Attention mechanism

杜岳華

2020.3.7

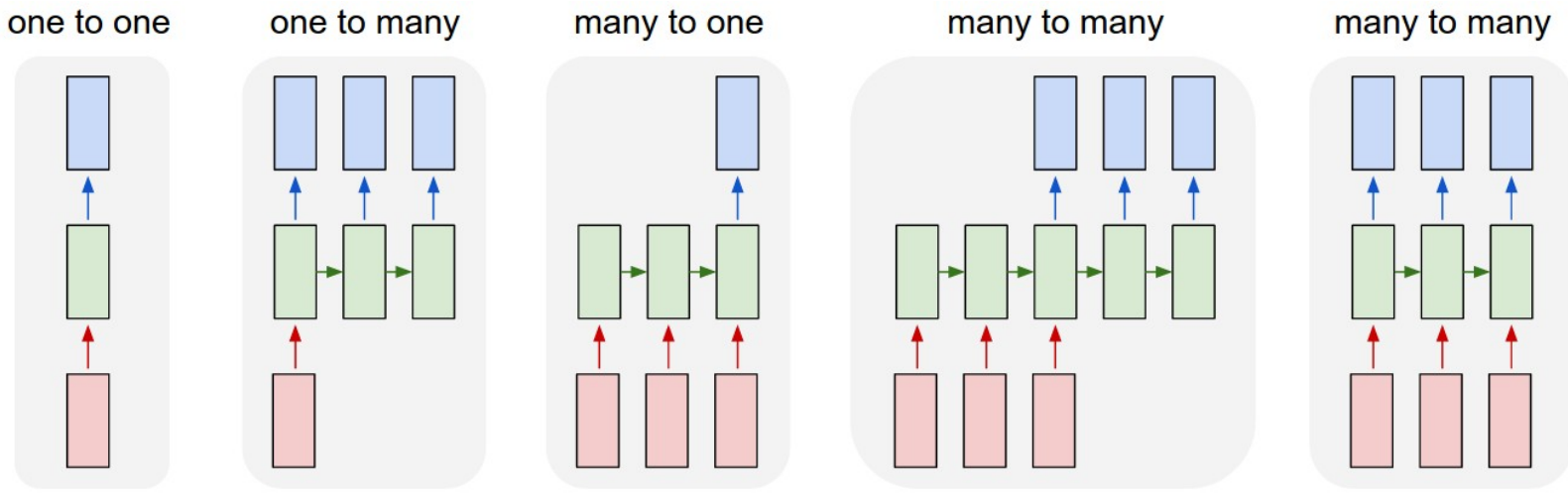
About me

- 中研院生物資訊學程博士生
- Julia Taiwan 社群主持人
- AI Tech 社群隱藏管理員
- Deep Learning 101 社群管理員
- 工研院 機器學習理論與實作 講師
- 著作：《Julia程式設計》、《Julia資料科學與科學計算》
- 專長：系統生物學、計算生物學、機器學習

Outline

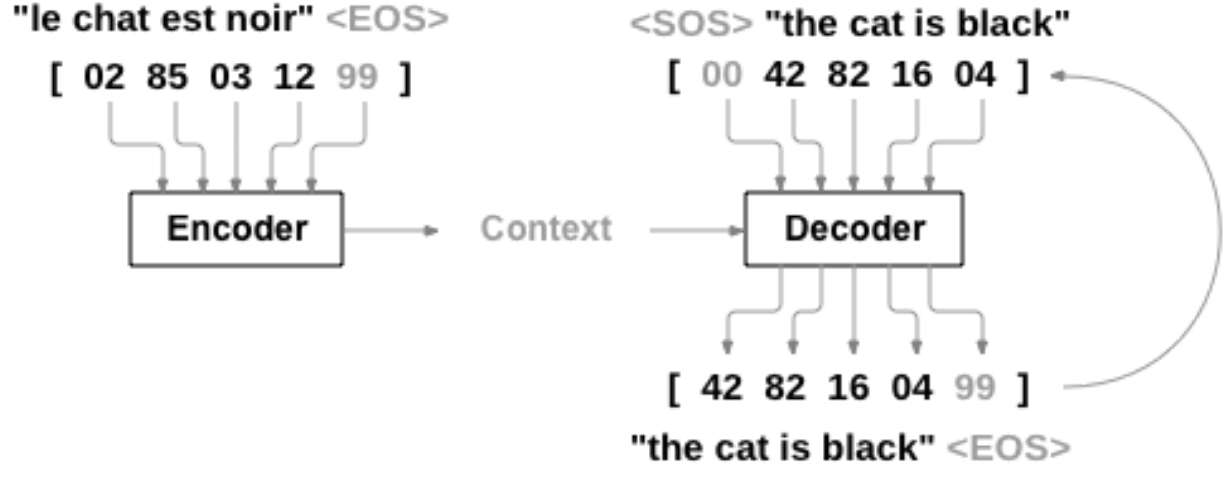
- RNN 的問題
- Seq2Seq encoder-decoder 架構
- Attention model 解決的問題
- Attention types
- Applications of attention
    - Translation
    - Summarization

