Middle-Out Decoding

Shikib Mehri and Leonid Sigal

Overview

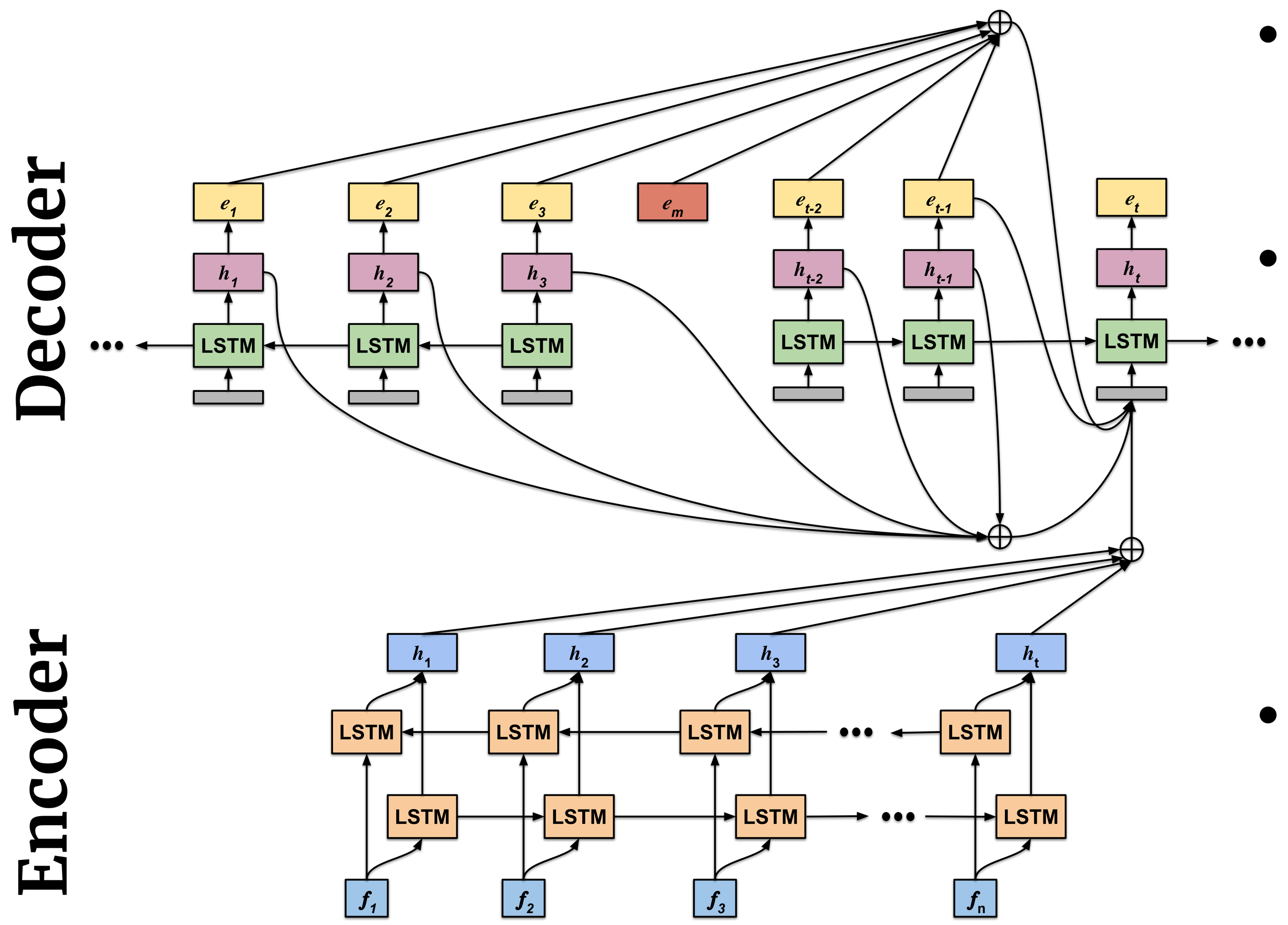
Motivation

- Left-to-right sequence generation results in earlier outputs having a **profound effect** on later outputs
- We speculate that this results in models that **lack diversity** and **cannot be effectively controlled**

Contribution

- We propose the ***middle-out decoder*** which begins from an initial middle-word and **simultaneously expands** the sequence in both directions
- To ensure the **consistency** and the **coherence** of the generated output sequence**,** we introduce a ***dual self-attention mechanism***
- We perform quantitative experimentation demonstrating **State-of-the-Art** results on **video captioning**
- Our analysis demonstrates interesting and valuable properties of the middle-out decoder, such as the ability to **effectively control** the sequence generation.

