Visual-Syntactic Embedding for Video Captioning

Jesus Perez-Martin, Jorge Pérez and Benjamin Bustos
Department of Computer Science, University of Chile
{jeperez, jperez, bebustos}@dcc.uchile.cl

Abstract

Video captioning is the task of predicting a semantic and syntactically correct sequence of words given some context video. The most successful methods for video captioning 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 address this limitation by considering 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. We tested our proposed architecture obtaining state-of-the-art results on two widely used video captioning datasets. This is a short version of a paper recently published at a Computer Vision Conference. The complete reference has been redacted to fulfill the double-blind restriction.*

1. Introduction

In this short paper we describe our SemSynAN architecture which is an encoder-decoder model for vide captioning that, besides considering visual and semantic features, incorporates in the decoder phase, a visual-syntactic representation extracted from the input video (see Figure 1). The three types of representations (visual, semantic, and syntactic) are combined with what we call *var-norm-compositional LSTM* and *adaptive fusion gates* that decide when and how to include each feature type in the token generation phase. Specifically, the main contributions of this approach are as follows:

1. 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, both into a common vector space (see Figure 2). 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 effective use of a ranking model to obtain syntactic representations of videos.

2. We propose the *Visual-Semantic-Syntactic Aligned Network* (SemSynAN) for video captioning that integrates global semantic and syntactic representations of the input video. It learns how to combine visual, semantic, and syntactic information in pairs (*i.e.*, visual-semantic, visual-syntactic, and semantic-syntactic) while generating output tokens. As our results show, this process produces more accurate descriptions, both semantically and syntactically.

3. We evaluate our method on two widely used datasets:

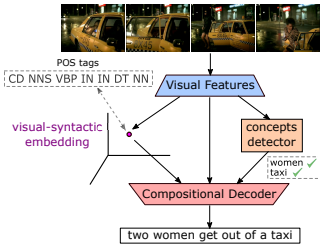

Figure 1. Example of video caption generation with Visual-Syntactic Embedding. The method computes high-level semantic and syntactic representations from the visual representation of the video. Next, the decoder generates a sentence from them.

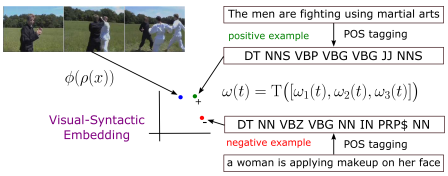

Figure 2. Visual-Syntactic Embedding. The model learns to map from video and POS sequences to a common space by the functions $\phi(\cdot)$ and $\omega(\cdot)$, preserving the relationship between visual content and positive syntactic structures.

the Microsoft Video Description (MSVD) [11] and the Microsoft Research Video-to-Text (MSR-VTT) dataset [24]. We improve the state-of-the-art in both datasets in all metrics except for one metric in MSR-VTT. For instance, in MSVD, we obtain a relative improvement of 10.8% for METEOR and 8.2% for CIDEr, and in MSR-VTT, a relative improvement of 2.6% for BLEU-4 and of 1.7% for METEOR.

For the sake of space we now present the model's performance over to widely used datasets. We refer the reader to Perez-Martin *et al*. [19] for more details about the approach.

2. Experiments and Results

Training Setup: To extract 2D-CNN features of the video, we use ResNet-152 [8] feature extractor pre-trained on ImageNet [4, 20]. For 3D-CNN, we use ECO [27] and R(2+1)D [21] feature extractors, both pre-trained on the Kinetics-400 dataset. On details, for frame-level representations, we concatenate the ResNet-152 and ECO features vectors, resulting in 3584-dimensional feature vectors. Concerning global representation, we average these features and concatenate it with the 512-dimensional R(2+1)D feature, obtaining a 4096-dimension global representation. To represent text descriptions, we obtain the vocabulary from the training set of each dataset. Next, we map each description to a sequence of vocabulary indices, putting the $\langle$eos$\rangle$ and $\langle$unk$\rangle$ tokens at the end and unknown words positions.