RNN 的問題



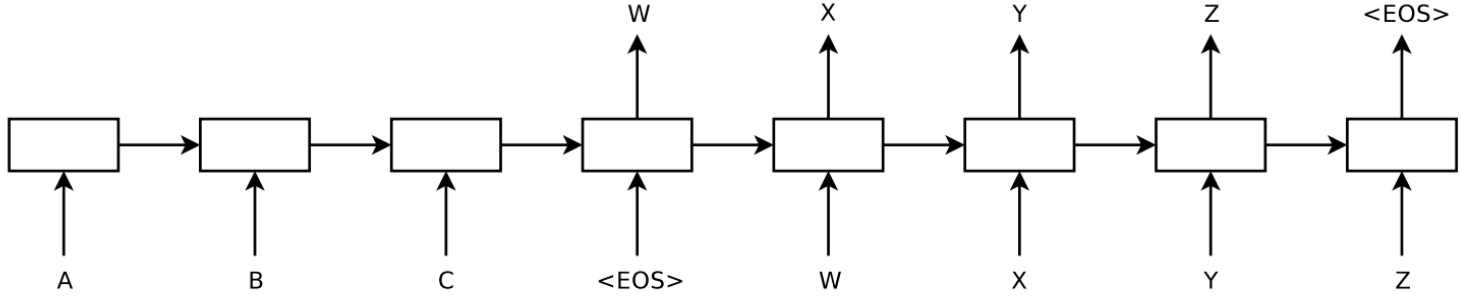
picture source (http://karpathy.github.io/2015/05/21/rnn-effectiveness/)

Seq2Seq encoder-decoder 架構



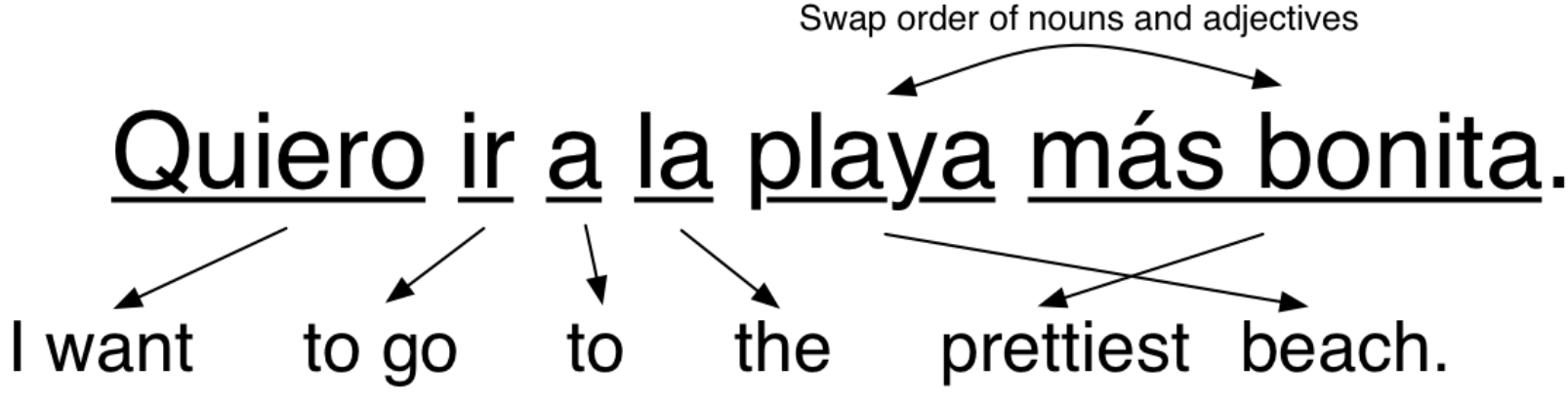
picture source (https://pytorch.org/tutorials/intermediate/seq2seq_translation_tutorial.html)

