A Multi-scale Multiple Instance Video Description Network

Huijuan Xu[1], Subhashini Venugopalan[2], Vasili Ramanishka[1], Marcus Rohrbach[3], Kate Saenko[1]
[1]UMass Lowell, [2]UT Austin, [3]UC Berkeley
[1]{hxu1, vramanis, saenko}@cs.uml.edu, [2]vsub@cs.utexas.edu, [3]rohrbach@eecs.berkeley.edu

Abstract

Generating natural language descriptions for in-the-wild videos is a challenging task. Most state-of-the-art methods for this problem borrow existing deep convolutional neural network (CNN) architectures to extract a visual representation of the input video, which cannot handle multiple receptive filed sizes in original image. In order to consider objects in different positions and at different scales simultaneously, we apply several fully convolutional neural networks (FCNs) to form a multi-scale network. Evaluation on a Youtube video dataset shows the advantage of our approach compared to the original single-scale model.

1. Introduction

The ability to automatically describe videos in natural language has many real-world applications, such as content-based video retrieval, descriptive video service for the visually impaired, and automated video surveillance. Most current visual description approaches make use of pre-trained deep convolutional neural networks (CNNs) as semantic feature extractors for each video frame. These CNN models are trained to predict a single object label on images where objects are usually center positioned and occupy most of the image. However, realistic videos are much more complex and contain several objects of different scales in different positions of each video frame, including small objects. To detect smaller objects and actions, receptive fields of different sizes (relative to the original image size) must be used.

In this paper, we propose the first end-to-end trainable video description network that incorporates spatially localized descriptors to capture concepts at multiple scales. We combine the traditional classification CNN that operates on the scale of the whole image with fully-convolutional network (FCN) of smaller receptive fields in the original image. We further incorporate a Multiple Instance Learning (MIL) mechanism to deal with the uncertainty of object scales and positions. The resulting semantic representation of the frames is encoded into a hidden state vector and then decoded into a sentence using a recurrent neural network proposed in [7]. We call our model the **Multi-scale Multi-instance Video Description Network (MM-VDN)**.

2. Approach

The MM-VDN architecture is shown in Figure 1. The input to the network is a sequence of video frames, and the output is a sequence of words. Each frame is processed by a multi-scale multi-instance convolutional network and embedded into a $N-$dimensional high-level semantic vector, corresponding to activations of N high-level concepts. A recurrent network accepts such semantic vectors from all frames in the video, and then decodes the resulting state into the output sentence. Unlike previous approaches that used a single-scale single-label architecture (top stream in our network), our network can handle ambiguity in the number, size and location of objects and activities in the scene.

The visual subnet (Figure 1) is structured as several multi-scale fully convolutional networks connected via the MIL mechanism. The first scale is the base pre-trained CNN classification network (AlexNet [4]) applied on the whole frame to capture scene-level semantics. Additional scales consist of the same CNN network but applied in a fully convolutional manner across upsampled versions of the original frame.

Our FCN conversion of AlexNet changes the last three fully-connected layers into convolutional layers, while the first five convolutional layers are kept the same. The weights in the last three fully-connected layers of AlexNet are converted to be the filter weights in the last three convolutional layers of the FCN. The weights of the first seven layers are shared across scales (shown by shading their outputs in Figure 1). The fc8 weights are not shared, to allow different concepts to be learned at different scales.

The MIL mechanism consists of several layers of max-pooling and allows the latent position and scale of concepts to be discovered simultaneously during learning. MIL allows weaker forms of supervision where training examples come in *sets*. We consider a set to be all image patches corresponding to all possible receptive field locations at all scales in a given frame. A MIL max pooling layer is defined on top of each FCN output score map to capture the maximum score, which infers the latent object positions. Then, we define an additional MIL element-wise max layer on top of the multi-scale FCNs to select among different

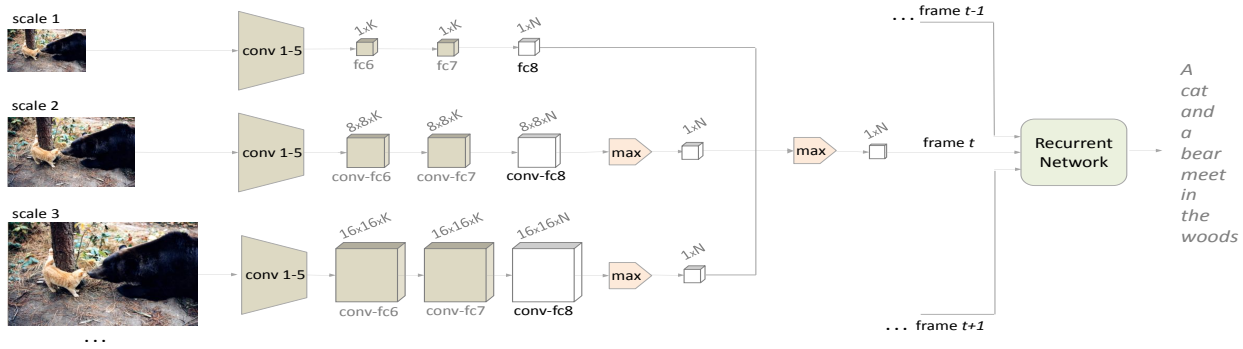

Figure 1. MM_VDN architecture

Input Image Size	Score Map Size	Height Ratio
227×227	1×1	100%
451×451	8×8	78.7%
707×707	16×16	50.2%

Table 1. Height ratio of the receptive field to the original input image for different input sizes in the FCN. The size of the receptive field is 355×355 for all input image sizes.

input scales for each embedding concept.

