Project final report

Project title: Real-time object detection in videos
Team number: 8 / Dig Vijay Kumar Yarlagadda (Class ID: 26)

The paper below is written following strict 6-page format for submission to IEEE International Conference on Multimedia and Expo 2017.

I am including GitHub URL and video URL below.

GitHub URL: https://github.com/digvijayky/Tensorflow-real-time-video-analysis

VideoURL:https://umkc.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=ffbee666-03f5-42f5-ba40-17b44beede34

Please find latex files of the paper in the following link:
https://drive.google.com/file/d/0BxeFK0-zgXsAdkoxdXBycFIzbW8/view?usp=sharing

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 developments 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

Video captioning is a challenging task due to change of video frames at high rates and lack of light and efficient method to caption frames in real-time. Object detection and image captioning are improved due to recent advancements in the field of deep learning. Microsoft ResNets [12] and Inception V4 [26] architectures are two important deep networks which has achieved state-of-art accuracy on the ImageNet [7] dataset. Many approaches such as Show and Tell [31], Deep visual semantic approach [15] and Attention-based EncoderDecoder Networks [4] are proposed but they are not adopted for video captioning and have very limited applications. The training process for video captioning is challengin due to size of large datasets such as YouTubeClips [11] and TACoS-MultiLevel [23] which contain thousands of lexical entries and dozens of hours of videos. As a result, the video captioning task is a complex task, and the generation performance of prior art methods is usually low on these large-scale datasets.

2. Related Work

The large volume of video data has motivated approaches to caption video. Many efforts are devoted to design effective feature representations and robust classifiers. Recently, researchers have attempted to apply deep learning techniques to the video domain across supervised deep learning and unsupervised feature learning.

Some methods have 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.

The success of CNN modules on image analysis tasks has stimulated the utilization of deep features for video classification. The idea is to treat a video clip as a collection of frames, and then for each frame feature representation could be derived by running a feed-forward pass till a certain ully-connected layer with state-of-the-art deep models pre-trained on ImageNet [7], including AlexNet [17], VGGNet [25], GoogleNet [27] and ResNet [12]. Finally, frame-level features are averaged into video-level representations as inputs of standard classifiers for recognition, such as the well-known SVMs or softmax classifiers.

Among the works on image-based video classification, Zha *et al.* [10] systematically studied the performance of image-based video recognition using features from different layers of deep models together with multiple kernels for classification. They demonstrated that off-the-shelf CNN features coupled with kernel SVMs can obtain decent recognition performance. Motivated by the advanced feature encoding strategies in images in [24], Xu *et al.* [33] proposed to obtain video-level representation through VLAD encoding, which can attain performance gain over the trivial averaging pooling approach.Donahue *et al.* [8] trained two two-layer LSTM networks for action recognition with features from the two-stream approach. They also tried to fine-tune the CNN models together with LSTM but did not obtain significant performance gain compared with only training the LSTM model. Wu *et al.* [32] fused the outputs of LSTM models with CNN models to jointly model spatial-temporal clues for video classification and observed that CNNs and LSTMs are highly complementary. Ng *et al.* [20] further trained a 5-layer LSTM

model and compared several pooling strategies. But these methods have traded accuracy and performance and did not achieve consisently high accuracies on all video datasets.

3. Model

Overview In this paper, we propose a deep network to generate sentences from videos. As this is a problem of machine translation, we can leverage a recurrent neural network based language sequence model to achieve good performance.

A convolutional neural network accepts a fixed-sized vector as input and produce a fixed-sized vector as output and perform this mapping using a fixed amount of computational steps. In contrast, recurrent neural networks operate over sequences and the number of times the transformation can be applied can be varied. Long Short Term Memory (LSTM) [13] network, a RNN architecture with learning algorithm, capable of learning long-term dependencies. We input features extracted from an image and apply the principles developed in neural machine translation to generate sentences. Gated Recurrent Unit (GRU) [5] has gated units which can modulate information without having separate memory cells. Chung et al. [6] has shown that performance of GRU is similar to that of LSTM. RNNs can also be used as generative models: they can learn from sequences of images and sentences and generate sentences for new images.