Seq2Seq encoder-decoder 架構



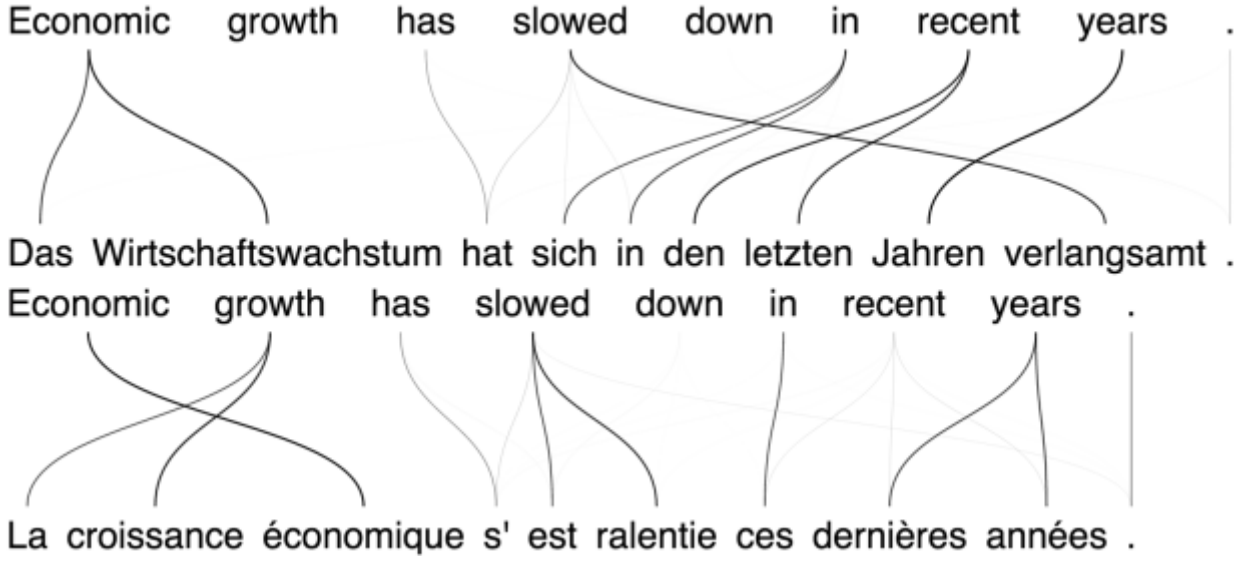
picture source (https://machinelearningmastery.com/encoder-decoder-recurrent-neural-network-models-neural-machine-translation/)

Attention model 解決的問題



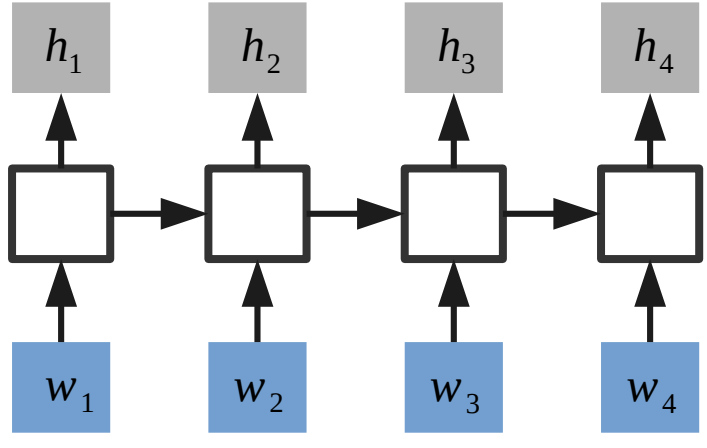
picture source (https://medium.com/@ageitgey/machine-learning-is-fun-part-5-language-translation-with-deep-learning-and-the-magic-of-sequences-2ace0acca0aa)

How to solve the problem?