Middle-Out Decoder

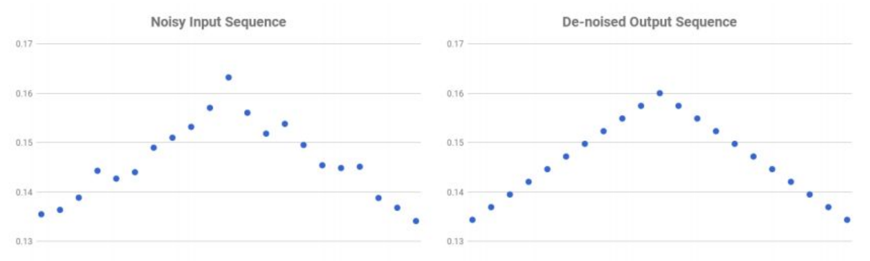


- Starting from an initial middle-word, e_m, our *middle-out decoder* generates the sequence **bidirectionally**
- Our *dual self-attention* ensures the coherence of the output by doing an attention over both
  - **embedded outputs** (Werlen et al., 2018)
  - **decoder hidden states** (Daniluk et al., 2017)
- Variable definitions
  - f is the frame feature vector
  - h is the LSTM hidden state
  - e is the embedded output words

Synthetic De-Noising

Dataset

We generate a synthetic dataset for the task of de-noising. We generate a symmetric sequence and add uniform noise to points.



Results

We evaluate a standard seq2seq network, as well as the middle-out decoder on this task -- demonstrating the overwhelming effectiveness of the latter.

Models	MSE	Symmetric MSE
Seq2Seq	1.52 x 1e-3	0.0780
Middle-Out	3.51 x 1e-4	1.32 x 1e-4

Video Captioning

Dataset

We utilize the **MSVD** (Youtube2Text) dataset (Chen and Dolan, 2011) which consists of 1970 youtube videos and captions.

Frame Features

The videos were sampled at *3fps* and passed through a pretrained Inception-v4 model (Szegedy et al., 2017), to obtain a 1536-dim feature vector for each frame.

Middle-Word

For the task of video captioning, we define the middle-word to be the **action verb** in the caption. We **train a classifier** to predict the middle-word (i.e., the action) given the video.

Models	METEOR	CIDEr-D	ROUGE-L	BLEU-4
LSTM + Visual Semantic Embeddings	31.0	-	-	45.3
Paragraph RNN	32.6	65.8	-	49.9
Hierarchical Recurrent Neural Encoder	33.9	-	-	46.7
Hierarchical LSTM + Adjusted Attention	33.6	-	-	**53.0**
Seq2Seq + Attention	34.0	78.9	70.0	47.4
Seq2Seq + Attention + Dual Self-Attention	**34.4**	**81.9**	**70.5**	**48.3**
Middle-Out + Attention	30.9	68.6	66.9	40.8
Middle-Out + Attention + Dual Self-Attention	**34.1**	**79.5**	**69.8**	**47.0**

Properties of the Middle-Out Decoder

Controllable

We evaluate the quality of the output when providing the **ground truth middle-word**. We compare to Grid Beam Search (Hokamp and Liu, 2017).

Models	METEOR	BLEU-4
Seq2Seq	34.5	47.4
Grid Beam Search	40.4	61.0
Middle-Out	**40.9**	**62.5**

We provide qualitative examples to demonstrate that it can **effectively utilize the middle-word** to attend to various parts of the input. We concatenate two videos together and try to generate the caption for either of the two videos.

	 *a magician is performing magic*	 *a man is playing piano*
Provided Middle-Word	performing	playing
Seq2Seq + Attention	*a man is playing the piano*	*a man is playing the piano*
Middle-Out + Self-Attention	*a man is performing on a stage*	*a man is playing the piano*

Dependance on Middle-Word

We simulate different **classification accuracies** to show that as the middle-word classification improves, decoder generates **progressively higher-quality** captions.

Accuracy	METEOR	CIDEr-D	ROUGE-L	BLEU-4
31.64%	34.1	79.5	69.8	47.0
50%	35.6	90.4	71.9	50.3
75%	38.4	105.6	75.1	56.2
100%	**40.9**	**124.4**	**78.6**	**62.5**

References

- Werlen, Lesly Miculicich, et al. "Self-Attentive Residual Decoder for Neural Machine Translation." *Proceedings of NAACL 2018.*
- Daniluk, Michał, et al. "Frustratingly short attention spans in neural language modeling." *Proceedings of ICLR 2017.*
- Chen, David L., and William B. Dolan. "Collecting highly parallel data for paraphrase evaluation." *Proceedings of ACL 2011.*
- Szegedy, Christian, et al. "Inception-v4, inception-resnet and the impact of residual connections on learning." *Proceedings of AAAI 2017.*
- Hokamp, Chris, and Qun Liu. "Lexically constrained decoding for sequence generation using grid beam search." *Proceedings of ACL 2017.*