For the visual-syntactic embedding, we set the dimension common space dimension 512, the *visual model*'s hidden sizes to 2048 and 1024, and the *syntactic model*'s hidden size to 1024. We trained the model on the MSR-VTT dataset using the *cosine distance* as the dist$(\cdot,\cdot)$ function, a learning rate of 1×10^{-5}, and a margin parameter of 0.1. Some methods like LSTM-E [15] use all ground-truth captions, while others like LJRV [14] and Dong *et al*. [5] randomly sample five ground-truth captions per video. We follow the latter strategy. Our results demonstrate that, in MSVD and MSR-VTT, five samples are sufficient for learning the cues about video captions' syntactic structure.

We use Adam optimizer with an initial learning rate of 4×10^{-5} for the MSR-VTT dataset and 2×10^{-5} for MSVD and a batch-size of 64. We trained for at least 50 epochs with early-stopping criteria of 10 epochs. Each VNC$_L$ layer has a hidden size of 1024, and we use a keep probability of 0.8 for their dropout masks and 0.5 in all other cases. We fine-tune the hyperparameters on the validation sets and select the best checkpoint for testing according to a linear combination of BLEU-4, METEOR, CIDEr, and ROUGE$_L$ measures. We implemented our method on PyTorch [18], and it is publicly available on GitHub[1].

Table 1. Ablation study on the testing set of MSVD. Each row reports the results by changing only one aspect of the method, *e.g*., *w/o (v-se, se-sy)* omits the v-se-LSTM and se-sy-LSTM layers.

Architecture	BLEU-4	METEOR	CIDEr	ROUGE$_L$
SemSynAN (ours)	**64.4**	**41.9**	**111.5**	**79.5**
w/o v-sy	59.4	39.4	107.2	77.0
w/o se-sy	58.3	39.5	106.8	76.3
w/o (v-se, se-sy)	48.5	34.3	75.8	72.1
w/o (v-se, v-sy)	56.7	37.2	92.0	74.9
w/o hfg	60.8	41.3	103.9	75.9
w/o R(2+1)D	61.9	39.7	105.4	77.0
w/o wl	50.5	39.8	98.1	73.5
w/o vd	49.2	39.4	101.4	74.0
w/o max	63.7	41.2	108.1	79.0

Ablation Study: Table 1 shows the results of nine ablated experiments that we performed on the MSVD dataset. Specifically, we evaluate our SemSynAN model by removing, in separate runs, one or two of our VNC$_L$ layers, the *fusion gates*, the *weighted-loss* function, the dropout masks, and the maximum sampling strategy. We refer the reader to Perez-Martin *et al*. [19] for details about each ablated experiment.

The first four rows of Table 1 demonstrate that our model is significantly enhanced by including the syntactic information on MSVD dataset, proving the proposed method's effectiveness. Overall, the performance of our model is improved with the incorporation of each component.

Comparison with State of the Art on MSVD: Table 2 shows the proposed approach's performance and other state-of-the-art methods on the MSVD dataset. The SCN-LSTM [6] and SAVCSS [2] methods process a semantic representation by visual-semantic compositional LSTM decoders without considering the syntactic information. The incorporation of syntactic representation with our compositional modules improves the performance in comparison to those approaches. Likewise, the superior performance of our sentence generator framework is demonstrated in comparison to models that exploit fixed encoding based on 2D-CNN and 3D-CNN features, such as LSTM-E, SCN-LSTM, SAVCSS. Two recent approaches [10, 22] use the syntactic information from the POS tagging structure but

[1]Our code is publicly available at https://github.com/jssprz/visual_syntactic_embedding_video_captioning

Table 2. Performance comparison with the state-of-the-art methods on the testing set of MSVD dataset.