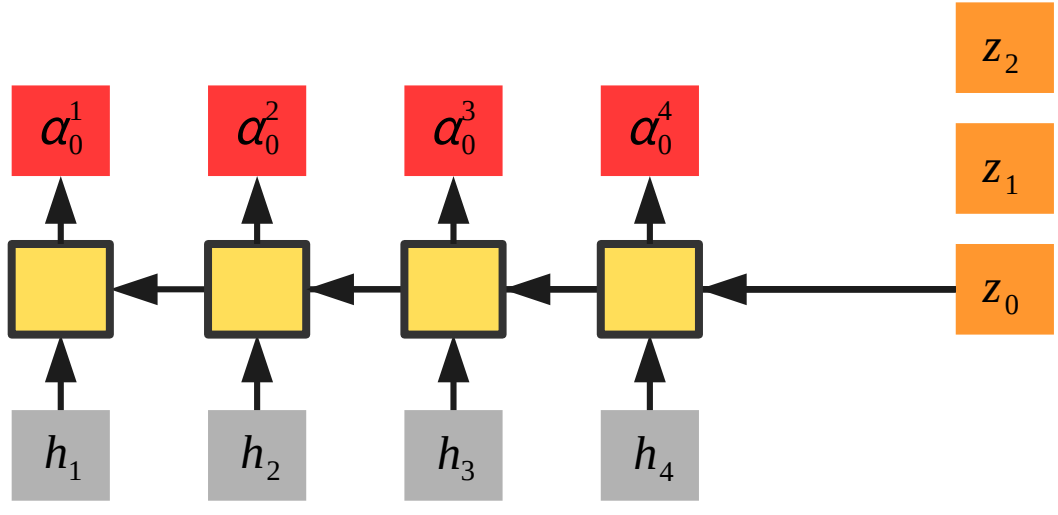


picture source (https://devblogs.nvidia.com/introduction-neural-machine-translation-gpus-part-3/)

