Advanced Deep Learning Model for Video Annotation

Dig Vijay Kumar Yarlagadda
University of Missouri-Kansas City
5100 Rockhill Rd, Kansas City, Missouri, USA
dy5kc@umkc.edu

Yugyung Lee
University of Missouri-Kansas City
5100 Rockhill Rd, Kansas City, Missouri, USA
leeyu@umkc.edu

Abstract

Video captioning is a challenging problem, historically it has been achieved by applying image captioning techniques for videos. Recent development in deep residual networks have achieved state-of-art accuracy in object recognition in images. In this paper, we describe the development of a deep network inspired by deep residual network with RNNs and Inception modules for captioning videos. Leveraging recent advancements in GPU computing, our network takes a memory based approach and can scale across a distributed cluster. Our approach generates coherent sentences and achieves state-of-art accuracy for caption generation in videos.

1. Introduction

Object detection and image captioning is a heavily researched topic in the field of deep vision. The closest work to Inception V4 is Microsoft RESNET which emphasizes the importance of building deeper networks for achieving state-of-art accuracy. Many approaches such as Imagenet, Show and Tell, Deep visual semantic approach and Attention-based EncoderDecoder Networks are proposed but they are not adopted for video captioning and very limited in applications.

2. System Architecture

Our system uses an image encoder using the Inception-ResNet v2 model, which achieves 95.3% accuracy on the ImageNet classification task. This gives many points of improvement in the BLEU-4 metric over the systems used previously for video captioning and image captioning by Orioal [3].

The data sets we used for our system are Microsoft COCO dataset for image caption generation, object segmentation and dataset for training the inception architecture and Flickr30k image Dataset for training and testing the generated image captions.

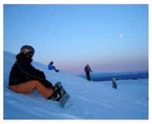
Image	**Human annotations test dataset**	**Generated by our system**
	A man in an orange hard hat and vest looks at an object in his hand while the man next to him stands on a ladder in order to reach the ceiling .	Two men wearing orange shirt stand next to each other near a ladder.
	A mother decides to take her child on a piggyback ride outside their apartment complex .	A child sit on top of a smiling woman.
	Two men in a tennis court in a city are standing on either side of the net talking to each other.	Two men stand next to each other on a tennis court with a fence.
	A snowboarder sits on a slope with skiers and boarders nearby .	Four people with boards sits on snow in evening.

Figure 1. Human annotations vs annotations generated by our system at 240,000 steps

The image model is tuned to satisfy the goal of a video captioning system: describing content in videos.

In the fine-tuning phase, the captioning system is improved by jointly training its vision and language components on human generated captions. This was inspired by the work of A. Karpathy [1]. This allows the captioning system to transfer information from the image that is specifically useful for generating descriptive captions, but which was not necessary for classifying objects. In particular, after fine-tuning it becomes better at correctly describing the colors of objects. Importantly, the fine-tuning phase must occur after the language component has already learned to generate captions - otherwise, the noisiness of the randomly initialized language component causes irreversible corruption to the vision component.

Although it is sometimes not clear whether a description should be deemed successful or not given an image, prior art has proposed several evaluation metrics. The most reliable is to ask for raters to give a subjective score on the usefulness of each description given the image. The rest of the metrics can be computed automatically assuming one has access to groundtruth, i.e. human generated descriptions. The most commonly used metric so far in the image description literature has been the BLEU score [38], which is a form of precision of word n-grams between generated and reference sentences. Even though this metric has some obvious drawbacks, it has been shown to correlate well with human evaluations.

The TensorFlow implementation achieves the same level of accuracy with significantly faster performance: time per training step is just 0.7 seconds in TensorFlow compared to 3 seconds in DistBelief on an Nvidia K20 GPU.

A natural question is whether our captioning system can generate novel descriptions of previously unseen contexts and interactions. The system is trained by showing it hundreds of thousands of images that were captioned manually by humans, and it often re-uses human captions when presented with scenes similar to what its seen before.

Our model does indeed develop the ability to generate accurate new captions when presented with completely new scenes, indicating a deeper understanding of the objects and context in the images. Moreover, it learns how to express that knowledge in natural-sounding English phrases despite receiving no additional language training other than reading the human captions.

The training process should be expose these methods to even larger datasets (e.g., YouTubeClips [6] and TACoS-MultiLevel) which contain thousands of lexical entries and dozens of hours of videos. As a result, the video captioning task becomes much more challenging, and the generation performance of these methods is usually low on these large-scale datasets.

We have used variation of deep residual networks along with inception module for our network, while their model is wider but not deep. Further, their model predicts multiple word at each frame, and combines them using SVO, but our model considers entire context into consideration and predicts entire sentences but not as individual words. This is different from word by word approach used by Subhashini *et al.* [2]

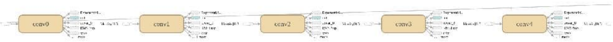


Figure 2. Convolutional modules

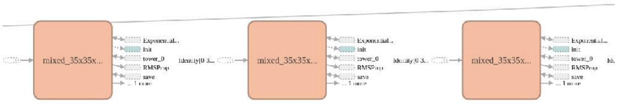


Figure 3. Recurrent Neural Network modules

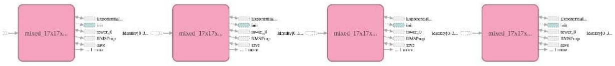


Figure 4. Dropout and regularization techniques

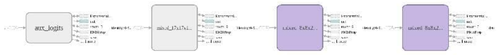


Figure 5. Final classification layer

Image captioning translation using RNNs can be treated as a special case of video captioning when each video has a single frame and no temporal structure. The amount of data handled by an image captioning method is much (dozens of times) less than that handled by a video captioning method. As a result, image captioning only requires computing object appearance features, but not action/motion features.The overall structure of an image captioner (instance-to-sequence) is also usually simpler than that of a video captioner (sequence-to-sequence). Some other methods, addressed the problem of retrieving sentences from training database to describe a sequence of images. They proposed a local coherence model for fluent sentence transitions, which serves a similar purpose of our paragraph generator.

A simple RNN is able to model temporal dependency for a small time gap, it usually fails to capture long-term temporal information. To address this issue, the GRU is designed to adaptively remember and forget the past. The very early video captioning method based on RNNs extends the image captioning methods by simply average pooling the video frames. Then the problem becomes exactly the same as image captioning. However, this strategy works only for short video clips where there is only one major event, usually appearing in one video shot from the beginning to the end. One difference between our framework and theirs is that we additionally exploit spatial attention. The other difference is that after weighing video features with attention weights, we do not condition the hidden state of our recurrent layer on the weighted features.

BLEU-4 SCORES	
Our system	46.3
LSTM-E	45.3
GRU-RCN	43.3
Glove + Deep Fusion	42.1
SA	41.9
LSTM - YT	33.3
NIC v2	32.1
FGM	13.7

Table 1. Comparison of BLEU scores for 4-grams

METEOR	
Our system	$tocalculate$
FGM	23.9
LSTM - YT	29.1
SA	29.6
LSTM-E	31.0
Glove + Deep Fusion	31.4
GRU-RCN	31.6

Table 2. Comparison of METEOR scores

References

[1] A. Karpathy and F. Li. Deep visual-semantic alignments for generating image descriptions. *CoRR*, abs/1412.2306, 2014.

[2] S. Venugopalan, L. A. Hendricks, R. J. Mooney, and K. Saenko. Improving lstm-based video description with linguistic knowledge mined from text. *CoRR*, abs/1604.01729, 2016.

[3] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge. *CoRR*, abs/1609.06647, 2016.