A Multitask Learning Encoder-Decoders Framework for Generating Movie and Video Captioning

Oliver Nina
Ohio State University
Columbus, OH 43210
nina.3@osu.edu

Washington Garcia
University of Florida
Gainsville, FL 32611
w.garcia@ufl.edu

Scott Clouse
ACT 3
Dayton, OH 45431
hsclouse@ieee.org

Alper Yilmaz
Ohio State University
Columbus, OH 43210
yilmaz.15@osu.edu

Abstract

Learning visual feature representations for video analysis is non-trivial and requires a large amount of training samples and a proper generalization framework. Many of the current state of the art methods for video captioning or movie description rely on encoding mechanisms through recurrent neural networks to encode temporal visual information extracted from video data. In this paper, we introduce a novel multitask encoder-decoder framework for generating automatic semantic description and captioning of video sequences. In contrast to current approaches, at training time our method relies on multiple distinct decoders to train a visual encoder in a multitask fashion. Our method shows improved performance over current state of the art methods in several metrics on both multi-caption and single-caption datasets. Our method is the first method to use a multi-task approach for encoding video features. Furthermore, our method demonstrates its robustness on the Large Scale Movie Description Challenge (LSMDC) 2017 where our method won the movie description task. Based on human subject evaluations from the competition, our method was ranked as the most helpful algorithm for the visually impaired. [1]

1 Introduction

Video captioning and description is a well known problem in computer vision that can be formulated as follows: given a video segment, we would like an intelligent system to automatically summarize and describe the central activity or activities of interacting animated or inanimate objects in a video in natural language. Semantic video description is a challenging problem and requires solving a number of different computer vision problems simultaneously: object detection and classification, action recognition, saliency detection, and natural language processing among others.

Having an automatic semantic video description system would be advantageous with many potential applications such as helping the visually impaired by providing automatic narration of videos. Current systems that transcribe videos or films rely heavily on human effort and are labor intensive and costly, limiting the content available to the blind.

[1] The work presented on this paper is currently under review as a journal submission. Much of the text has been reduced and modified to fit the theme of the AI for Social Good workshop. A preprint version of our paper is found at https://arxiv.org/abs/1809.07257

Preprint. Work in progress.

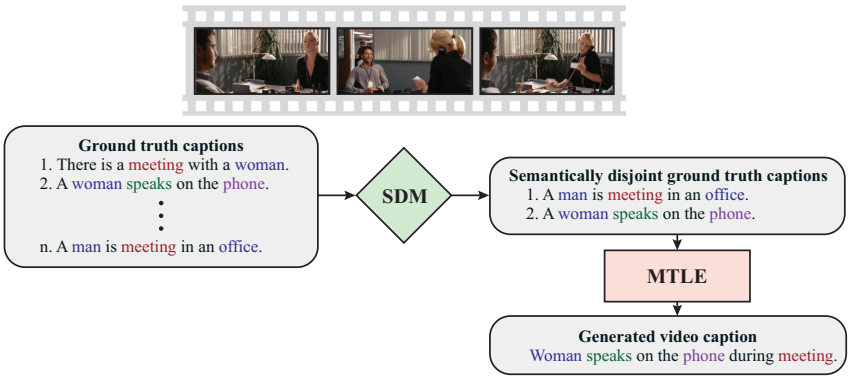

Figure 1: Example of Multitask Learning Encoder (MTLE)-based video captioning. Ground truth captions are compared with a semantic distance matrix (SDM) to find semantically disjoint caption samples, which are passed to our MTLE method to produce a a video's caption.

In this work, we propose a novel multitask framework that is comprised of an encoder, whose weights are the focus of the learning process, and multiple decoders that correspond to the encoder and whose loss function and selection process bestow a broader semantic meaning of the video in question. We call our method *MTLE* which stands for multitask learning encoder because our model focuses on generalizing the encoder weights and because of the way in which the encoder and decoders are arranged. Decoders and encoder are trained in a multitask fashion on semantically disjoint labels to enhance the system's semantic knowledge. Such disjoint labels are chosen systematically to broaden the semantic information fed to the decoders in order to generalize the encoder. Our encoder consists of a bi-directional recurrent neural network, more specifically a long short term memory (LSTM) unit trained with two or more distinct decoders. Each decoder uses labeled samples, or video captions and convert them into textual features which are later projected into a measurable semantic space.