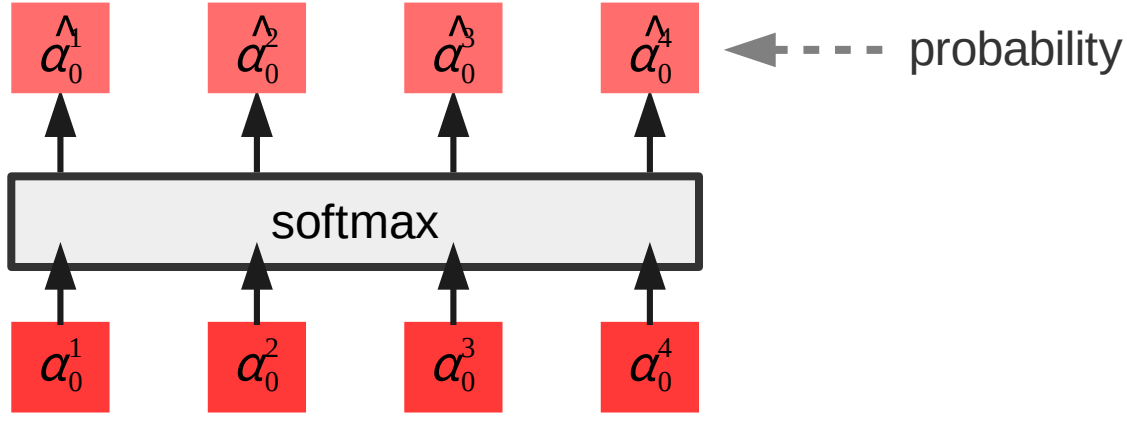
Attention mechanism



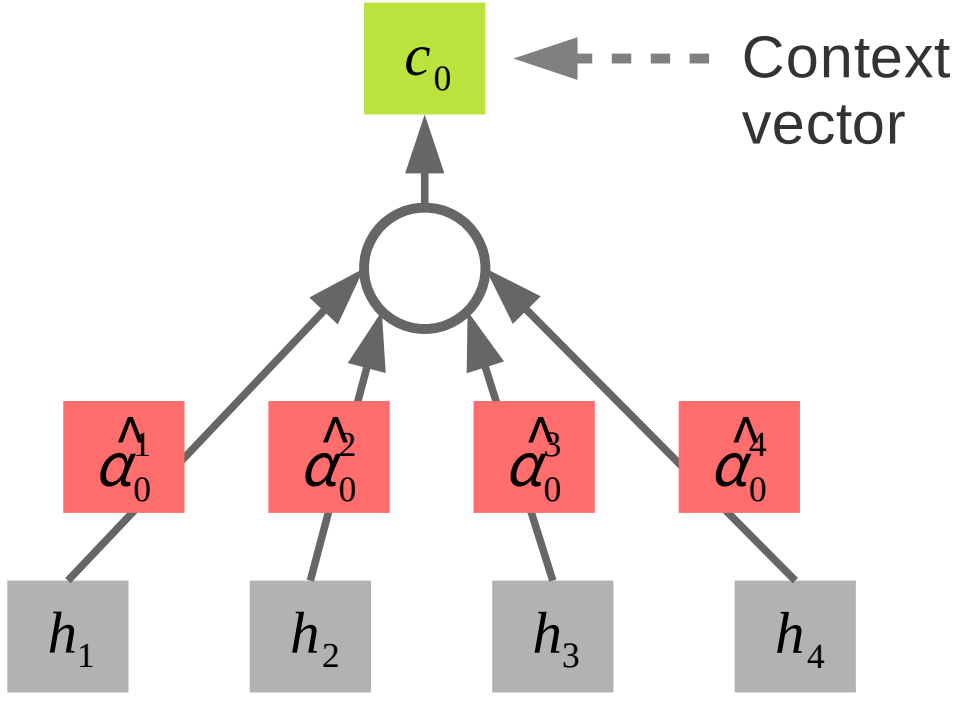
Attention mechanism



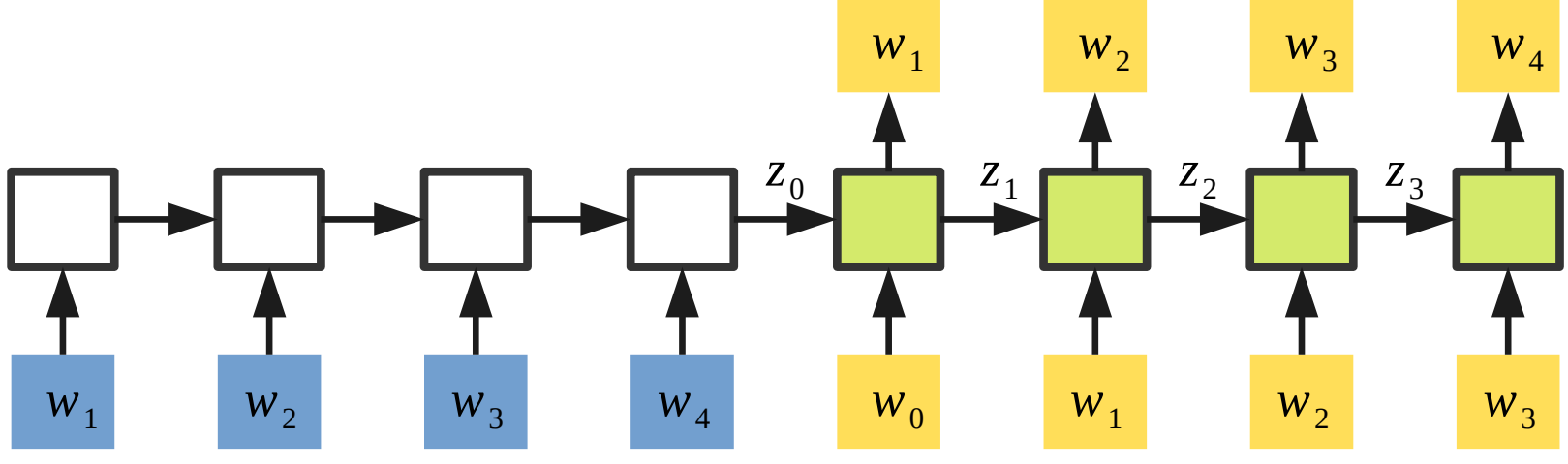
Attention mechanism



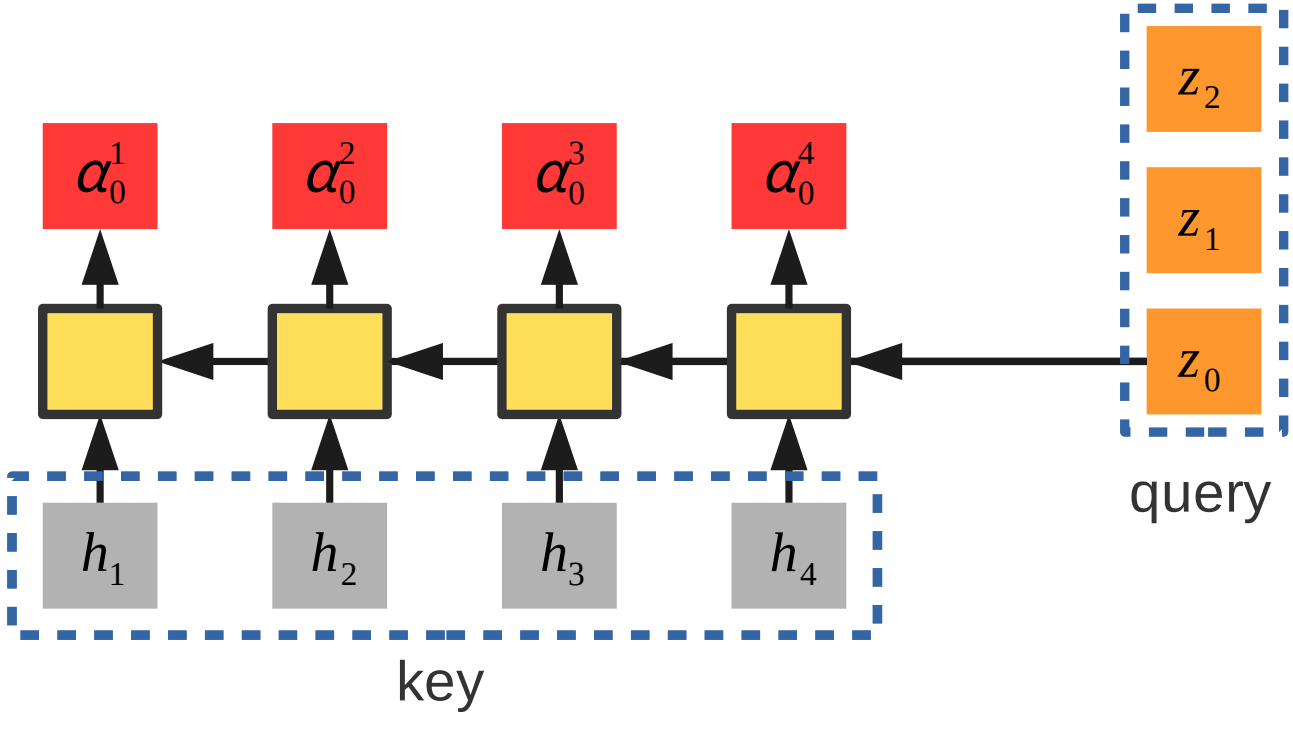
Attention mechanism



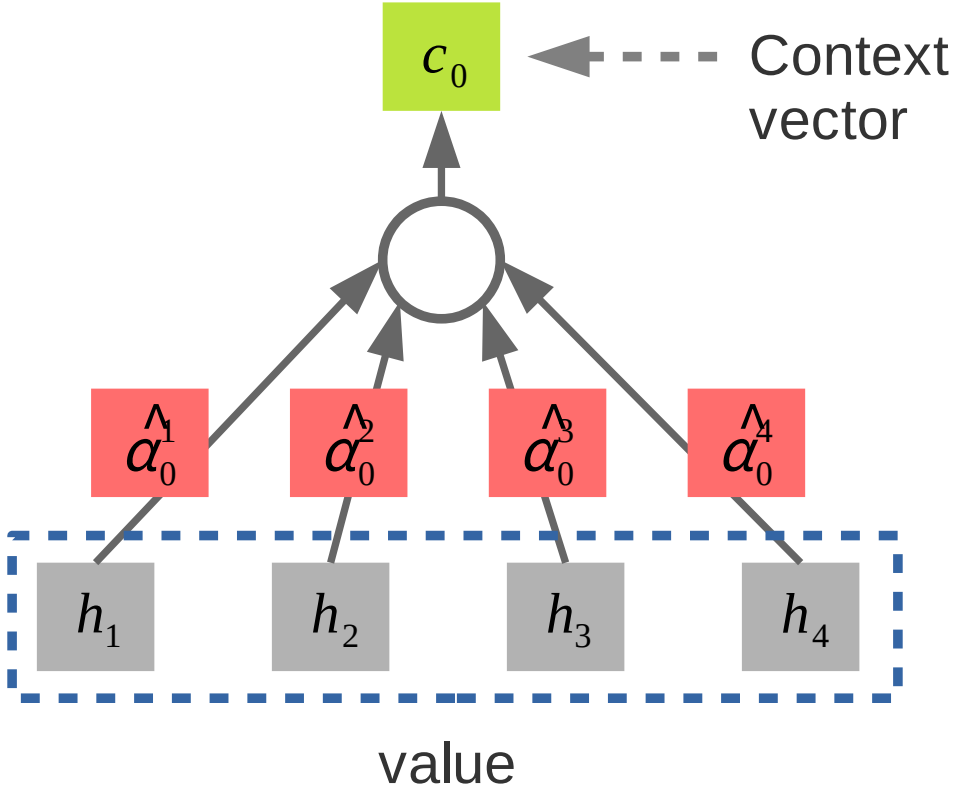
Attention mechanism



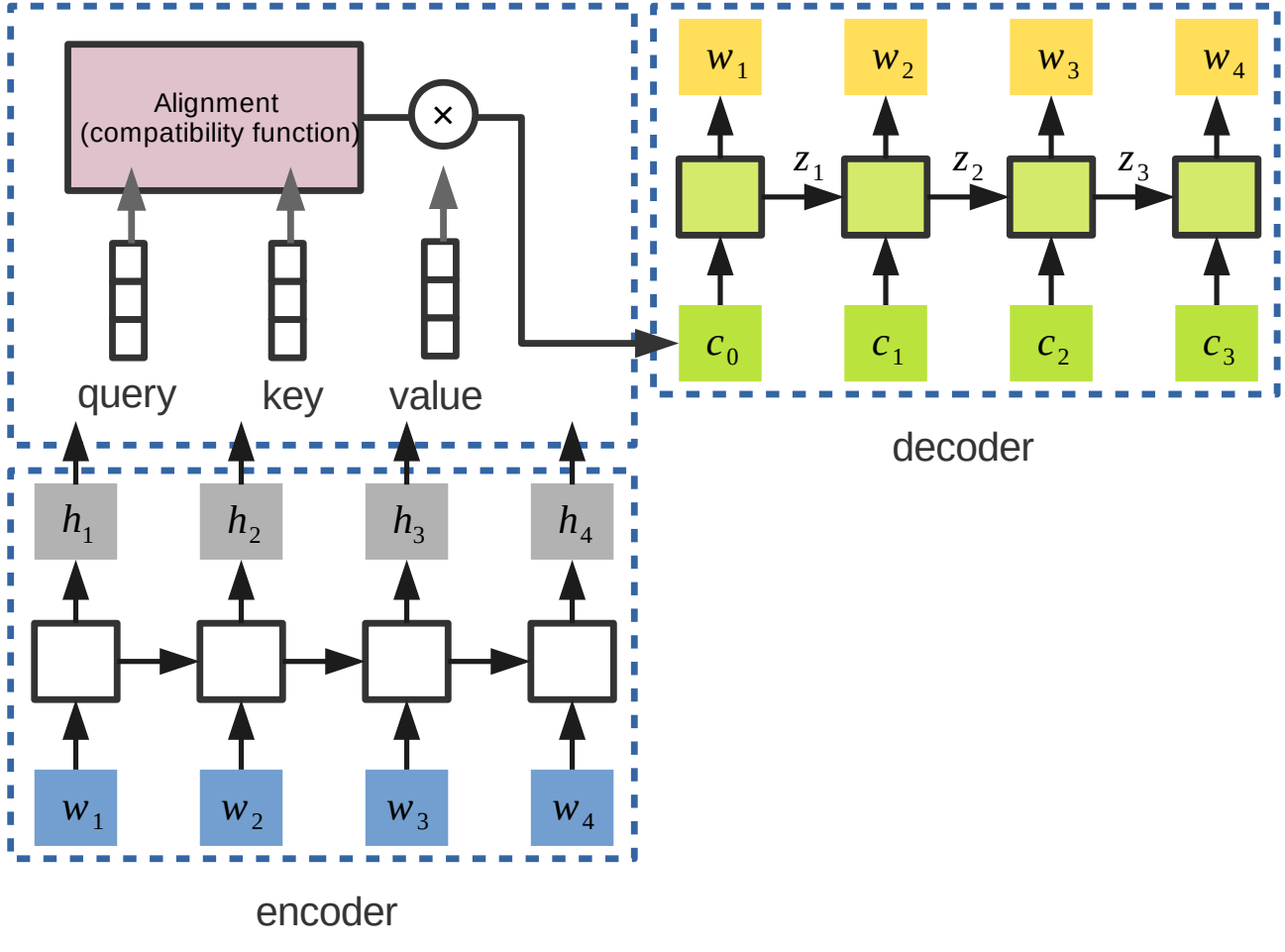
Attention mechanism



Attention mechanism



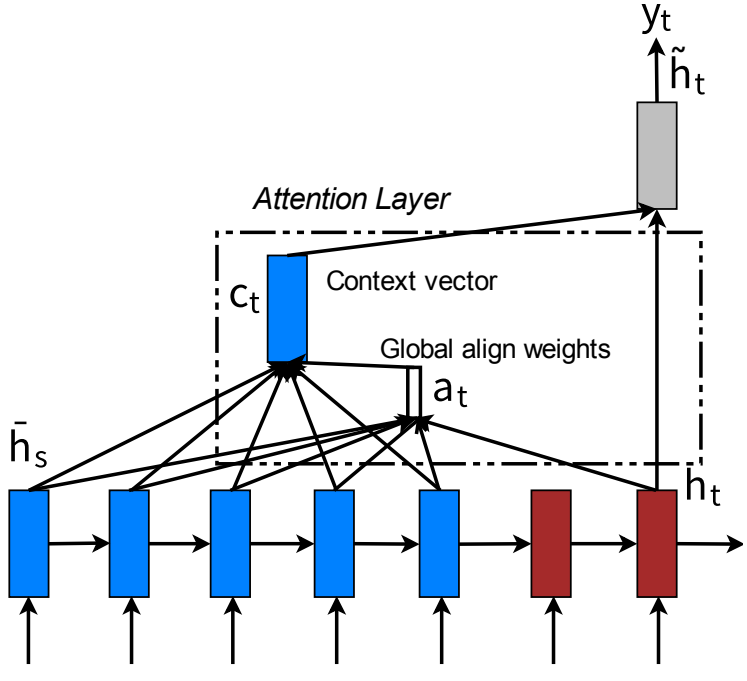
Attention mechanism



