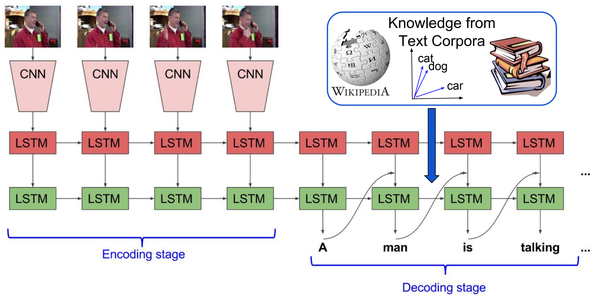
Improving LSTM-based Video Description with Linguistic Knowledge Mined from Text

Subhashini Venugopalan[1], Lisa Anne Hendricks[2], Raymond Mooney[1], Kate Saenko[3]
[1] UT-Austin [2] UC-Berkeley [3] UMass-Lowell

Goals

Incorporate linguistic knowledge mined from external text to improve Video Description. [1]



We propose to add linguistic knowledge from external text corpora to enhance the quality of LSTM-based video description networks.

Datasets

We evaluate our approach on a large, realistic collection of YouTube videos and movies.

(a) YouTube Video corpus

(b) MPII Movie Description Dataset

The YouTube dataset, collected by (Chen and Dolan, ACL 2011) consists of 1970 videos, where each video is accompanied by about 41 human descriptions (sentences), see (a) above. We also show results on large movie description corpora like the Montreal and MPII movie description datasets, see (b) above.

References

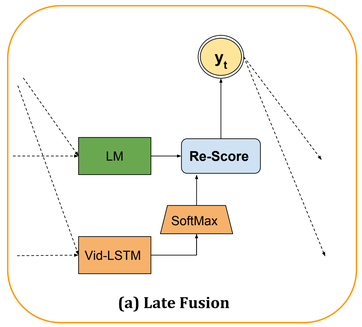
[1] S. Venugopalan, L. A. Hendricks, R. Mooney, and K. Saenko. Improving lstm-based video description with linguistic knowledge mined from text. *arXiv:1604.01729, EMNLP*, 2016.

[2] S. Venugopalan, L. A. Hendricks, M. Rohrbach, R. Mooney, T. Darrell, and K. Saenko. Captioning images with diverse objects. *arXiv:1606.07770*, 2016.

Techniques

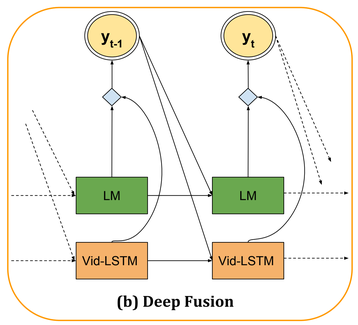
We propose multiple approaches to integrate language models trained on unannotated text corpora with existing video-captioning systems.

Our **early fusion** approach, simply initializes the weights of the video-captioning LSTM with an LSTM language model trained on large external text corpora.



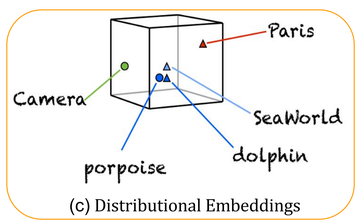
(a) Late Fusion

Our **late fusion** approach combines the language model and the video captioning model by re-scoring the softmax output of the video-to-text model.



(b) Deep Fusion

Our **deep fusion** model learns jointly from the hidden representations of the Language Model and S2VT video-to-text model (Vid-LSTM). So, the video description network is trained with the knowledge of the LSTM LM.

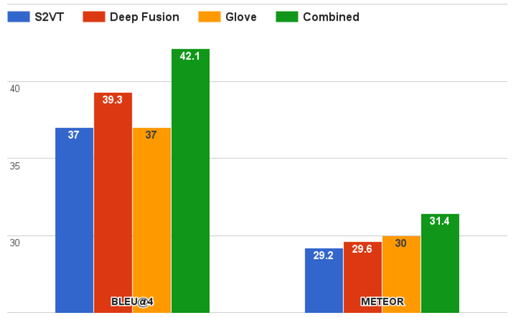


(c) Distributional Embeddings

Additionally, instead of the typical "one-hot" representations for words, we use distributional word embeddings. Embeddings such as `word2vec` and Glove are dense vector-space representations of words that capture fine-grained semantic and syntactic regularities

Evaluations

We evaluate our models on automatic machine translation metrics - METEOR and BLEU. We also obtain human judgements on 'Relevance' and 'Grammar'.