Our method is also designed to accommodate single caption datasets. The objective function includes a regularization term that both leverages multiple caption scenarios and augments training samples when there is a limited amount of training data. Thus, our method does not depend on a large number of training labels and can handle datasets with limited number of annotations such as the Large Scale Movie Description Challenge (LSMDC) (Rohrbach et al., 2017). Figure 1 shows the outline of our method. Our proposed method shows improvements over the current baseline in public datasets that contain single or multiple annotations per video including LSMDC (Rohrbach et al., 2017), MSR-VTT (Xu et al., 2016), TRECVID (NIST, 2017), MSVD (Chen and Dolan, 2011) among others.

2 Proposed Method

2.1 MTL Encoder-N-Decoder

Multitask learning (MTL) studies the problem of estimating multiple functions jointly by exploiting shared structures in order to improve generalization (Caruana, 1998). In our approach, we treat a single encoder-decoder framework as a function f and compose several decoders with a shared encoder to approximate the final loss through MTL. However, because the relation between their tasks is non-linear, solving this MTLE problem is non-trivial (Ciliberto et al., 2017). Nonetheless, we consider the problem as a linear MTL approximation in order to solve it using a convex optimization procedure.

Formally, given a set of functions: $f_1, ..., f_n : \mathcal{X} \rightarrow \mathbb{S}$ and a corresponding set of training samples $(\mathbf{V}_l, \mathbf{X}_n)$, with $\mathbf{X}_i \in \mathbb{S}$ and $\mathbf{X}_i \in \mathbb{V}$, we define $\mathbb{S} : \mathbb{S} \subseteq \mathbb{R}$ as the semantic space of all captions. The visual space is defined as $\mathbb{V} : \mathbb{V} \subseteq \mathbb{R}$ with $\mathbf{V}_l = \{\mathbf{v}^1, ..., \mathbf{v}^t\}$,where l corresponds to the l-th video in the training data with its corresponding caption $\mathbf{X}_n$. In this sense, for each f function, the following holds: $f_n(\mathbf{V}_l) = \mathbf{X}_n$.

We model task relations as a set of $\mathcal{P}$ functions with a constraint function $\gamma : \mathbb{V}^l \rightarrow \mathbb{S}^{\mathcal{P}}$ and require $\gamma(f_1(\mathbf{V_1}), ..., f_n(\mathbf{V_1})) = f'(\mathbf{V_1})$, where $f'(\mathbf{V}_l)$ corresponds to the "true" semantic position of $\mathbf{V}_l$ in $\mathcal{S}$.

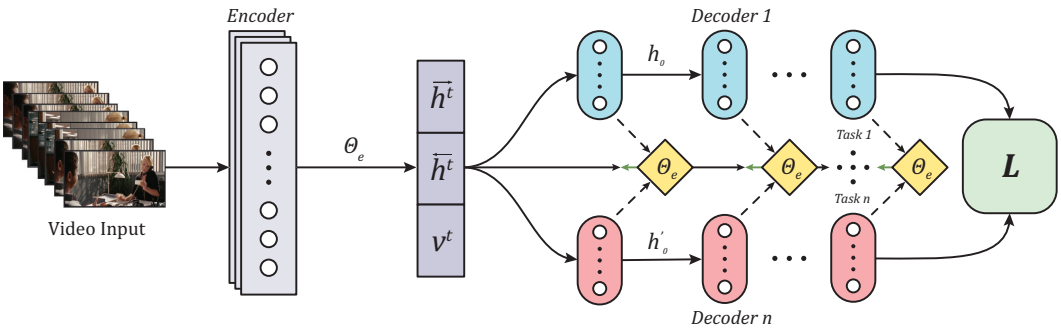

Figure 2: Overview of our MTLE method. Video frames are passed to a CNN encoder for feature vector generation. From the encoder, feature vectors are passed to an RNN-based bi-directional attention encoder. The output of this process is the concatenated visual feature encoding of the video, which is passed to a multitask conditional decoder with soft attention. The parameters of the RNN-based bi-directional attention encoder are denoted by θ_e, and are trained simultaneous to the multitask decoder. Green arrows denote back-propagation.