3.1. Residual network layer

Residual network [12] is a standard feed-forward convolutional neural network with added skip connections that bypass few convolutional layers at a time. Each bypass gives rise to a residual block in which the convolution layers predict a residual that is added to the input of that block. These bypass or shortcut connections which are added parallel to convolutional layers allowing gradients to propagate to the early layers of the network, thereby solving one of the major problems with RNN: vanishing gradients problem. The connections can be average pooling for size reduction and zero padding for size enlargement among others. They can train faster and achieve better accuracy than other neural networks. Figure 2 shows basic building block of ResNet [9] introduced in [12].

Recent studies by A Veit *et al*. [29] and Liao *et al*. [18] showed similarities between residual networks and recurrent neural networks. As video captioning can be effectively done through RNNs, we use deep residual networks on images and LSTMs for generating sentences.

Our system uses an image encoder tuned from Inception-ResNet v2 model, which achieves 95.3% accuracy on the ImageNet classification task. The image model is tuned to satisfy the goal of a video captioning system: describing content in videos. This gives many points of improvement in the BLEU-4 metric [22] over the systems used previously for video captioning and image captioning by Vinyals *et al*. [31].

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*. [30].

3.2. Sentence generation using LSTMs

Each image frame extracted from the video is processed using the ResNet, it is sent over to the LSTM model which generates sentences from the image features it has received. Image is input to all LSTMs only once with description of the image contents, the same parameters are shared across all LSTMs. As LSTMs are connected in sequence, output of an LSTM is fed to next LSTM in the sequence. The sequence of LSTMs have a start word and stop word which designates the start and end of the sentence. By emitting the stop word the LSTM signals that a complete sentence has been generated. Vinyals *et al*. [31] has noted that feeding image at each time step will result in overfitting. The loss we minimize is the probability of predicting a correct word. We are trying to minimize the number of words in a sentence, as we noted that more words in a sentence doesn't result in better descriptions, instead we notice a reduction in BLEU score.

The LSTM captioning system is tuned by training its vision and language components on human generated captions as in [16]. Only the information required for generating captions is input to LSTMs using transfer learning [21] techniques.The tuning is done after language components has learned to generate sentences to avoid corruption of LSTMs by deep network.

We are using beam search approach with a beam size of 16 to iteratively consider the set of best sentences as candidates to generate sentences and keep only the best sentences among the generated sentences. Beam search is a form of greedy search that does not give an exact highest probability output sequence, but lets us get some number of candidates, called the beam size. We then compute the set of most likely first words, instead of a single most likely word. We then continue to compute the list of most likely words using the previous words as the candidates. This approach significantly improves accuracy.

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 less than that handled by a video captioning method. As a result, image captioning only requires computing object appearance

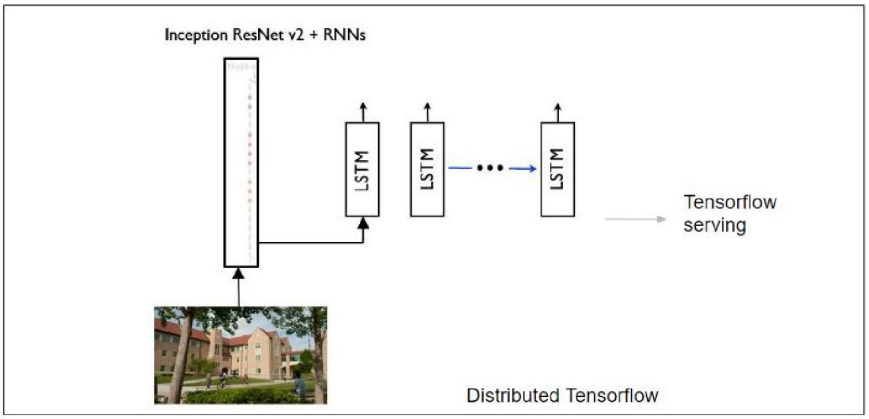

Figure 1. System overview

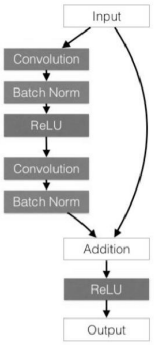

Figure 2. ResNet basic block

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).

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.

4. Experiments

4.1. Evaluation metrics

There are several metrics available for evaluating machine translation such as BLEU [22], METEOR [1] and CIDEr [28] being popular among others. BLEU score computes precision of word n-grams (n=1,2,3,4) between generated and reference sentences.