When the input image of the FCN has been upscaled, each score in the output score map corresponds to a smaller region in the original unscaled image. Thus, by using a FCN coupled with an upscaled input image, we can capture smaller objects. We further combine several FCNs with different input image sizes (scales), and apply MIL across the scales to capture concepts of different scales simultaneously. The ratio of the receptive field height to the original image height for several different input scales is shown in Table 1.

3. Experiments

We perform our experiments on the Microsoft Research Video Description Corpus (MSVD) [2]. The dataset contains 1,970 short Youtube video clips paired with multiple human-generated natural language descriptions ($\sim$ 40 English sentence descriptions per video). The 1,970 videos are split into training set (1,200 videos), validation set (100 videos) and testing set (670 videos), as used by the prior work on the same video description task [3][6][9][8]. We report the test accuracy after choosing the model on validation set.

In our model, the input image for the AlexNet part is resized to 256×256 and cropped to 227×227 to generate five candidate patches (four corners and one center) without mirroring. The input size for other scales is directly set to be the input size listed in Table 1, without cropping or mirroring. The AlexNet weights are initialized with the 566-category ImageNet pretrained model, and conv1 to fc7 weights are kept fixed during training. The FCN weights are initialized with the reshaped 566-category ImageNet pre-trained model weights, and conv1 to conv-fc7 weights are also kept fixed. The fc8 and conv-fc8 layers directly connect with the LSTM recurrent neural network via max operations, and only fc8, conv-fc8 and the LSTM parameters are fine-tuned on the training videos. We have also tried to fine-tune the weights in conv-fc6 and conv-fc7, but the result was poor compared with keeping these layers fixed.

Methods	BLEU	METEOR
FGM [6] [8]	13.68	23.9
LSTM-YT [8]	31.19	26.87
S2VT RGB ($AlexNet$) [7]	-	27.9
MM-VDN ($AlexNet + 8 \times 8$)	**37.64**	**29.00**

Table 2. Comparison to other baselines. (FGM) is the factor graph model in [6]; (LSTM-YT) is the LSTM model in [8]; (S2VT RGB) is the basic RGB sequence to sequence model with AlexNet in [7]; (MM-VDN) is our model.

For a single scale, the original AlexNet whole-frame scale is better than the other two scales alone. We investigate the combinations of different input scales 1×1, 8×8, and 16×16 in the FCN model. The combination of 1×1 and 8×8 performs the best among all possibile combinations and gets a boost of 4.7 in BLEU value and a boost of 1 in METEOR value. Scale combination is a dataset-specific problem, and the best scale combination requires cross-validation. Our model is flexible enough to integrate different input scales and combinations of several scales depending on the dataset.

We use BLEU[5] and METEOR[1] scores to evaluate the generated sentences against all reference sentences. The comparison of the best MM-VDN model from the ablation study ($AlexNet + 8 \times 8$ score maps) to the other three baselines is listed in Table 2. Our MM-VDN model provides a distinct improvement in both BLEU and METEOR compared to the other three baselines.

4. Conclusion

This paper proposed a Multi-scale Multi-instance Video Description Network (MM-VDN).The model is shown to be effective on the task of video description generation, compared to the single scale whole-frame classification CNN.

References

[1] S. Banerjee and A. Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005. 2

[2] D. L. Chen and W. B. Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 190–200. Association for Computational Linguistics, 2011. 2

[3] S. Guadarrama, N. Krishnamoorthy, G. Malkarnenkar, S. Venugopalan, R. Mooney, T. Darrell, and K. Saenko. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *Proceedings of the 14th International Conference on Computer Vision (ICCV-2013)*, pages 2712–2719, Sydney, Australia, December 2013. 2

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 1

[5] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002. 2

[6] J. Thomason, S. Venugopalan, S. Guadarrama, K. Saenko, and R. Mooney. Integrating language and vision to generate natural language descriptions of videos in the wild. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING)*, August 2014. 2

[7] S. Venugopalan, M. Rohrbach, J. Donahue, R. J. Mooney, T. Darrell, and K. Saenko. Sequence to sequence - video to text. *CoRR*, abs/1505.00487, 2015. 1, 2

[8] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. J. Mooney, and K. Saenko. Translating videos to natural language using deep recurrent neural networks. *CoRR*, abs/1412.4729, 2014. 2

[9] R. Xu, C. Xiong, W. Chen, and J. J. Corso. Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In *Proceedings of AAAI Conference on Artificial Intelligence*, 2015. 2