Sequence to Sequence – Video to Text

答辩人：肖静

老师：倪文龙

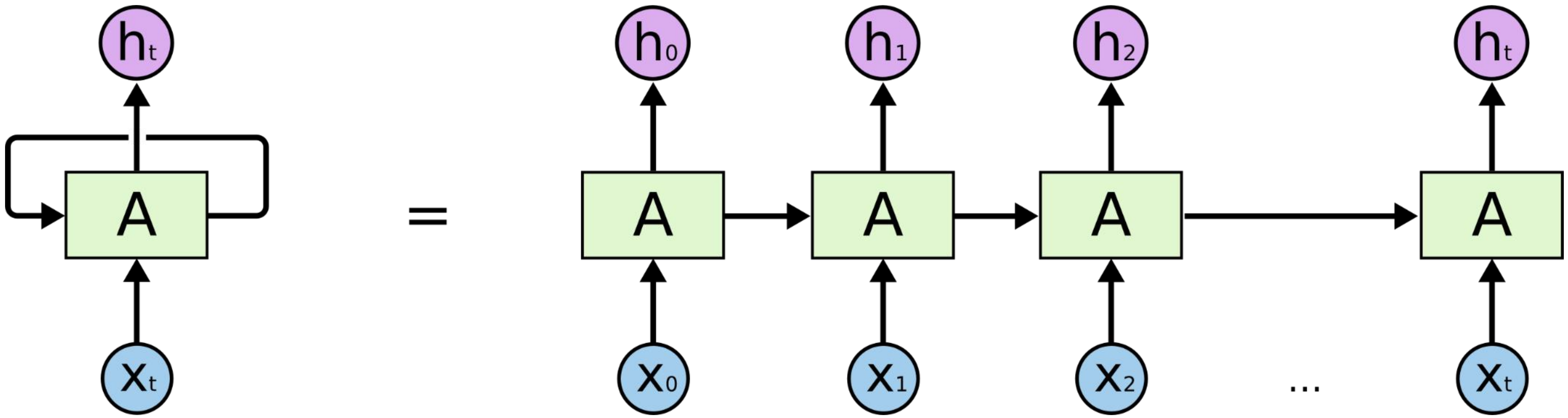
Content

PART 04

Experiment

Recurrent Neural Networks

- A recurrent neural network can be thought of as multiple copies of the same network, each passing a message to a successor.

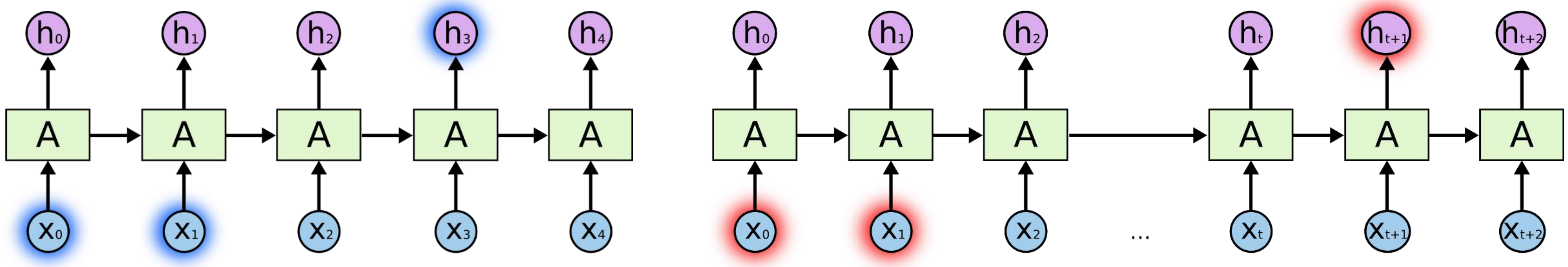


- if we unroll the loop:
  - This chain-like nature reveals that recurrent neural networks are intimately related to sequences and lists.

Experiment

The Problem of Long-Term Dependencies

- when the gap between the relevant information and the place that it's needed is small, RNNs can learn to use the past information.

