Deliberation Networks: Sequence Generation Beyond One-Pass Decoding

Yingce Xia, Fei Tian, Lijun Wu, Jianxin Lin, Tao Qin, Nenghai Yu, Tie-Yan Liu

Lisa Kuhn

Recent Advances In Sequence-To-Sequence Learning

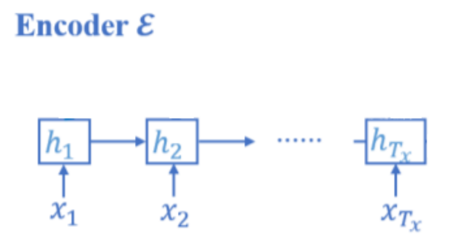
Introduction

- Deliberation: process of polishing (e.g. an essay) while looking at how a local element fits in its global environment
- Output of a neural network depends partly on what has been output already, not what will be output
- Extended encoder-decoder model with second decoder to do the deliberation process

Encoder

- Encodes input sequence x to hidden states $H = \{h_1, h_2, ..., h_{T_x}\}$

$$h_i = \mathrm{RNN}(x_i, h_{i-1})$$



First-pass Decoder

- Generates hidden states $\hat{s}$, first-pass sequence $\hat{y}$
- Attention Model:

$$ctx_e = \sum_{i=1}^{T_x} \alpha_i h_i$$

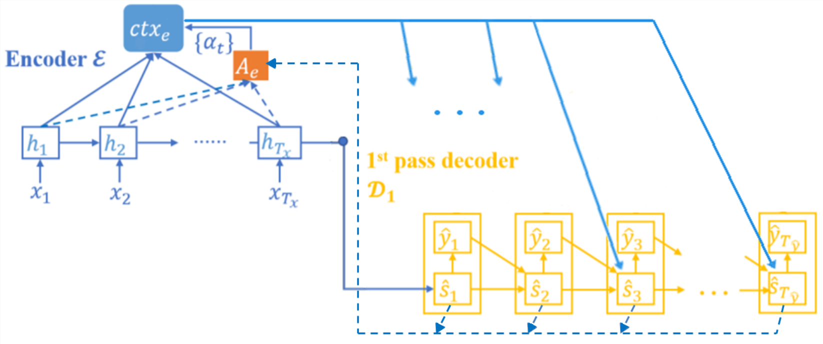
$$\alpha_i \propto \exp(\upsilon_\alpha^T \tanh(W_{att,h}^c h_i + W_{att,\hat{s}}^c \hat{s}_{j-1}))$$

- Hidden state calculation:

$$\hat{s}_j = \text{RNN}([\hat{y}_{j-i}; ctx_e], \hat{s}_{j-1})$$

- Apply affine transformation on $[\hat{s}_j; ctx_e; \hat{y}_{j-1}]$
- Softmax layer $\rightarrow$ sample out $\hat{y}_j$ from multinomial distribution

First-pass decoder



Second-pass Decoder

- Takes:
  - Previous hidden state s_{t-1}
  - Previously decoded word y_{t-1}
  - Source contextual information ctx'_e
  - First-pass contextual information ctx_c
- ctx'_e computed similarly to ctx_e, last hidden state of second-pass decoder is used (s_{t-1} instead of $\hat{s}_{j-1}$)

Second-pass Decoder

- Attention Model:

$$ctx_c = \sum_{j=1}^{T_{\hat{y}}} \beta_j [\hat{s}_j ; \hat{y}_j]$$

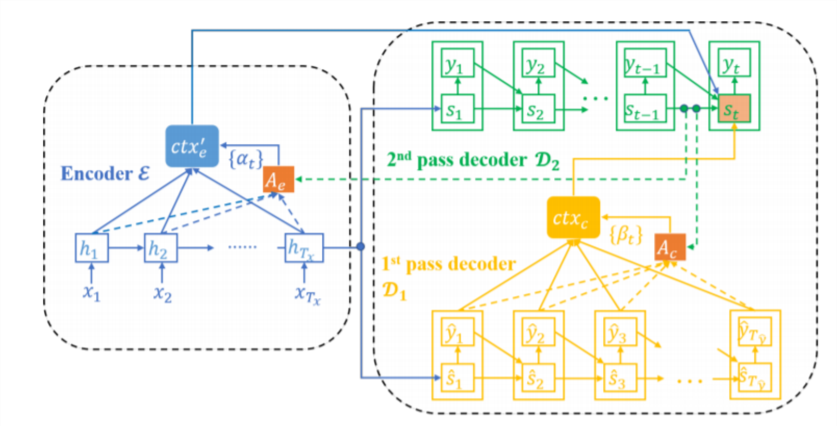
$$\beta_j \propto \exp(\upsilon_\beta^T \tanh(W^d_{att,\hat{s}\hat{y}} [\hat{s}_j ; \hat{y}_j] + W^d_{att,s\,t-1}))$$

- Hidden state calculation:

$$s_t = \mathrm{RNN}([y_{t-1}; ctx'_e; ctx_c], s_{t-1})$$

- To generate y_t further transform $[s_t; ctx'_e; ctx_c; y_{t-1}]$ similar to sampling of $\hat{y}_j$

Framework



Encoder $\mathcal{E}$
ctx'_e
$\{\alpha_t\}$
A_e
h_1
h_2
h_{T_x}
x_1
x_2
x_{T_x}
y_1
y_2
y_{t-1}
y_t
s_1
s_2
s_{t-1}
s_t
2nd pass decoder $\mathcal{D}_2$
ctx_c
$\{\beta_t\}$
A_c
1st pass decoder $\mathcal{D}_1$
$\hat{y}_1$
$\hat{y}_2$
$\hat{y}_3$
$\hat{y}_{T_{\hat{y}}}$
$\hat{s}_1$
$\hat{s}_2$
$\hat{s}_3$
$\hat{s}_{T_{\hat{y}}}$