METEOR evaluates a generated sentence by calculating a score based on explicit word-to-word matches between generated sentence and human annotation by creating a mapping between words in the sentences.

Table 1. Comparison of BLEU-4 scores

System	BLEU-4 score
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

4.2. Datasets

Microsoft COCO dataset Microsoft COCO: Common Objects in Context [19] dataset consists of 330,000 images described using 1.5 million captions [3]. We use this dataset for training object detection, object segmentation, inception architecture and testing the generated captions.

Flickr30k image dataset Flickr30k dataset [34] consists of 31,783 images described by 158,915 crowd-sourced

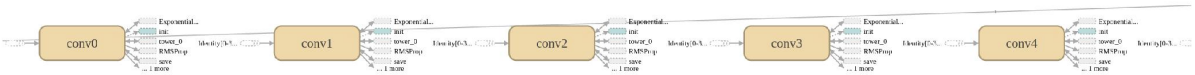

Figure 3. Convolutional modules

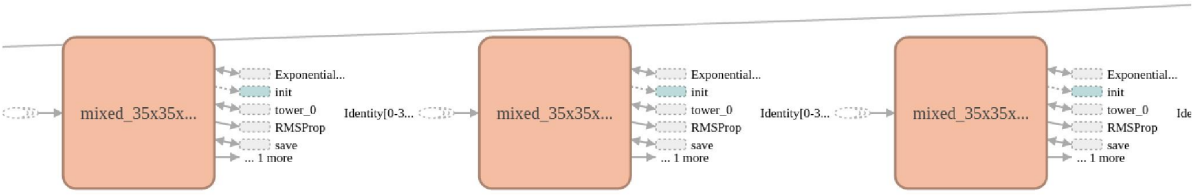

Figure 4. Recurrent Neural Network modules

captions. We use this dataset in complement to Microsoft COCO dataset to train and validate the video captions.

Microsoft Research Video Description Corpus MSVD [2] dataset consists of 1,970 videos captioned on Amazon Mechanical Turk using 120K single sentences. We use this dataset for evaluating our network by computing BLEU and METEOR scores.

The evaluation metrics on MSVD dataset are presented in tables 1 and 2. Our system achieve state-of-art performance across both the metrics.

4.3. Results

Table 2. Comparison of METEOR scores

System	METEOR score
Our system	**35.1**
FGM	23.9
LSTM - YT	29.1
SA	29.6
LSTM-E	31.0
Glove + Deep Fusion	31.4
GRU-RCN	31.6

Our deep network achieves state-of-art performance with decent amount of training and generate mostly accurate, concise and sensible sentences. In this section, we explain our details of training and performance of generated sentences.

4.3.1 Training details

The system is trained on a machine with Intel Core i7 processor (3.5GHz), 16GB RAM, and an NVIDIA GeForce GTX 960M GPU containing of 640 cores and 4GB GDDR5 graphics memory. Our deep network is trained for 240,000 steps. Training our network took more than 300 hours to complete on this machine.

4.3.2 Generation Results

The captions generated are presented in Figure 7. We noted our system generates sentences with less number of words than competing systems. We trained our network such that actions are included while common objects are excluded. Our system can generate novel descriptions of unseen contexts and interactions. As the system is trained using hundreds of thousands of images that were captioned manually by humans on Amazon Mechanical Turk, it re-uses human captions for previously seen scenes. It thus demonstrates the ability to generate captions on new scenes, indicating that the system has deep understanding of objects and context in images. It also forms natural phrases in English without having any additional training on language datasets.

Our TensorFlow implementation achieves high level of accuracy with good performance: time per training step is just 0.8 seconds in TensorFlow. For each ten video frames, captions are generated in less than a second. With a powerful GPU we can process images and generate captions at higher frame rates.

