Attentive Visual Semantic Specialized Network for Video Captioning

Millennium Institute Foundational Research on Data

dcc CIENCIAS DE LA COMPUTACIÓN UNIVERSIDAD DE CHILE

Jesus Perez-Martin, Benjamin Bustos, Jorge Pérez

Millennium Institute Foundational Research on Data, Department of Computer Science, University of Chile

ABSTRACT

We present the **Attentive Visual Semantic Specialized Network (AVSSN)** for video captioning, which is an encoder-decoder model based on our **Adaptive Attention Gate (AAG)** and **Specialized LSTM (S-LSTM)** layers. This architecture can selectively decide when to use visual or semantic information into the text generation process. The adaptive gate select the relevant information for providing a better temporal state representation than the existing decoders. Besides, the model is capable of learning to improve the expressiveness of generated captions attending to their length, using a caption-length-related loss function. We evaluate the effectiveness of the proposed approach on the **Microsoft Video Description (MSVD)** and **Microsoft Research Video-to-Text (MSR-VTT)** datasets, achieving state-of-the-art performance for popular evaluation metrics: BLEU-4, METEOR, CIDEr, and ROUGE$_L$

CONTACT

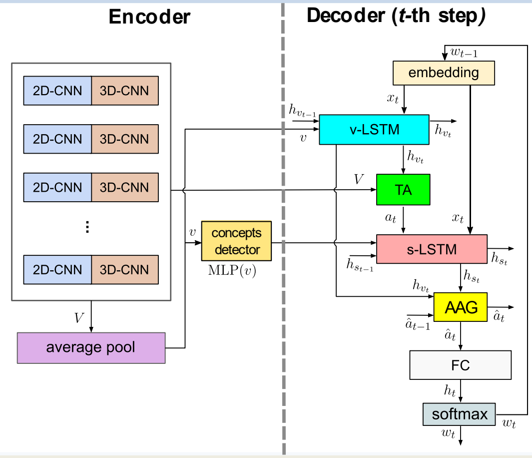
Jesús Pérez-Martín

Email: jeperez@dcc.uchile.cl

Twitter: @jes_prz

Website: https://users.dcc.uchile.cl/~jeperez

PROPOSED APPROACH

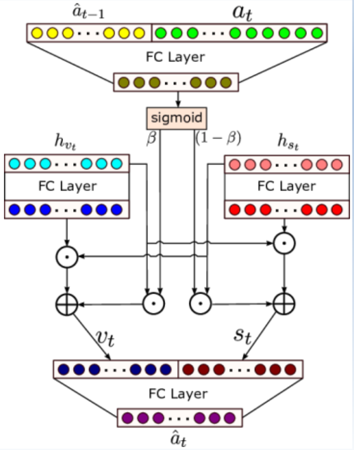


Visual-related Layer (v-LSTM): processes the visual information of the video at each step.

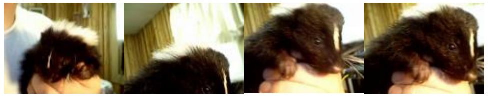
Semantic-related Layer (s-LSTM): processes the global semantic information, focusing on learning the visual and language context information.

Adaptive Attention Gate (AAG): determines the most accurate information to generate the word at each step. A cross-activation strategy with residual connections erges the related information within the specialized hidden states h_{v_t} and h_{v_t}.

$$\beta = \sigma(W_3 \cdot [\hat{a}_{t-1}, a_t] + b_2),$$
$$s_t = (h_{v_t} \cdot W_4) \odot h_{s_t} + \beta \odot h_{v_t},$$
$$v_t = (h_{s_t} \cdot W_5) \odot h_{v_t} + (1-\beta) \odot h_{s_t},$$
$$\hat{a}_t = W_6 \cdot [s_t, v_t] + b_3$$



QUALITATIVE RESULTS



GT 1: a man is holding a pet animal
GT 2: someone holds a baby skunk
wo. s-LSTM: a woman is petting a small animal
wo. AAG: a person is petting a small animal
wo. argmax: a person is petting a baby sloth
ours: a man is holding a small animal

GT 1: a little girl is pushing a stroller through a grocery store
GT 2: a kid pushes a stroller
wo. s-LSTM: a girl is dancing
wo. AAG: a bayby is playing
wo. argmax: a small child is walking
ours: a girl is pushing a stroller

GT 1: a man is seasoning some bacon
GT 2: a person seasons some meat
wo. s-LSTM: a woman is adding spices to a bowl of meat
wo. AAG: a man is seasoning some meat
wo. argmax: a man is sprinkling spices on a piece of meat
ours: a man is sprinkling spices on some bacon

LENGTH-RELATED LOSS FUNCTION

For the learning process, we operate with explicit supervision at the sequence level. We weight the standard CELoss by the length of the reference captions [3].
We minimize

$$\mathcal{L}_\theta = \frac{1}{L^\gamma} \sum_{t=1}^{L} \log p_\theta(w_t | w_{z<t})$$

QUANTITATIVE RESULTS

On **MSVD** [1], we surpass the SOTA by **3.7%** for METEOR, and **4.5%** for CIDEr

On **MSR-VTT** [2], we improve the SOTA by **2.7%**, **5.7%** and **2.3%** for BLEU4, METEOR and ROUGEL

Dataset	BLEU-4	METEOR	CIDEr	ROUGE_L
MSVD	62.3	39.2	107.7	78.3
MSR-VTT	45.5	31.4	50.6	64.3

CONCLUSIONS

- Learning to decide which of the visual and semantic representations is more important for predicting each word improves the quality of descriptions.
- Our adaptive attention gate (AAG) effectively determines the essential information to keep or disregard for generating the word in each step.
- Our method achieves state-of-the-art results on the MSVD and MSR-VTT datasets.

REFERENCES

1. Chen et *al.*, *"Collecting highly parallel data for paraphrase evaluation,"* in ACL, 2011

2. Xu et *al.*, *"MSR-VTT: A Large Video Description Dataset for Bridging Video and Language,"* in CVPR, 2016

3. Chen el *al.*, *"A Semantics-Assisted Video Captioning Model Trained with Scheduled Sampling,"* FRAI, 2020