IMFD_IMPRESEE at TRECVID 2020: Description Generation by Visual-Syntactic Embedding *

Jesus Perez-Martin^{1,2}, Benjamin Bustos^{1,2}, Jorge Pérez^{1,2}, and Juan Manuel Barrios^{1,3}
^{1}Department of Computer Science, University of Chile
^{2}Millennium Institute for Foundational Research on Data
^{3}Impresee Inc., CA, USA
{jeperez, jperez, bebustos}@dcc.uchile.cl, juan.barrios@impresee.com

Abstract

In this paper we present an overview of our participation in TRECVID 2020 Video to Text Description Challenge [1]. Specifically, we participated in the Description Generation subtask by extending of our recent paper [21]. We address the limitation of previous video captioning methods that have a strong dependency on the effectiveness of semantic representations learned from visual models, but often produce syntactically incorrect sentences which harms their performance on standard datasets. We consider syntactic representation learning as an essential component of video captioning. We construct a visual-syntactic embedding *by mapping into a common vector space a visual representation, that depends only on the video, with a syntactic representation that depends only on Part-of-Speech (POS) tagging structures of the video description. We integrate this joint representation into an encoder-decoder architecture that we call Visual-Semantic-Syntactic Aligned Network (SemSynAN), which guides the decoder (text generation stage) by aligning temporal compositions of visual, semantic, and syntactic representations. Considering different datasets for training the model, such as VATEX and TGIF, our results represent third place by teams on the TRECVID 2020 Challenge for METEOR and CIDEr-D metrics. We also show that paying more attention to syntax improves the quality of generated descriptions.*

1. Video to Text: Description Generation

In this year's TRECVID Challenge [1] we extended our work [20, 21] by training and validating the model in other datasets: VATEX, TGIF, and VTT20. We developed our method under the assumption that integrating semantic-concept representations with syntactic representations can improve the quality of generated sentences. Now, we briefly present the model and we refer the reader to Perez-Martin *et al.* [21] for details.

For tasks like *video retrieval from descriptions* and *video descriptions retrieval from videos* [5, 6, 8, 15, 16], the joint visual-semantic embeddings have a successful application. These embeddings are constructed by combining two models: a *language model* that maps the captions to a language representation vector, and a *visual model* that obtains a visual representation vector from visual features. Both models are trained for projecting those representations into a joint space, minimizing a distance function. Dong *et al.* [6] obtain high-performance in retrieval tasks by using the same multi-level architecture for both models, and training with the *triplet-ranking-loss* function [7].

For *video description generation*, these embeddings have not been widely explored [10, 14, 18]. In LSTM-E [18], a joint embedding component is utilized to bridge the gap between visual content and sentence semantics. This embedding is trained by minimizing the *relevance loss* and *coherence loss* simultaneously. In SibNet [14], autoencoder for visual information, and a visual-semantic embedding for semantic information are exploited. These joint embeddings only consider the implicit contextual information of word vectors. To improve the perplexity and syntax correctness of generated sentences, we learn a new representation of videos with suitable syntactic information.

We propose a model to create *visual-syntactic embeddings* by exploiting the Part-of-Speech (POS) templates of video descriptions. We do this by learning two functions: $\phi(\cdot)$ that maps videos, and $\omega(\cdot)$ that maps (POS tags of) captions,

*This work was funded by ANID - Millennium Science Initiative Program - Code ICN17_002 and ANID/Doctorado Nacional/2018-21180648.

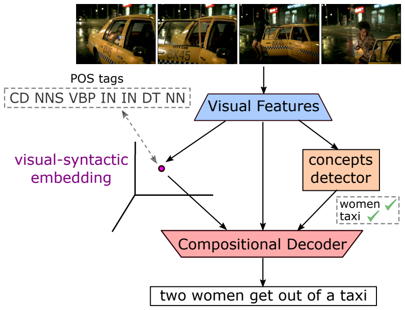

Figure 1. Video description generation with visual-syntactic embedding. The method computes high-level semantic and syntactic representations from the visual representation of the video. Next, the decoder generates a sentence from them.

both into a common vector space. The learning process is based on a *match and rank* strategy, and ensures that videos and their corresponding captions are mapped close together in the common space. Then, when producing features for the decoder architecture (see the next point), we can use the function $\phi(\cdot)$ to map the input video and generate our desired visual-syntactic embedding. To the best of our knowledge, this is the first approach to jointly learn embeddings from videos and (POS tags of) descriptions. Moreover, our proposal constitutes the first instance of an effective use of a ranking model to obtain syntactic representations of videos.

