Paper Reading: Review Networks for Caption Generation

Jie Hao

June 11, 2018

Introduction

- Review network: a novel extension of the encoder-decoder framework.
- The review network implement some review steps with attention mechanism on the encoder hidden states and outputs the **fact vectors**, which are more compact, abstractive, and global representation.
- Experiments: image captioning and source code captioning.

Previous: Attentive encoder-decoder models

Issues:

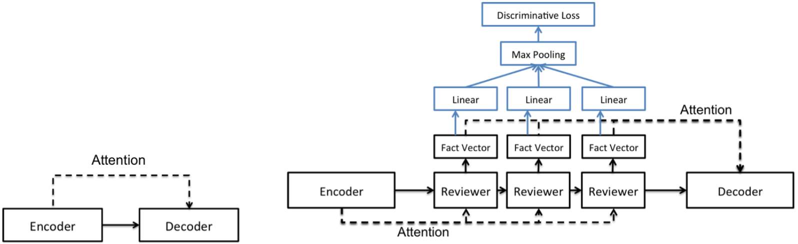
- Lacks global modeling abilities because attention mechanism proceeds in a sequential manner.

  For example, at the generation step t, the decoded token is conditioned on the attention results at current step $\tilde{h}_t$, but has no information about future attention results $\tilde{h}_{t'}$, $t^{'} > t$.

- Did not integrate discriminative supervision, which is beneficial for generative models.

Comparison between previous model and paper's model

Reviewer: $\mathbf{f}_t = g_t(H, \mathbf{f}_{t-1})$, where g_t is a modified LSTM unit with attention mechanism at review step t.

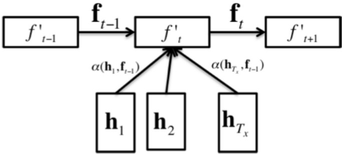


(a) Attentive Encoder-Decoder Model.

(b) Review Network. Blue components denote optional discriminative supervision. T_r is set to 3 in this example.

Figure 1: Model Architectures.

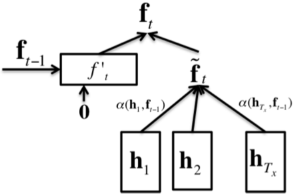
Attentive Input Reviewer



$$\tilde{\mathbf{f}} = att(H, \mathbf{f}_{t-1}) = \Sigma_{i=1}^{|H|} \frac{\alpha(\mathbf{h}_i,\mathbf{f}_{t-1})}{\Sigma_{i'}^{|H|}\alpha(\mathbf{h}_{i'},\mathbf{f}_{t-1})}\mathbf{h}_i \ , \ g_t(H, \mathbf{f}_{t-1}) = f'_t(\tilde{\mathbf{f}}_t, \mathbf{f}_{t-1}).$$

Here, α is a function that determines the weights for the hidden state. f'_t is an LSTM unit at step t.

Attentive Output Reviewer



$$\tilde{\mathbf{f}} = att(H, \mathbf{f}_{t-1}),\ g_t(H, \mathbf{f}_{t-1}) = f'_t(\mathbf{0}, \mathbf{f}_{t-1}) + \mathbf{W}\tilde{\mathbf{f}}_t$$

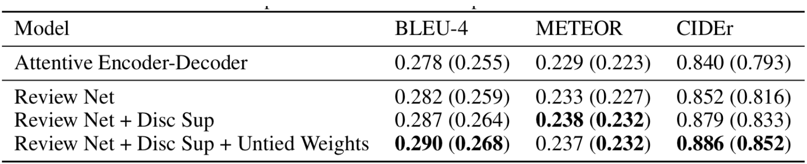
Discriminative Supervision

$$\mathcal{L}_d = \frac{1}{Z}\Sigma_{j \in W}\Sigma_{i \neq j} max(0, 1 - (s_j - s_i)).$$

Here s_i is the score of world i after the max pooling layer, and W is the set of all words that occur in **y**.

Experiments

MSCOCO dataset:



Model	BLEU-4	METEOR	CIDEr
Attentive Encoder-Decoder	0.278 (0.255)	0.229 (0.223)	0.840 (0.793)
Review Net	0.282 (0.259)	0.233 (0.227)	0.852 (0.816)
Review Net + Disc Sup	0.287 (0.264)	**0.238** (**0.232**)	0.879 (0.833)
Review Net + Disc Sup + Untied Weights	**0.290** (**0.268**)	0.237 (**0.232**)	**0.886** (**0.852**)

Inspiration for current work

- Use Review Network to improve the current Transformer model. Specifically, we use fact vectors as an auxiliary input to the decoder.
- Implement a discriminative loss to the system.