following equations express how to compute a reconstruction loss.

$$r_{attn} = W_{rec} \sum_{k=1}^{N} a_k \phi(r_{ijk}) \quad (7)$$

$$L_{rec} = \frac{1}{D} \sum_{d=1}^{D} (\psi(e_i)^d - r_{attn}^d) \quad (8)$$

Table 1: Results of visual grounding of accuracy at three thresholds 0.1, 0.3 and 0.5.

	0.5	0.3	0.1
MIL	0.094	0.227	0.494
Rec	0.080	0.192	0.402
Random	0.077	0.181	0.408

In Equation 7, the sum of a visual feature from region proposals multiplied by attention weights a_k is transformed into a reconstructed embedding of a query entity r_{attn}. In Equation 8, L_{rec} is essentially a mean squared error of a reconstructed query entity and an embedded query entity.

3 IMPLEMENTATION DETAILS

For each video frame, 20 region proposals were extracted from the How2 dataset [9] using Mask-RCNN [6]. The Mask-RCNN uses ResNeXt101 [10] as its backbone. For each region proposal, the ResNet 152 model was used to extract fixed-length vectors. The dimension of each visual feature was 2,048. The word embedding model was trained on the training set of the How2 speech transcripts using the fastText library [2], and each query entity was embedded into a 100 dimensional vector.

4 RESULTS

Table 1 shows results of visual grounding using the MIL-based approach, the reconstruction based approach and the system which chooses region proposals at random. The systems were evaluated in terms of intersection of a selected region proposal and a gold standard bounding box divided by union of a region proposal and a gold standard (IoU). When an IoU value exceeded thresholds of 0.5, 0.3 or 0.1, a system prediction was regarded as correct. As can be seen in the table, both MIL and reconstruction approaches generally produced slightly better results than a simple random approach. A possible explanation for poor results of the two models is that those approaches have been applied to caption grounding and showed reasonable results, but have not been applied to speech grounding. For speech grounding, it is possible that entities are sometimes weakly associated with visual objects. Therefore, existing models may need modification for speech grounding to efficiently learn relationships between entities and objects.

5 CONCLUSIONS

This paper describes the DCU-ADAPT participation in Eyes and Ears Together at MediaEval 2019. We employed machine learning approaches previously applied to caption grounding, and investigated whether those models can work on speech grounding as well. It was found that whlie they still perform better than the random baseline, they require modification to better capture weak relationships between entities in speech transcripts and visual objects.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *3rd International Conference on Learning Representations (ICLR)*.

[2] Piotr. Bojanowski, Edouard. Grave, Armand. Joulin, and Ttomas. Mikolov. 2016. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5 (2016), 135–146.

[3] Marc-André Carbonneau, Veronika Cheplygina, Eric Granger, and Ghyslain Gagnon. 2018. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition* 77 (2018), 329 – 353. https://doi.org/10.1016/j.patcog.2017.10.009

[4] De-An Huang, Shyamal Buch, Lucio Dery, Animiesh Garg, Li Fei-Fei, and Juan Carlos 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.

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

[6] Francisco Massa and Ross Girshick. 2018. maskrcnn-benchmark: Fast, modular reference implementation of Instance Segmentation and Object Detection algorithms in PyTorch. https://github.com/facebookresearch/maskrcnn-benchmark. (2018). Accessed: 07 June 2019.

[7] Yasufumi Moriya, Ramon Sanabria, Florian Metze, and Gareth Jones J. F. MediaEval 2019: Eyes and Ears Together. In *Proceedings of MediaEval 2019*.

[8] Anna Rohrbach, 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.

[9] Ranom Sanabria, Ozan Caglayan, Shurti Palaskar, Desmond Elliott, Loic Barrault, Lucia Specia, and Florian Metze. 2018. How2: A Large-scale Dataset For Multimodal Language Understanding. In *Workshop on Visually Grounded Interaction and Language (ViGIL)*. NeurIPS. http://arxiv.org/abs/1811.00347

[10] Saining Xie, Ross Girshick, Piotr Dollar, Zhuowen Tu, and Kaiming He. 2017. Aggregated Residual Transformations for Deep Neural Networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.