1.1. The Encoder

For encoding the input video x, we propose an architecture of three stages (Figure 1). The first stage consists in compressing the video into a global representation, which combines two standard visual features extractors: 2D-CNN feature vectors and 3D-CNN feature vectors. The second stage in our encoder consists of producing a semantic representation of the video. To produce it, based on video captioning studies like Chen *et al*. [3] and Gan *et al*. [9], we use a standard concept detector.

1.1.1 Visual-Syntactic Embedding

The third stage in our encoder architecture produces what we call *visual-syntactic embedding*. We claim that cues about the syntactic structure of the video's descriptions can be directly extracted from a video without necessarily extracting explicit information about the entities or objects participating. For example, there may be several videos in our dataset that, because of their structure, share a description pattern of the form

$$\langle\text{object1}\rangle \quad \langle\text{object2}\rangle \quad \langle\text{action}\rangle \quad \langle\text{object3}\rangle$$

as in a description like "⟨The dog⟩ and ⟨the cat⟩ ⟨are lying⟩ on ⟨the floor⟩." We propose to train a model to compute a suitable syntactic representation of descriptions directly from the input video. We attack this representation learning problem as a *Part-Of-Speech template retrieval* problem.

Given a pair $(x, y) \in \mathcal{D}$, our strategy is to learn how to map a visual representation of x and the sequence of POS tags of y into a d-dimensional common space. Specifically, the aim is to learn two mapping functions $\phi(\cdot)$ and $\omega(\cdot)$ that map from visual features and POS template vectors, respectively, to the joint embedding space. The learning process is based on a *match and rank* strategy, ensuring that videos and their corresponding captions are mapped close together in the common space. In our architecture, we use an improved version of triple-ranking loss, which penalizes the model taking into account the hardest negative examples [6, 7, 21].

1.2. The Decoder

Given the output of our encoder, *i.e*., the averaged feature representation, the concept detector vector and the visual-syntactic encoding, our decoder network generates the natural language description. We define it as a recurrent architecture that generates the tokens each word at a time. This recurrent architecture has four components to dynamically decide when to use visual-semantic, visual-syntactic, or semantic-syntactic temporal information in the generation process (Figure 2).

In detail, our decoder has three specialized recurrent layers based on compositional-LSTM network [3, 9], and an additional layer to combine the outputs of the three recurrent layers defined by two levels of what we call *fusion gates*. The role of this combination layer is to adaptively mix the outputs of the three recurrent layers, while the role of the recurrent layers is to capture temporal states related to a specific pair of feature information (*i.e*., visual-semantic, visual-syntactic, and

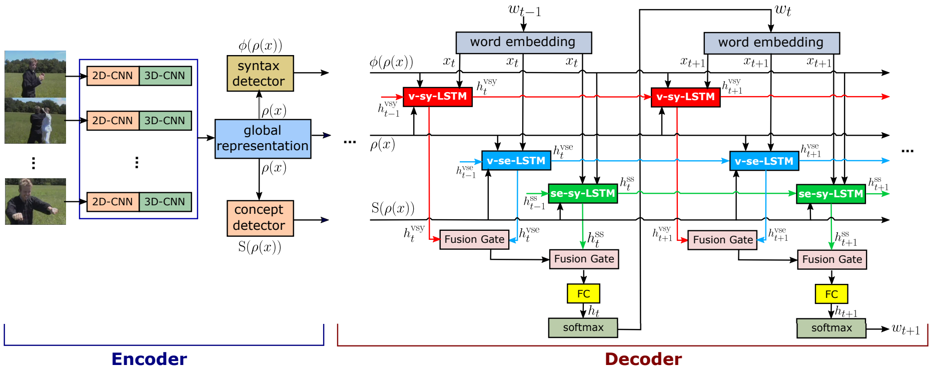

Figure 2. Proposed video captioning model. Firstly, we extract 2D-CNN and 3D-CNN visual features and a global representation $\rho(x)$. Next, the method predicts semantic and syntactic representations of the video by $\mathrm{S}(\rho(x))$ and $\phi(\rho(x))$, respectively. Then, the decoder generates the t-th word combining these three vectors in pairs. A different VNC$_L$ layer processes each pair.

semantic-syntactic). Unlike other architectures [2, 11, 17], our recurrent layers are not deeply connected (the output of one is not the input of another). Intuitively, each one is in charge of combining two different information channels separately. So, we compute the three layers in parallel without increasing the execution time.