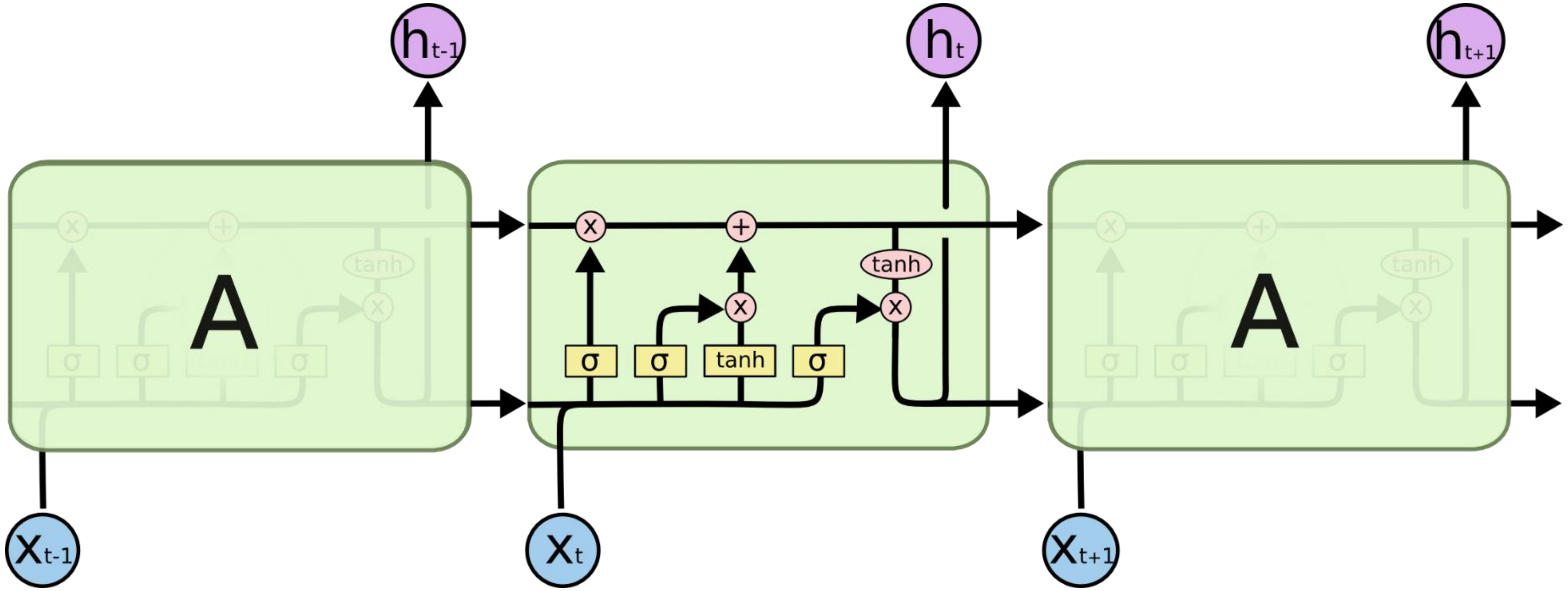


- Unfortunately, as that gap grows, RNNs become unable to learn to connect the information.

Experiment

LSTM Networks

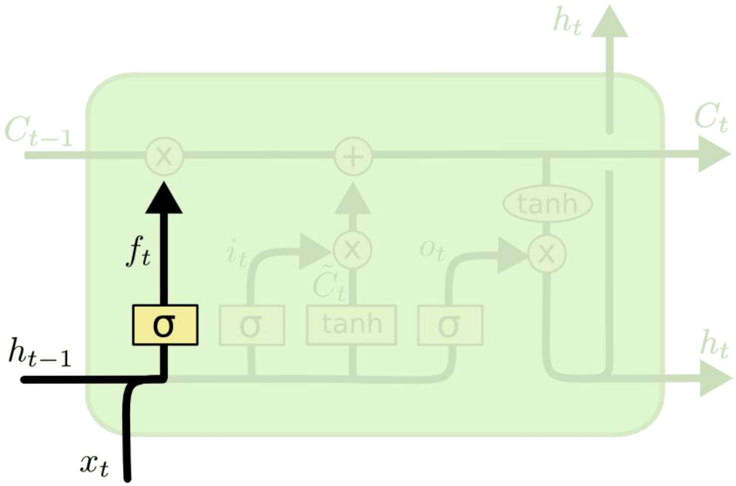
- Long Short Term Memory networks – usually just called "LSTMs" – are a special kind of RNN, capable of learning long-term dependencies.