Evaluations using automated metrics on the Youtube video dataset.

Results - Youtube

Model	METEOR	B-4	Relevance	Grammar
S2VT	29.2	37.0	2.06	3.76
Early Fusion	29.6	37.6	-	-
Late Fusion	29.4	37.2	-	-
Deep Fusion	29.6	39.3	-	-
Glove	30.0	37.0	-	-
Glove+Deep				
- Web Corpus	30.3	38.1	2.12	4.05*
- In-Domain	30.3	38.8	2.21*	4.17*
Ensemble	**31.4**	**42.1**	**2.24***	**4.20***
Groundtruth			4.52	4.47

Youtube dataset: METEOR and BLEU@4 in %, and human ratings (1-5) on relevance and grammar. Best results in bold, * indicates significant over S2VT.

Results - Movies

Model	MPII-MD		M-VAD	
	METEOR	Grammar	METEOR	Grammar
S2VT†	6.5	2.6	6.6	2.2
Early Fusion	6.7	-	6.8	-
Late Fusion	6.5	-	6.7	-
Deep Fusion	6.8	-	6.8	-
Glove	6.7	3.9*	6.7	3.1*
Glove+Deep	6.8	**4.1***	6.7	**3.3***

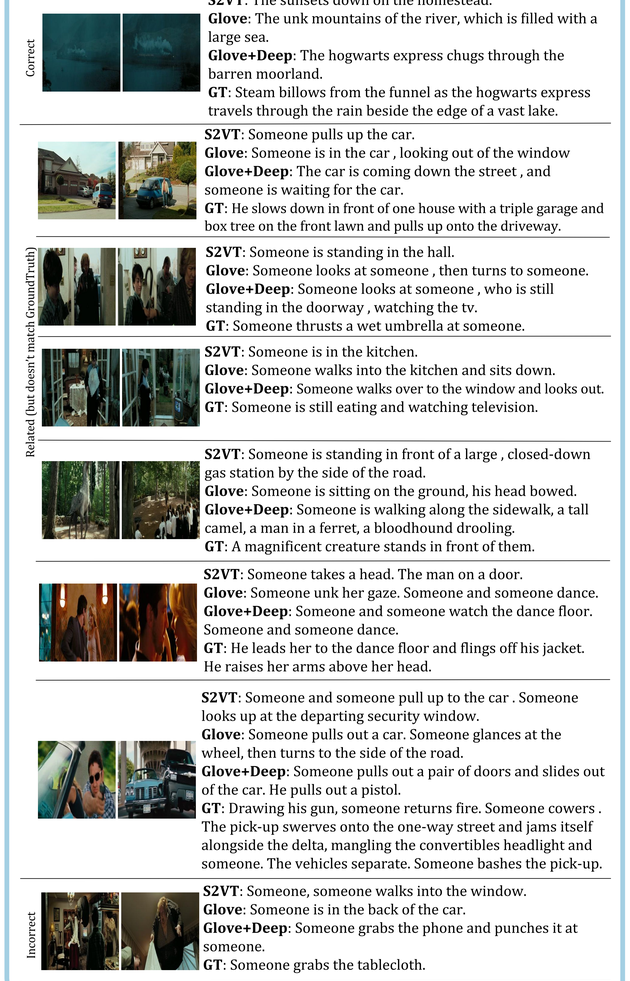
Movie dataset: METEOR in %, and human ratings (1-5) on grammar (higher is better). Best results in bold, * indicates significant over S2VT.

CODE

Project Page with Code and Examples
`http://vsubhashini.github.io/language_fusion.html`

Examples

Results on the Movie Description Corpora.



Future Directions

We also show that knowledge from external sources can be used to generate captions for obejcts without paired image-caption data. [2]