Approach	BLEU-4	METEOR	CIDEr	ROUGE$_L$
LSTM-E [15]	45.3	31.0	-	-
SCN-LSTM [6]	51.1	33.5	77.7	-
TDDF [25]	45.8	33.3	73.0	69.7
MTVC [16]	54.5	36.0	92.4	72.8
BAE [1]	42.5	32.4	63.5	-
MFATT-TM-SP [13]	52.0	33.5	-	-
ECO [27]	53.5	35.0	85.8	-
SibNet [12]	54.2	34.8	88.2	71.7
Joint-VisualPOS [10]	52.8	36.1	87.8	71.5
GFN-POS_RL(IR+M) [22]	53.9	34.9	91.0	72.1
hLSTMat [7]	54.3	33.9	73.8	-
SAVCSS [2]	61.8	37.8	103.0	76.8
DSD-3 DS-SEM [9]	50.1	34.7	76.0	73.1
ORG-TRL [26]	54.3	36.4	95.2	73.9
SemSynAN (ours)	**64.4**	**41.9**	**111.5**	**79.5**

Table 3. Performance comparison with the state-of-the-art methods on the testing set of MSR-VTT dataset. * denotes results that were obtained by reinforcement learning of that metric.

Approach	BLEU-4	METEOR	CIDEr	ROUGE$_L$
TDDF [25]	37.3	27.8	43.8	59.2
MTVC [16]	40.8	28.8	47.1	60.2
CIDEnt_RL [17]	40.5	28.4	51.7*	61.4
HRL [23]	41.3	28.7	48.8*	61.7
PickNet [3]	38.9	27.2	42.1	59.5
MFATT-TM-SP [13]	39.1	26.7	-	-
SibNet [12]	40.9	27.5	47.5	60.2
Joint-VisualPOS [10]	42.3	29.7	49.1	62.8
GFN-POS_RL(IR+M) [22]	41.3	28.7	**53.4***	62.1
hLSTMat [7]	39.7	27.0	43.4	-
SAVCSS [2]	43.8	28.9	51.4*	62.4
DSD-3 DS-SEM [9]	45.2	29.9	51.1	64.2
ORG-TRL [26]	43.6	28.8	50.9	62.1
SemSynAN (ours)	**46.4**	**30.4**	51.9	**64.7**

do not directly consider temporal relations between the visual, semantic, and syntactic representations. In the proposed approach, the semantic and syntactic representations are adaptively fused with the visual features, determining the most accurate information for generating each word. Hence, it is seen that our SemSynAN provides better scores than the previous syntax-based approaches. Specifically, our method has a relative BLEU-4 improvement of 4.2% $\left(\frac{64.4-61.8}{61.8}\right)$, METEOR of 10.8% $\left(\frac{41.9-37.8}{37.8}\right)$, CIDEr of 8.2% $\left(\frac{111.5-103.0}{103.0}\right)$, and ROUGE$_L$ of 3.5% $\left(\frac{79.5-76.8}{76.8}\right)$.

Comparison with State of the Art on MSR-VTT: Table 3 compares our SemSynAN model's performance with the recently published results on the MSR-VTT dataset. Our approach surpasses the methods that exploit the POS tagging structure of video captions [10, 22] and the approaches based on visual-semantic embeddings [12] and compositions [2, 6]. Unlike CIDEnt_RL [17], HRL [23], GFN-POS_RL(IR+M) [22], and SAVCSS [2], we do not use reinforcement learning to directly maximize any metric. However, our approach improves the results in terms of all metrics except CIDEr, where GFN-POS_RL(IR+M) [22] rich a better score by reinforcing this score. Specifically, our model has a relative BLEU-4 improvement of 2.6% $\left(\frac{46.4-45.2}{45.2}\right)$, METEOR of 1.7% $\left(\frac{30.4-29.9}{29.9}\right)$, and ROUGE$_L$ of 0.8% $\left(\frac{64.7-64.2}{64.2}\right)$. While, in terms of relative CIDEr, our approach outperforms the models without reinforcement learning by 1.6% $\left(\frac{51.9-51.1}{51.1}\right)$.

Qualitative Analysis: Figure 3 shows our model's predictions for three video examples of the MSVD dataset. To observe the improvement in the captions generated by our model, we compared these predictions with the outputs of two of our ablated models, *i.e.*, **w/o v-sy** and **w/o (v-se, se-sy)**. We highlighted some words and POS tags, where the model combined the semantic and syntactic information correctly. In these three examples, we can notice that our proposal generates better descriptions than the ablated models. In the first example, our approach generates the syntactic pattern "NN CC NN". In the second and third examples,

Figure 3. Three representative samples from the test split of MSVD, which cover ground-truth captions and their POS structure, two of our ablation models, and our proposal. Highlighted, the words and POS tags that the model predicted correctly.

different to the ablated models, our approach predicts the syntactic patterns "NN IN DT NN" and "NN IN PRP$ NN" respectively. In the last example, **w/o v-sy** and **w/o (v-se, se-sy)** fail to generate the noun "face".