Our problem imposes a constraint in the range of γ, mainly, $\mathcal{X} \rightarrow \mathcal{C}$ to take values in the constraint set:

$$\mathcal{C} = \left\{ \mathbf{y} \in \mathbb{S}^n \mid \gamma(\mathbf{y}) = f'(\mathbf{V}_l) \right\} \subseteq \mathbb{S}^n, \tag{1}$$

Thus the goal is to find a good approximation $\hat{f} : \mathcal{X} \rightarrow \mathcal{C}$ for the following multitask *expected risk* minimization problem:

$$\begin{gathered} \min_{f:\mathcal{X}\rightarrow\mathcal{C}} \mathcal{E}(f), \\ \mathcal{E}(f) = \frac{1}{N}\sum_{n=1}^{N} \frac{1}{M_n} \sum_{m=1}^{M_n} \mathcal{L}(f_n(\mathbf{x}_{mn}), \mathbf{y}_{mn}) d\rho_n(\mathbf{x}, \mathbf{y}), \end{gathered} \tag{2}$$

where $\mathcal{L} : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ is the loss function of prediction errors for each task $n = 1, ..., N$, ρ_n is the distribution on $\mathcal{X} \times \mathbb{R}$ from where M training points $(\mathrm{x}_{mn}, \mathrm{y}_{mn})_{n=1}^{m_n}$ have been sampled independently. If $\mathcal{C}$ is a non-linear subset, the minimization in (2) is difficult to solve through convex optimization. Hence, we assume that $\mathcal{C}$ is a linear subset. Furthermore, similar to (Dinuzzo et al., 2011), we approximate (2) by calculating a matrix of caption pairs that generates positive semi-definite matrix $\mathcal{A}$ that encourages linear relations between the tasks discussed in (Dinuzzo et al., 2011). Furthermore, in contrast to (Dinuzzo et al., 2011), we treat matrix $\mathcal{A}$ as a sampling distribution rather than a second term to the solution as follows:

$$\min_{f=(f_1,...,f_n)\in\mathcal{H}^n} \lambda \sum_{n=1}^{N} \sum_{m=1}^{M} \mathcal{L}(f(\mathbf{x}_m), \mathbf{y}_m) \cdot \mathcal{A}_{nc} \langle n, c \rangle, \tag{3}$$

where λ is a normalization term, $\mathcal{L}$ is the loss function of the system, $\mathcal{A} = (\mathcal{A}_{nc})_{c,n=1}^{N}$ and is called the Semantic Distance Matrix, $\mathcal{H}$ represents a reproducing kernel Hilbert space and in order to evaluate the minimization as a linear combination, we assume that the functions f_n are part of $\mathcal{H}$.

The total cost is obtained by adding the total loss $\mathcal{L}$ of each n independent task of the system along with a regularization term α. Our method without a semantic distance constraint could allow for any number of decoder tasks. Having a large number of decoder tasks could be advantageous for improving the generalization of the weights onto a larger semantic space. However, because of hardware limitations and for simplicity, we set $n = 2$ in our model. Thus, in such setup, our loss function is approximated as follows:

$$\sum_{n} \mathcal{L}(f(\mathbf{x}), \mathbf{y}) \cdot \mathcal{A} \langle f_n, f_c \rangle \approx \delta_1(\boldsymbol{\nu}, y) + \delta_2(\boldsymbol{\nu}, y) + \alpha, \tag{4}$$

where δ represent the loss of each of the decoders, a negative log likelihood. Substituting each term by its corresponding value we obtain:

$$\delta_1(\boldsymbol{\nu}) \;+\; \delta_2(\boldsymbol{\nu}) \;+\; \alpha \quad = \quad \sum_i -\log P^n_{\hat{\mathbf{x}}} \;+\; \sum_i -\log P^c_{\hat{\mathbf{x}}} \;+\; \eta \sum_i |P^n_{\hat{\mathbf{x}}} - P^c_{\hat{\mathbf{x}}}|. \quad (5)$$

where f_n represents the main reference task and f_c is the complement task to f_n, i corresponds to the index of each caption word, η is a parameter set to [0,1] depending if the videos belong to a single or multi-caption dataset respectively. $P^n_{\hat{\mathbf{x}}}$ represents the probability of a centroid caption in semantic space related to task n.

3 Results

Table 1: Human Evaluation from LSMDC 2017

	Human Score
Reference (Human)	4.46
MTLE (Ours)	**2.50**
Fcrerank (Kaufman et al., 2017)	2.18
PostProp (Dong et al., 2016)	2.17
FuseNet (Rohrbach et al., 2017)	1.96
attn2l (Rohrbach et al., 2017)	1.68

