Paper Reading: Deconvolution-Based Global Decoding for Neural Machine Translation (COLING 2018)

Junyang Lin, Xu Sun, Xuancheng Ren, Shuming Ma, Jinsong Su, Qi Su

Peking University & Xiamen University

Introduction

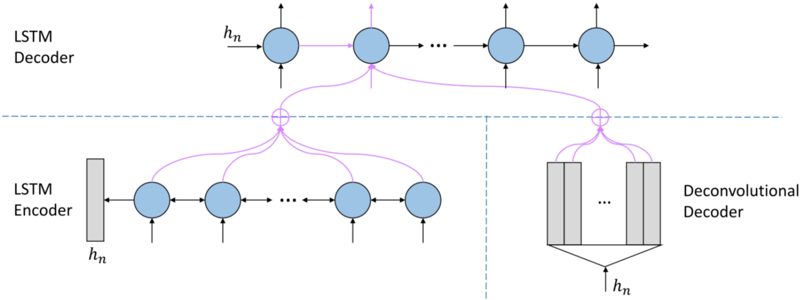
- A new NMT model that decodes the sequence with guidance of its structural prediction of the context of the target sequence.
- The model contains a deconvolution-based decoder to provide global information of the target-side contexts to the RNN decoder, so that the model is able to perform global decoding.
- Experiments:
  Chinese-English translation : Bleu: 2.82 $\uparrow$
  English-Vietnamese translation: Bleu: 1.54 $\uparrow$

Previous model problem:

- Conventional decoder translates words in a sequential order, the current generation is highly dependent on the previous generation and it is short of the knowledge about future generation.

  Briefly, translation is in need of the global information from the target-side context, but the decoding pattern of the conventional Seq2Seq model in NMT does not meet the requirement.

Model Architecture

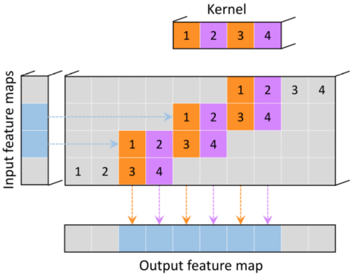


Model: Encoder

Bidirectional LSTM:

$$\overrightarrow{h_i} = LSTM(x_i, \overrightarrow{h_{i-1}}, C_{i-1})$$
$$\overleftarrow{h_i} = LSTM(x_i, \overleftarrow{h_{i-1}}, C_{i-1})$$

Deconvolution-Based Decoder



1d deconvolution on a input of size 2 with a kernel of size 4, a padding of 1 and, a stride of 2.

$\iff$ 1d convolution: i=6, k=4, s=2, p=2.

Deconvolution-Based Decoder

The goal of deconvolution operation is to generate a word embedding matrix $E \in R^{T*dim}$ where T refers to the sentence length designed for the output text sentence, which is a hyper-parameter.

At the l*th* layer, deconvolution generates a matrix $E_l \in R^{T_l*dim}$ where

$$T_l = T_{l-1} * s_l + k_l - 2 * p_l$$

RNN-based Decoder

Unidirectional LSTM:
At each time step, the decoder generates a word y_t by sampling from a conditional probability distribution of the target vocabulary P_{vocab} where,

$$
\begin{aligned}
P_{vocab} &= softmax(W_o v_t) \\
v_t &= g(s_t, c_t, \tilde{c}_t) \\
s_t &= LSTM(y_{t-1}, s_{t-1}, C_{t-1})
\end{aligned}
$$

where $g(\dot{})$ refers to non-linear activation function, c_t and $\widetilde{c}_t$ are the outputs of the attention mechanism.

Experiments

Dataset: NIST translation task for the Chinese-to-English translation.

Model	MT-03	MT-04	MT-05	MT-06	Ave.
Moses	32.43	34.14	31.47	30.81	32.21
RNNSearch	33.08	35.32	31.42	31.61	32.86
Lattice	34.32	36.50	32.40	32.77	34.00
Coverage	34.49	38.34	34.91	34.25	35.49
InterAtten	35.09	37.73	35.53	34.32	35.67
MemDec	36.16	**39.81**	35.91	35.98	36.97
Seq2Seq+Attention	35.32	37.25	33.52	33.54	34.91
+DeconvDec	**38.04**	39.75	**36.77**	**36.32**	**37.73**

Experiments

Dataset: IWLST2015 for the English-to-Vietnamese translation task.

Model	BLEU
RNNSearch-1	23.30
RNNSearch-2	26.10
LabelEmb	26.80
NPMT	27.69
Seq2Seq+Attention	26.93
+DeconvDec	**28.47**

Conclusion

- With deconvolution-based decoder, the model can effectively exploit the information for the inference of syntactic structure and semantic meaning in the translation.
- The model generates less repetitive translation an demonstrates higher robustness to the sentences of different lengths.
- Inspiration for our current work: this paper provides a new global representation method for the translation task.