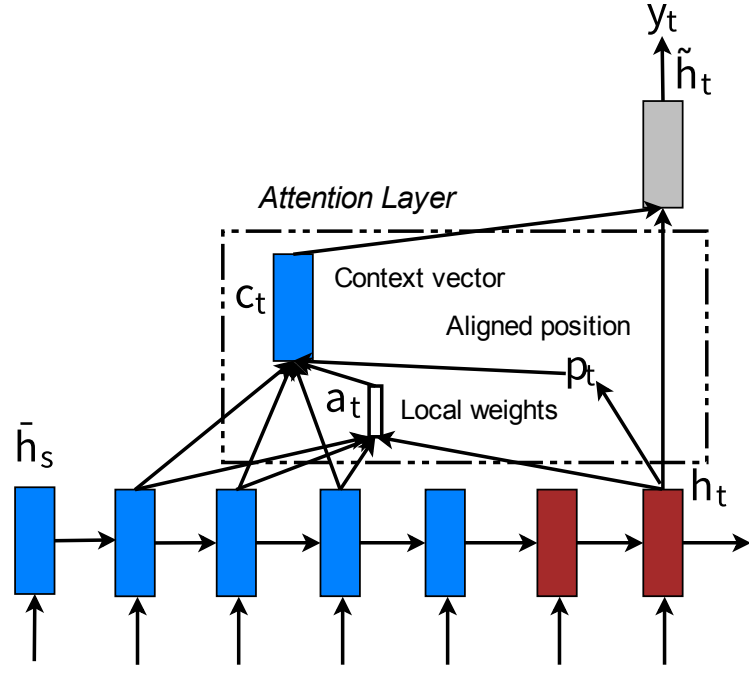
Attention types

- Global/local attention
- Hard/soft attention
- Self-attention

Global/local attention



Global attention



Local attention

Hard/soft attention

Soft attention

- Alignment weights are learned to attend over all data
- $0 \leq w \leq 1$
- Pro: model is smooth and differentiable
- Con: large computation if input is large

Hard attention

- Select part of data to attend or not at a time
- 0 or 1
- Pro: less inference time