Experiment

LSTM Networks

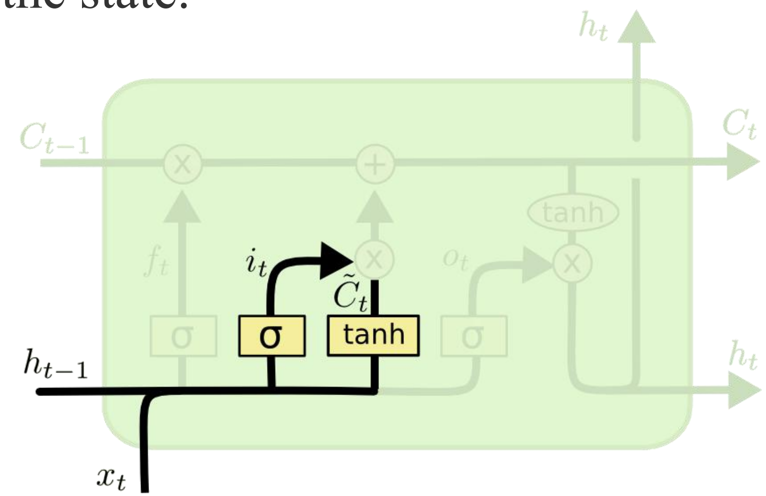
- The first step in our LSTM is to decide what information we're going to throw away from the cell state.
- This decision is made by a sigmoid layer called the "forget gate layer."



$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right)$$

LSTM Networks

- The next step is to decide what new information we're going to store in the cell state.
- This has two parts. we'll combine these two to create an update to the state.
  - First, a sigmoid layer called the "input gate layer" decides which values we'll update.
  - Next, a tanh layer creates a vector of new candidate values, $\tilde{C}_t$,that could be added to the state.

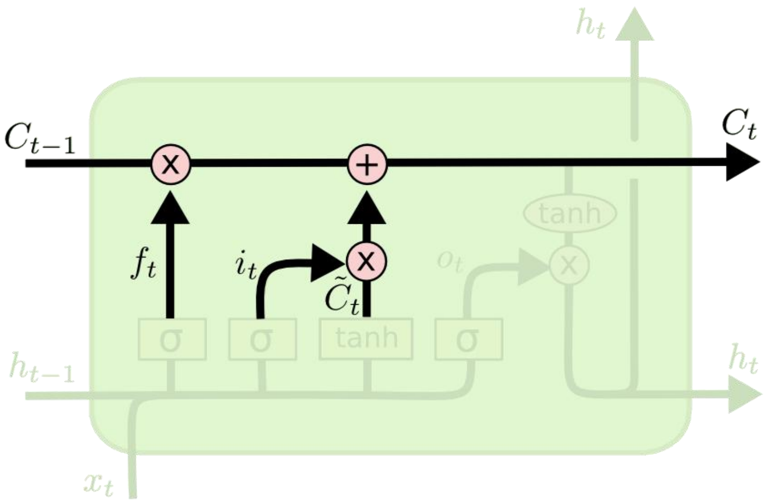


$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \ + \ b_i\right)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \ + \ b_C)$$

LSTM Networks

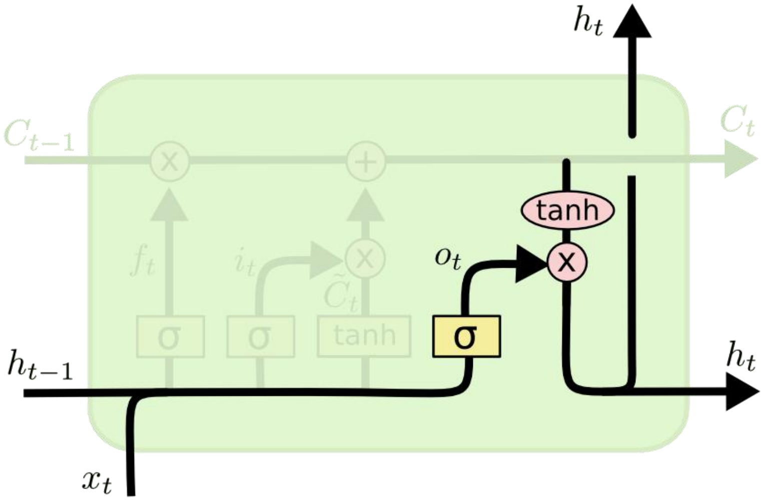
- It's now time to update the old cell state, C_{t-1}, into the new cell state. C_t.
- We multiply the old state by f_t, forgetting the things we decided to forget earlier. Then we add $i_t * \tilde{C}_t$. This is the new candidate values, scaled by how much we decided to update each state value.