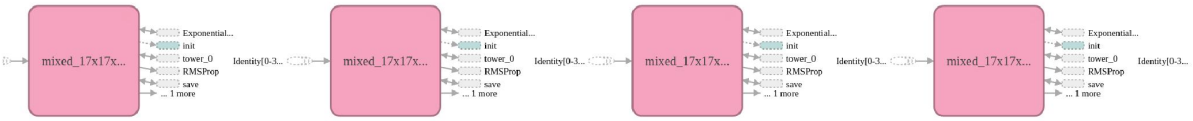

Figure 5. Dropout and regularization techniques

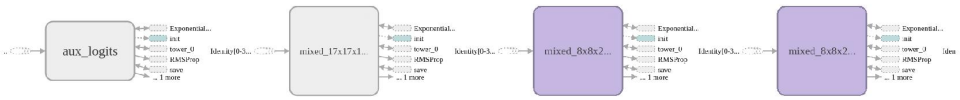

Figure 6. Final classification layer

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 7. Comparison of human annotations and annotations generated by our system trained at 240K steps

5. Conclusion

Our results prove that deep networks can achieve state-of-art accuracies in video captioning task. The main advantage of our system is that it can run in real-time and generate concise and accurate sentences. As our system is build on principles of transfer learning, it can also be used across wide range of similar applications.

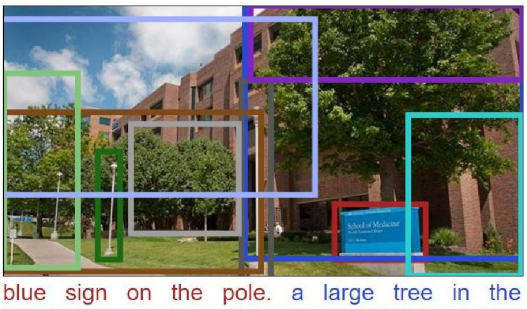

Figure 8. Comparison of sentences generated by dense captioning system by Johnson *et al.* [14] and our system

References

[1] S. Banerjee and A. Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. 2005.

[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.

[3] X. Chen, H. Fang, T. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick. Microsoft COCO captions: Data collection and evaluation server. *CoRR*, abs/1504.00325, 2015.

[4] K. Cho, A. Courville, and Y. Bengio. Describing multimedia content using attention-based encoder-decoder networks.

IEEE Transactions on Multimedia, 17(11):1875–1886, 2015.

[5] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *CoRR*, abs/1409.1259, 2014.

[6] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.

[8] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2625–2634, 2015.

[9] M. S. Ebrahimi and H. K. Abadi. Study of residual networks for image recognition.

[10] B. Fernando, A. EDU, and S. Gould. Learning end-to-end video classification with rank-pooling.

[11] 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 IEEE International Conference on Computer Vision*, pages 2712–2719, 2013.

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[13] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[14] J. Johnson, A. Karpathy, and L. Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. *arXiv preprint arXiv:1511.07571*, 2015.

[15] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137, 2015.

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

[17] 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.

[18] Q. Liao and T. A. Poggio. Bridging the gaps between residual learning, recurrent neural networks and visual cortex. *CoRR*, abs/1604.03640, 2016.

[19] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014.

[20] J. Y. Ng, M. J. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond short snippets: Deep networks for video classification. *CoRR*, abs/1503.08909, 2015.

[21] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.

[22] 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.

[23] A. Rohrbach, M. Rohrbach, W. Qiu, A. Friedrich, M. Pinkal, and B. Schiele. Coherent multi-sentence video description with variable level of detail. In *German Conference on Pattern Recognition*, pages 184–195. Springer, 2014.

[24] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek. Image classification with the fisher vector: Theory and practice. *International journal of computer vision*, 105(3):222–245, 2013.

[25] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[26] C. Szegedy, S. Ioffe, and V. Vanhoucke. Inception-v4, inception-resnet and the impact of residual connections on learning. *arXiv preprint arXiv:1602.07261*, 2016.

[27] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[28] R. Vedantam, C. Lawrence Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4566–4575, 2015.

[29] A. Veit, M. Wilber, and S. Belongie. Residual networks are exponential ensembles of relatively shallow networks. *arXiv preprint arXiv:1605.06431*, 2016.

[30] 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.

[31] 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.

[32] Z. Wu, X. Wang, Y.-G. Jiang, H. Ye, and X. Xue. Modeling spatial-temporal clues in a hybrid deep learning framework for video classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 461–470. ACM, 2015.

[33] Z. Xu, Y. Yang, and A. G. Hauptmann. A discriminative cnn video representation for event detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1798–1807, 2015.

[34] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

Future work

As our approach is unique with custom layered architecture and works well for all categories of videos, we could develop many custom networks for different tasks like action recognition or object tracking and use it across a wide range of applications. Also, by applying transfer learning techniques the knowledge learned by our network can be used in conjunction with other networks across domains.