For more details about the definition of each component of the decoder we we refer the reader to Perez-Martin *et al.* [21].

1.3. Syntax-weighted Loss

As a language generation task, video description models are usually trained by the Cross-Entropy minimization (CELoss) [12]. However, the weak relationship of the CELoss function with the popular evaluation metrics of video captioning [19, 22], constitutes a limitation of its use.

To overcome this limitation while aiming to consider syntactic information in the training phase, we propose the syntax-weighted loss function. Our function improves the loss used by Chen *et al.* [3], considering the distance between the syntactic representation and the POS structure of the generated description. Given a video x, the ground-truth caption $y = (y_1, y_2, \ldots, y_L)$ of x, and the POS tagging t of the generated description, we define the weight

$$w = \max\left\{1, L^{\beta} - \big(\mathrm{dist}\big(\phi(\rho(x)), \omega(t)\big) + 1\big)^{\gamma}\right\},$$

and we minimize

$$-\frac{1}{w}\sum_{i=1}^{L} \log p_{\theta}(y_i | y_{z<i}), \tag{1}$$

where $\beta \in [0, 1]$ and $\gamma \in [0, 1]$ are hyperparameters used to manage the balance between the length (conciseness) and syntactic correctness of generated descriptions. Greater β implies longer captions, and greater γ implies better syntax.

2. Experiments

For the *visual-syntactic embedding*, we set the dimension of the common space to 512. We pre-train the embedding on the MSR-VTT dataset using the *cosine distance*, a learning rate of 1×10^{-5} and a margin parameter of 0.1.

To extract 2D-CNN features of the video, we use ResNet-152 [13] feature extractor pre-trained on ImageNet [4, 23]. For 3D-CNN, we use ECO [25] and R(2+1)D [24] feature extractors, both pre-trained on Kinetics-400.

Table 1 shows the results we obtained in each run on the VTT20 test set. We submit four runs as following:

Run 1 Training during 40 epochs on 80% of VTT20 train set and validating on the other 20%.

Run 2 and 4 Regularizing, these runs train during 29 and 46 epochs on the full VTT20 train set (without validation).

Run 3 Using VATEX dataset and the 80% of VTT20 as our training set. We trained for three epochs only and we validated on the other 20%.

Table 1. Performance comparison of our runs on the TRECVID VTT 2020 test set.

Run	Training Dataset	Validation Dataset	epochs	BLEU-4	METEOR	CIDEr	CIDEr-D	Spice
1	MSRVTT+VTT20(80%)	VTT20(20%)	40	**0.0115**	0.2105	0.125	0.060	0.057
2	MSRVTT+VTT20	-	29	0.0113	**0.2187**	**0.136**	**0.065**	**0.060**
3	MSRVTT+VTT20(80%)+VATEX	VTT20(20%)	3	0.0075	0.1938	0.087	0.047	0.040
4	MSRVTT+VTT20	-	46	0.0105	0.2071	0.124	0.062	0.055

3. Conclusions

In this paper, we presented an encoder-decoder model for video description capable of generating sentences with more precise semantics and syntax. As part of this model, we proposed a technique to retrieve POS tagging structures of video descriptions while obtaining a high-level syntactic representation from visual information. We show that paying more attention to syntax improves the quality of descriptions. There are several promising areas that we consider for future work, such as improving visual-syntactic embedding by learning to relate syntactic information to a graph-based representation of visual content.

References

[1] George Awad, Asad A. Butt, Keith Curtis, Yooyoung Lee, Jonathan Fiscus, Afzal Godil, Andrew Delgado, Jesse Zhang, Eliot Godard, Lukas Diduch, Jeffrey Liu, Alan F. Smeaton, Yvette Graham, Gareth J. F. Jones, Wessel Kraaij, and Georges Quénot. TRECVID 2020: comprehensive campaign for evaluating video retrieval tasks across multiple application domains. In *Proceedings of TRECVID 2020*. NIST, US, 2020.