Training

Training

- For this setting, data log likelihood specialized to

$$(1/n) \sum_{(x,y) \in D_{XY}} \mathcal{J}(x, y; \theta_e, \theta_1, \theta_2)$$

where

$$\mathcal{J}(x, y; \theta_e, \theta_1, \theta_2) = \log \sum_{y' \in \mathcal{Y}} P(y|y', E(x; \theta_e); \theta_2) P(y'|E(x; \theta_e); \theta_1)$$

- $\mathcal{Y}$ is the collection of all possible target sentences

Problem

- Derivation of $\mathcal{J}$ w.r.t θ_1:

$$\nabla_{\theta_1}\mathcal{J}(x, y; \theta_e, \theta_1, \theta_2) = \frac{\sum_{y' \in \mathcal{Y}} P(y|y', E(x; \theta_e); \theta_2)\nabla_{\theta_1} P(y'|E(x; \theta_e); \theta_1)}{\sum_{y' \in \mathcal{Y}} P(y|y', E(x; \theta_e); \theta_2) P(y'|E(x; \theta_e); \theta_1)}$$

- Computationally not feasible because of $\mathcal{Y}$
- Solution: Monte Carlo based method to optimize lower bound of $\mathcal{J}$ function $\rightarrow \tilde{\mathcal{J}}$

$$\tilde{\mathcal{J}}(x, y; \theta_e, \theta_1, \theta_2) = \sum_{y' \in \mathcal{Y}} P(y'|E(x; \theta_e); \theta_1) \log P(y|y', E(x; \theta_e); \theta_2)$$

Algorithm

Algorithm 1: Algorithm to train the deliberation network

Input: Training data corpus D_{XY}; minibatch size m; optimizer *Opt*$(\cdots)$ with gradients as input ;

while *models not converged* **do**

- Randomly sample a mini-batch of m sequence pairs $\{x^{(i)}, y^{(i)}\}\ \forall i \in [m]$ from D_{XY};
- For any $x^{(i)}$ where $i \in [m]$, sample $y'^{(i)}$ according to distribution $P(\cdot|E(x^{(i)};\theta_e);\theta_1)$;
- Perform parameter update: $\Theta \leftarrow \Theta + Opt(\frac{1}{m}\sum_{i=1}^{m} G(x^{(i)}, y^{(i)}, y'^{(i)}; \Theta))$.

Full Training Process

- Pre-train standard encoder-decoder based NMT models until convergence
- Deliberation network encoder initialized by encoder of standard model
- Both deliberation network decoders are initialized by the decoder of the pre-trained model
- Train deliberation network until convergence
- Use beam search to sample output by first decoder

Results

Neural Machine Translation

- English-to-French translation on WMT’14 and newstest datasets
- Chinese-to-English translation on n LDC corpus and NIST datasets
- Removed sentences with more than 50 words, limit source and target words
- Two models: shallow and deep model

Shallow Model

- Based on single-layer GRU model RNNSearchch
- Baselines:
    - Standard NMT model RNNSearch
    - Standard NMT model with two stacked decoding layers
    - Review Network

Shallow Model Results

Table 1: BLEU scores of En→Fr translation

Algorithm	$\mathcal{M}_{\text{base}}$	$\mathcal{M}_{\text{dec}\times 2}$	$\mathcal{M}_{\text{reviewer}\times 4}$	$\mathcal{M}_{\text{delib}}$
BLEU	29.97	30.40	30.76	**31.67**

Table 2: BLEU scores of Zh→En translation

Algorithm	NIST04	NIST05	NIST06	NIST08
$\mathcal{M}_{\text{base}}$	34.96	34.57	32.74	26.21
$\mathcal{M}_{\text{delib}}$	36.90	35.57	33.90	27.13

Translation Examples

Source Sentence

Aiji shuo, zhongdong heping xieyi yuqi jiang you yige xinde jiagou.

Reference Translation

Egypt says a new framework is expected to come into being for the Middle East peace agreement.

Translation Base Model

egypt's middle east peace agreement is expected to have a new framework, he said.

First-pass decoder output

egypt's middle east peace agreement is expected to have a new framework, egypt said.

Second-pass decoder Output

egypt says the middle east peace agreement is expected to have a new framework

Deep Model

- Deep LSTM model
- Only on En-Fr translation task
- Apply BPE techniques: split training sentences in sub-word units
- Restrict source and target sentence lengths within 64 subwords
- Encoder and Decoders are 4-layer LSTMs with residual connections

Deep Model Results

System	Configurations	BLEU
GNMT [31]	Stacked LSTM (8-layer encoder + 8 layer decoder) + RL finetune	39.92
FairSeq [4]	Convolution (15-layer) encoder and (15-layer) decoder	40.51
Transformer [26]	Self-Attention + 6-layer encoder + 6-layer decoder	41.0
this work	Stack LSTM (4-layer encoder and 4-layer decoder)	39.51
	Stack 4-layer NMT + Dual Learning	40.53
	Stack 4-layer NMT + Dual Learning + Deliberation Network	**41.50**

Text Summarization Results

Algorithm	ROUGE-1	ROUGE-2	ROUGE-L
$\mathcal{M}_{\text{base}}$	27.45	10.51	26.07
$\mathcal{M}_{\text{dec}\times 2}$	27.93	11.09	26.50
$\mathcal{M}_{\text{reviewer}\times 4}$	28.26	11.25	27.28
$\mathcal{M}_{\text{delib}}$	30.90	12.21	29.09

Questions