# Learning When to Concentrate or Divert Attention: Self-Adaptive Attention Temperature for Neural Machine Translation

Junyang Lin[1,2] Xu Sun[2] Xuancheng Ren[2] Muyu Li[2] Qi Su[1]

[1]School of Foreign Languages, Peking University
[2]MOE Key Laboratory of Computational Linguistics, Peking University
{linjunyang, xusun, renxc, limuyu0110, sukia}@pku.edu.cn

## Abstract

A new NMT model with self-adaptive attention temperature;

Attention varies at each time step based on the temperature;

Improved results on the benchmark datasets;

Analysis shows that temperatures vary when translating words of different types.

## Motivation

Focus on the source text should be different when translating words of various types, such as content word and function word;

Conventional attention mechanism uses the same computation for the decoding of each time step;

A learnable temperature can modify the softness of attention distribution.

## Sequence-to-Sequence as Baseline

**Encoder:** Bidirectional LSTM.;

**Decoder:** LSTM for sequential decoding. Training is with teacher forcing;

**Attention mechanism:** global attention for the relevant source-side information.

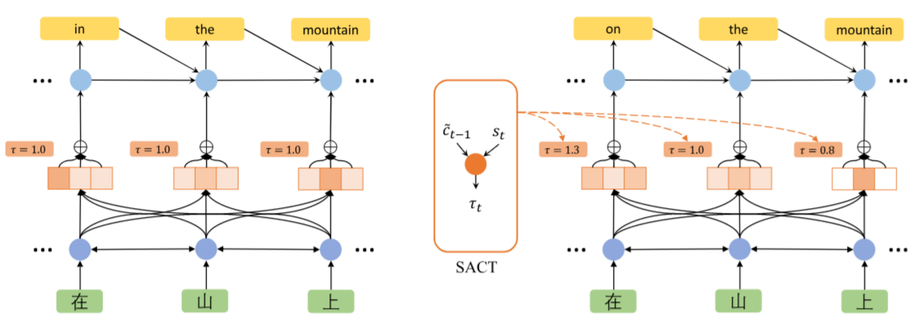
## Attention Temperature

Learn the attention temperature based on attention history and the current state, with a hyperparameter determining the bounds.

$$\tau_t = \lambda^{\beta_t}$$
$$\beta_t = \tanh(W_c \tilde{c}_{t-1} + U_s s_t)$$
$$\tilde{c}_t = \sum_{i=1}^{n} \tilde{\alpha}_{t,i} h_i$$
$$\tilde{\alpha}_{t,i} = \frac{exp(\tau_t^{-1} e_{t,i})}{\sum_{j=1}^{n} exp(\tau_t^{-1} e_{t,j})}$$

## Structure Comparison

Attention temperature adjusts the attention distribution automatically when translating different types of words.




(a) Example for the conventional attention mechanism with the fixed temperature.

(b) Example for the proposed attention mechanism with self-adaptive control of temperature.

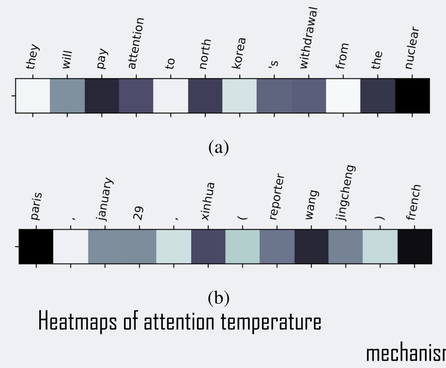
## Experiments

Chinese-to-English translation (LDC corpora).

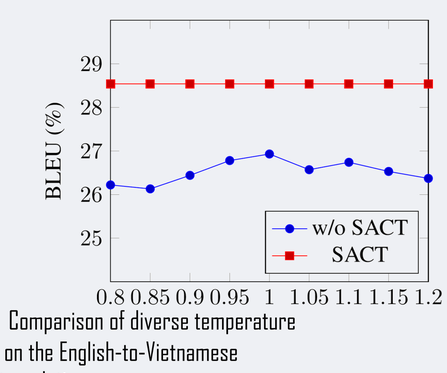
| Model | MT-03 | MT-04 | MT-05 | MT-06 | Ave. |
|---|---|---|---|---|---|
| Moses | 32.43 | 34.14 | 31.47 | 30.81 | 32.21 |
| RNNSearch | 33.08 | 35.32 | 31.42 | 31.61 | 32.86 |
| Coverage | 34.49 | 38.34 | 34.91 | 34.25 | 35.49 |
| MemDec | 36.16 | 39.81 | 35.91 | **35.98** | 36.97 |
| Seq2Seq | 35.32 | 37.25 | 33.52 | 33.54 | 34.91 |
| **+SACT** | **38.16** | **40.48** | **36.81** | 35.95 | **37.85** |

## Analysis

Attention heatmaps and comparison of fixed and self-adaptive temperature




Heatmaps of attention temperature




Comparison of diverse temperature mechanisms on the English-to-Vietnamese translation

## Examples

| |
|---|
| **Source:** 中国 大陆 手机 用户 成长 将 减缓 |
| **Gold:** growth of mobile phone users in mainland china to slow down |
| **Seq2Seq:** mainland cell phone users slow down |
| **SACT:** the growth of cell phone users in chinese main-land will slow down |

| |
|---|
| **Source:** 自 去年 12 以来, 受 委内瑞拉 国内 大罢工 和 伊拉克 战争 的影响, 国际 市场 原油 价格 持续 上涨 。 |
| **Gold:** since december last year , the price of crude oil on the international market has kept rising due to the general strike in venezuela and the threat of war in iraq . |
| **Seq2Seq:** since december last year , the international mar-ket has continued to rise in the international market and the threat of the iraqi war has continued to rise . |
| **SACT:** since december last year , the interna-tional market of crude oil has continued to rise because of the strike in venezuela and the war in iraq . |

## Conclusion

A new model with the self-adaptive attention temperature for the softness of attention distribution;

Improved results on the datasets and showed that attention temperature differs for decoding diverse words;

Try to figure out better demonstration for the effects of temperature.