[2] Lorenzo Baraldi, Costantino Grana, and Rita Cucchiara. Hierarchical Boundary-Aware Neural Encoder for Video Captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3185–3194. IEEE, 7 2017.

[3] Haoran Chen, Ke Lin, Alexander Maye, Jianming Li, and Xiaolin Hu. A Semantics-Assisted Video Captioning Model Trained with Scheduled Sampling. 8 2019.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, Miami, FL, US, 2009. IEEE.

[5] Jianfeng Dong, Xirong Li, and Cees G. M. Snoek. Predicting Visual Features From Text for Image and Video Caption Retrieval. *IEEE Transactions on Multimedia*, 20(12):3377–3388, 12 2018.

[6] Jianfeng Dong, Xirong Li, Chaoxi Xu, Shouling Ji, Yuan He, Gang Yang, and Xun Wang. Dual Encoding for Zero-Example Video Retrieval. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9338–9347. IEEE, 6 2019.

[7] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. VSE++: Improving Visual-Semantic Embeddings with Hard Negatives. In *British Machine Vision Conference (BMVC)*, 2018.

[8] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K. Srivastava, Li Deng, Piotr Dollar, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig. From captions to visual concepts and back. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 07-12-June, pages 1473–1482. IEEE, 6 2015.

[9] Zhe Gan, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, and Li Deng. Semantic Compositional Networks for Visual Captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2017-Janua, pages 1141–1150. IEEE, 7 2017.

[10] Lianli Gao, Zhao Guo, Hanwang Zhang, Xing Xu, and Heng Tao Shen. Video Captioning with Attention-Based LSTM and Semantic Consistency. *IEEE Transactions on Multimedia*, 19(9), 2017.

[11] Lianli Gao, Xiangpeng Li, Jingkuan Song, and Heng Tao Shen. Hierarchical LSTMs with Adaptive Attention for Visual Captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–19, 1 2019.

[12] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. 2016.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2016-Decem, pages 770–778. IEEE, 6 2016.

[14] Sheng Liu, Zhou Ren, and Junsong Yuan. SibNet: Sibling convolutional encoder for video captioning. In *MM 2018 - Proceedings of the 2018 ACM Multimedia Conference*, pages 1425–1434. Association for Computing Machinery, Inc, 10 2018.

[15] Antoine Miech, Ivan Laptev, and Josef Sivic. Learning a Text-Video Embedding from Incomplete and Heterogeneous Data, 4 2018.

[16] Niluthpol C. Mithun, Juncheng Li, Florian Metze, and Amit K. Roy-Chowdhury. Joint embeddings with multimodal cues for video-text retrieval. *International Journal of Multimedia Information Retrieval*, 8(1):3–18, 3 2019.

[17] Pingbo Pan, Zhongwen Xu, Yi Yang, Fei Wu, and Yueting Zhuang. Hierarchical Recurrent Neural Encoder for Video Representation with Application to Captioning. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1029–1038. IEEE, 6 2016.

[18] Yingwei Pan, Tao Mei, Ting Yao, Houqiang Li, and Yong Rui. Jointly Modeling Embedding and Translation to Bridge Video and Language. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4594–4602. IEEE, 6 2016.

[19] Ramakanth Pasunuru and Mohit Bansal. Reinforced Video Captioning with Entailment Rewards. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 979–985, Stroudsburg, PA, USA, 2017. Association for Computational Linguistics.

[20] Jesus Perez-Martin, Benjamin Bustos, and Jorge Pérez. Attentive Visual Semantic Specialized Network for Video Captioning. In *International Conference on Computer Vision*, 2020.

[21] Jesus Perez-Martin, Jorge Pérez, and Benjamin Bustos. Improving Video Captioning with Temporal Composition of a Visual-Syntactic Embedding. In *Winter Conference on Applications of Computer Vision*, 2021.

[22] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence Level Training with Recurrent Neural Networks. In *International Conference on Learning Representations*, 11 2016.

[23] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 12 2015.

[24] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A Closer Look at Spatiotemporal Convolutions for Action Recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6450–6459. IEEE, 6 2018.

[25] Mohammadreza Zolfaghari, Kamaljeet Singh, and Thomas Brox. ECO: Efficient Convolutional Network for Online Video Understanding. In *Computer Vision – ECCV 2018*, pages 713–730. Springer International Publishing, 2018.