$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

LSTM Networks

- Finally, we need to decide what we're going to output.
  - First, we run a sigmoid layer which decides what parts of the cell state we're going to output.
  - Then, we put the cell state through tanh and multiply it by the output of the sigmoid gate, so that we only output the parts we decided to.

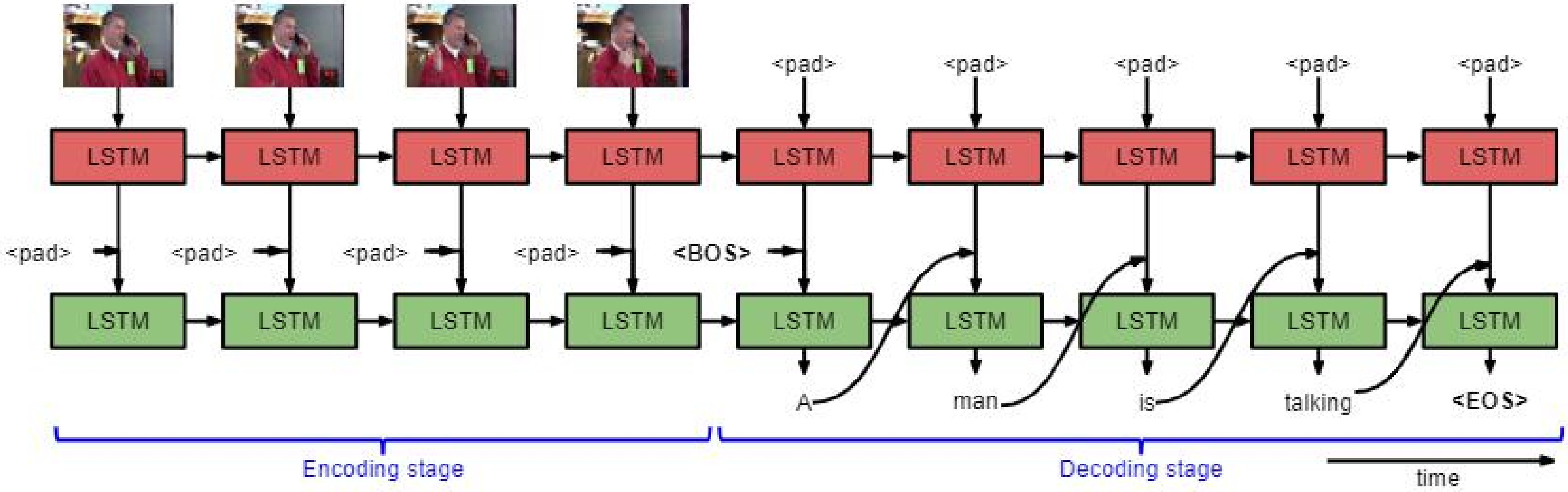


$$o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right)$$

$$h_t = o_t * \tanh\left(C_t\right)$$

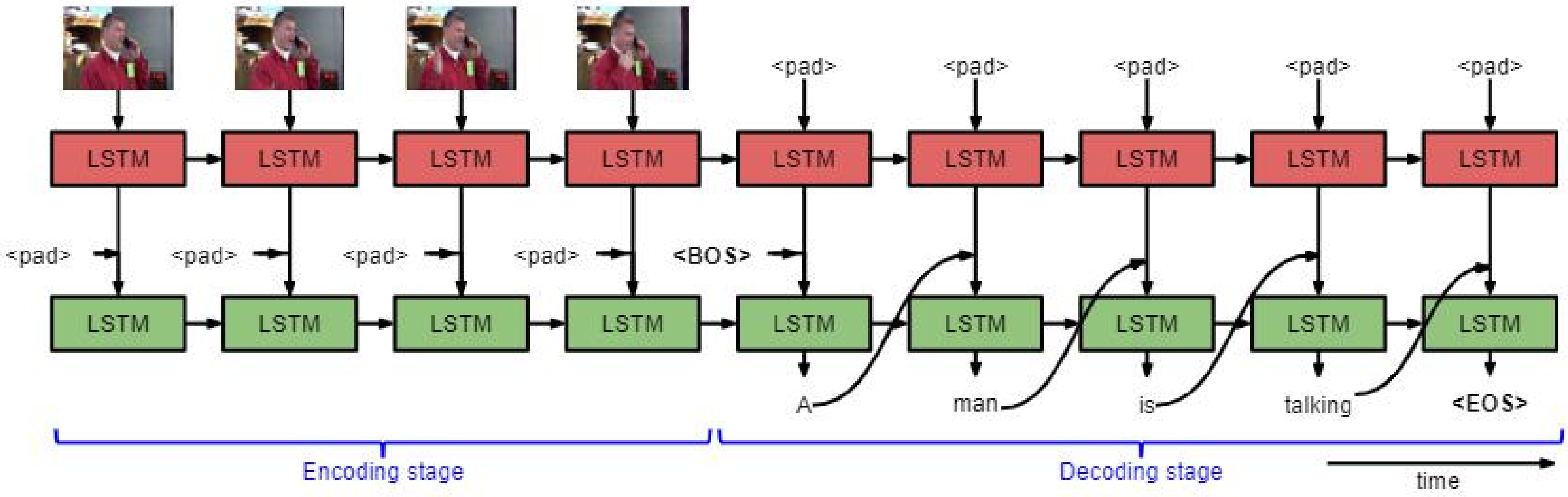
Experiment

S2VT



➢ In the first several time steps, the top LSTM layer (colored red) receives a sequence of frames and encodes them while the second LSTM layer receives the hidden representation (ht) and concatenates it with null padded input words (zeros), which it then encodes.

S2VT



- After all the frames in the video clip are exhausted, the second LSTM layer is fed the beginning-of-sentence (<BOS>) tag, which prompts it to start decoding its current hidden representation into a sequence of words.

Experiment

Video description datasets

- Microsoft Video Description Corpus (MSVD)
  - A collection of Youtube clips collected on Mechanical Turk by requesting workers to pick short clips depicting a single activity.
- MPII Movie Description Dataset (MPII-MD)
  - A collection contains around 68,000 video clips extracted from 94 Hollywood movies. Each clip is accompanied with a single sentence description.
- Montreal Video Annotation Dataset (M-VAD)
  - A collection of about 49,000 short video clips from 92 movies.

Experiment

Evaluation metrics And Related approaches

Model	METEOR	
FGM [36]	23.9	(1)
Mean pool		
- AlexNet [39]	26.9	(2)
- VGG	27.7	(3)
- AlexNet COCO pre-trained [39]	29.1	(4)
- GoogleNet [43]	28.7	(5)
Temporal attention		
- GoogleNet [43]	29.0	(6)
- GoogleNet + 3D-CNN [43]	29.6	(7)
S2VT (ours)		
- Flow (AlexNet)	24.3	(8)