Table 2: MSVD Dataset

Model	BLEU	METEOR	ROUGE	CIDEr
FGM (Thomason et al., 2014)	0.137	0.239	-	-
DR-LSTM (Venugopalan et al., 2015)	0.312	0.269	-	-
Yao (Yao et al., 2015)	0.403	0.290	-	0.480
S2VT MT (Venugopalan et al., 2016)	0.421	0.314	-	-
h-RNN-VGG (Yu et al., 2016)	0.443	0.311	-	-
HRNE (Pan et al., 2016)	0.438	0.331	-	-
SCN (2017) (Gan et al., 2017)	0.511	0.335	-	-
Pasunuru (2017) (Pasunuru and Bansal, 2017)	0.545	**0.360**	-	-
Pre-Split (Yao et al., 2015)				
Baseline C3D	0.411	0.286		
Baseline Googlenet	0.4717	0.3213	0.6896	0.6958
Baseline Resnet	0.500	0.331	0.696	0.734
Baseline NASnet–A Large	0.502	0.339	0.698	0.824
Baseline PNASnet–5 Large	0.507	0.334	0.693	0.776
MTLE Googlenet	0.492	0.322	0.689	0.702
MTLE Resnet	**0.550**	0.336	0.702	0.786
MTLE NASnet–A Large	0.537	0.343	**0.706**	**0.835**
MTLE PNASnet–5 Large	0.516	0.337	0.699	0.798

4 Conclusion

This paper presents a novel multitask encoder-n-decoder framework for semantic movie and video description. Our method learns an improved video feature encoder by leveraging the diversity of captions setup in a multitask paradigm to solve a conjoint multitask loss function through a convex optimization. Based on impartial human evaluations, our method was ranked as the highest and most useful method for helping the visually impaired wining the LSMDC 2017 competition among other top research groups worldwide and is currently the state of the art method for this application.

References

Caruana R (1998) Multitask learning. In: Learning to learn, Springer, pp 95–133

Chen DL, Dolan WB (2011) Collecting highly parallel data for paraphrase evaluation. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL-2011), Portland, OR

Ciliberto C, Rudi A, Rosasco L, Pontil M (2017) Consistent multitask learning with nonlinear output relations. CoRR abs/1705.08118, URL http://arxiv.org/abs/1705.08118, 1705.08118

Dinuzzo F, Ong CS, Pillonetto G, Gehler PV (2011) Learning output kernels with block coordinate descent. In: Proceedings of the 28th International Conference on Machine Learning (ICML-11), pp 49–56

Dong J, Li X, Lan W, Huo Y, Snoek CG (2016) Early embedding and late reranking for video captioning. In: Proceedings of the 2016 ACM on Multimedia Conference, ACM, pp 1082–1086

Gan Z, Gan C, He X, Pu Y, Tran K, Gao J, Carin L, Deng L (2017) Semantic compositional networks for visual captioning. In: The IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), pp 1141–1150, DOI 10.1109/CVPR.2017.127

Kaufman D, Levi G, Hassner T, Wolf L (2017) Temporal tessellation: A unified approach for video analysis. In: The IEEE International Conference on Computer Vision (ICCV)

NIST (2017) TRECVID 2017 Video To Text (VTT) Task. video description dataset. http://www-nlpir.nist.gov/projects/tv2017/Tasks/vtt/

Pan P, Xu Z, Yang Y, Wu F, Zhuang Y (2016) Hierarchical recurrent neural encoder for video representation with application to captioning. In: The Conference on Computer Vision and Pattern Recognition, pp 1029–1038

Pasunuru R, Bansal M (2017) Multi-task video captioning with video and entailment generation. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pp 1273–1283

Rohrbach A, Torabi A, Rohrbach M, Tandon N, Pal C, Larochelle H, Courville A, Schiele B (2017) Movie description. International Journal of Computer Vision URL http://resources.mpi-inf.mpg.de/publications/D1/2016/2310198.pdf

Thomason J, Venugopalan S, Guadarrama S, Saenko K, Mooney R (2014) 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), Dublin, Ireland

Venugopalan S, Xu H, Donahue J, Rohrbach M, Mooney R, Saenko K (2015) Translating videos to natural language using deep recurrent neural networks

Venugopalan S, Hendricks LA, Mooney R, Saenko K (2016) Improving lstm-based video description with linguistic knowledge mined from text. In: Conference on Empirical Methods in Natural Language Processing (EMNLP)

Xu J, Mei T, Yao T, Rui Y (2016) MSR-VTT: A large video description dataset for bridging video and language. The IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)

Yao L, Torabi A, Cho K, Ballas N, Pal C, Larochelle H, Courville A (2015) Describing videos by exploiting temporal structure. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV)

Yu H, Wang J, Huang Z, Yang Y, Xu W (2016) Video paragraph captioning using hierarchical recurrent neural networks. In: The Conference on Computer Vision and Pattern Recognition, pp 4584–4593