3. Conclusions

In this paper, we presented an encoder-decoder model for video captioning named SemSynAN 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. Our method guarantees the contextual relation between the words in the sentence, controlling the semantic meaning and syntactic structure of generated captions. The experimental results demonstrate that our approach improves the state of the art on two of the most utilized evaluation benchmarks on video captioning.

References

[1] 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

[2] Haoran Chen, Ke Lin, Alexander Maye, Jianming Li, and Xiaolin Hu. A Semantics-Assisted Video Captioning Model Trained with Scheduled Sampling. *Frontiers in Robotic and AI*, 7, 8 2020. 2, 3

[3] Yangyu Chen, Shuhui Wang, Weigang Zhang, and Qingming Huang. Less Is More: Picking Informative Frames for Video Captioning. In *Computer Vision – ECCV 2018*, pages 367–384. Springer International Publishing, 2018. 3

[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. 2

[5] 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. 2

[6] 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. 2, 3

[7] 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. 3

[8] 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. 2

[9] M Hemalatha and C Chandra Sekhar. Domain-Specific Semantics Guided Approach to Video Captioning. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1587–1596, 3 2020. 3

[10] Jingyi Hou, Xinxiao Wu, Wentian Zhao, Jiebo Luo, and Yunde Jia. Joint Syntax Representation Learning and Visual Cue Translation for Video Captioning. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 2, 3

[11] David L. Chen and William 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

[12] Sheng Liu, Zhou Ren, and Junsong Yuan. SibNet: Sibling Convolutional Encoder for Video Captioning. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1425–1434, New York, NY, USA, 10 2018. ACM. 3

[13] Xiang Long, Chuang Gan, and Gerard De Melo. Video Captioning with Multi-Faceted Attention. In *Transactions of the Association for Computational Linguistics*, pages 173–184, 2018. 3

[14] Mayu Otani, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Naokazu Yokoya. Learning Joint Representations of Videos and Sentences with Web Image Search. In *Computer Vision – ECCV 2016*, pages 651–667. Springer International Publishing, 2016. 2

[15] 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. 2, 3

[16] Ramakanth Pasunuru and Mohit Bansal. Multi-Task Video Captioning with Video and Entailment Generation. In *55th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 1273–1283, Stroudsburg, PA, USA, 2017. Association for Computational Linguistics. 3

[17] 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. 3

[18] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary Devito Facebook, A I Research, Zeming Lin, Alban Desmaison, Luca Antiga, Orobix Srl, and Adam Lerer. Automatic differentiation in PyTorch. 2017. 2

[19] Jesus Perez-Martin, Benjamin Bustos, and Jorge Pérez. Improving Video Captioning with Temporal Composition of a Visual-Syntactic Embedding. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021. 2

[20] 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. 2

[21] 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. 2

[22] Bairui Wang, Lin Ma, Wei Zhang, Wenhao Jiang, Jingwen Wang, and Wei Liu. Controllable Video Captioning with POS Sequence Guidance Based on Gated Fusion Network. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 2, 3

[23] Xin Wang, Wenhu Chen, Jiawei Wu, Yuan-Fang Wang, and William Yang Wang. Video Captioning via Hierarchical Reinforcement Learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4213–4222. IEEE, 6 2018. 3

[24] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5288–5296, 2016. 2

[25] Xishan Zhang, Yongdong Zhang, Dongming Zhang, Jintao Li, and And Qi Tian. Task-Driven Dynamic Fusion: Reducing Ambiguity in Video Description. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6250–6258. IEEE, 2017. 3

[26] Ziqi Zhang, Yaya Shi, Chunfeng Yuan, Bing Li, Peijin Wang, Weiming Hu, and Zhengjun Zha. Object Relational Graph with Teacher-Recommended Learning for Video Captioning. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13278–13288, 2020. 3

[27] 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. 2, 3