- RGB (AlexNet)	27.9	(9)
- RGB (VGG) random frame order	28.2	(10)
- RGB (VGG)	29.2	(11)
- RGB (VGG) + Flow (AlexNet)	29.8	(12)

Table 2. MSVD dataset (METEOR in %, higher is better).

- Quantitative evaluation of the models are performed using the METEOR metric which was originally proposed to evaluate machine translation results.
- The METEOR score is computed based on the alignment between a given hypothesis sentence and a set of candidate reference sentences.
- METEOR compares exact token matches, stemmed tokens, paraphrase matches, as well as semantically similar matches using WordNet synonyms.

PART 05

Contribution

Contribution

- This is the first approach to video de-scription that uses a general sequence to sequence model. This allows our model to handle a variable number of input frames, learn and use the temporal structure of the video and learn a language model to generate natural, grammatical sentences.

- The model is learned jointly and end-to-end, incorporating both intensity and optical flow inputs, and does not require an explicit attention model. We demonstrate that S2VT achieves state-of-the-art performance on three diverse datasets, a standard YouTube corpus MSVD and the M-VAD and MPII Movie Description datasets.

PART 06

Conclusion

Conclusion

- This paper proposed a novel approach to video description. In contrast to related work, they construct descriptions using a sequence to sequence model.
  - Frames are first read sequentially and then words are generated sequentially.
  - This allows us to handle variable-length input and output while simultaneously modeling temporal structure.
- The model achieves state-of-the-art performance on the MSVD dataset, and outperforms related work on two large and challenging movie-description datasets.
- Despite its conceptual simplicity, the model significantly benefits from additional data, suggesting that it has a high model capacity.

Thank you !