- Con: model is non-differentiable

Alignment (compatibility function)

query: q_j, key: k_i

Location-based

$$\alpha_j^i = softmax(W_\alpha q_j)$$

Content-based

$$score(q_j, k_i) = cos([q_j; k_i])$$

Additive

$$score(q_j, k_i) = v_\alpha^T tanh(W_\alpha [q_j; k_i])$$

Alignment (compatibility function)

General

$$score(q_j, k_i) = q_j^T W_\alpha k_i$$

Dot-product

$$score(q_j, k_i) = q_j^T k_i$$

Scaled dot-product

$$score(q_j, k_i) = \frac{q_j^T k_i}{\sqrt{n}}$$

Applications of attention

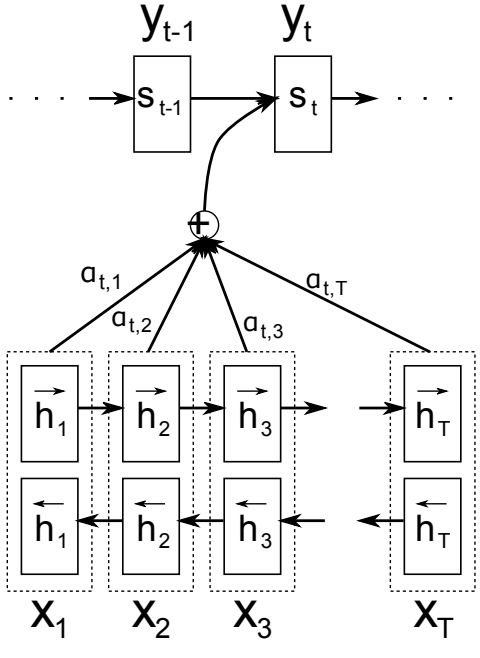
- Summarization: Rush 2015 (https://arxiv.org/abs/1509.00685)
- Translation: Bahdanau 2014 (https://arxiv.org/abs/1409.0473), Luong 2015 (https://arxiv.org/abs/1508.04025)
- Image caption: Xu 2015 (https://arxiv.org/abs/1502.03044)
- ...

Neural Machine Translation by Jointly Learning to Align and Translate

Yoshua Bengio

ICLR 2015

Bidirectional LSTM as encoder



Attention v.s. Seq2Seq

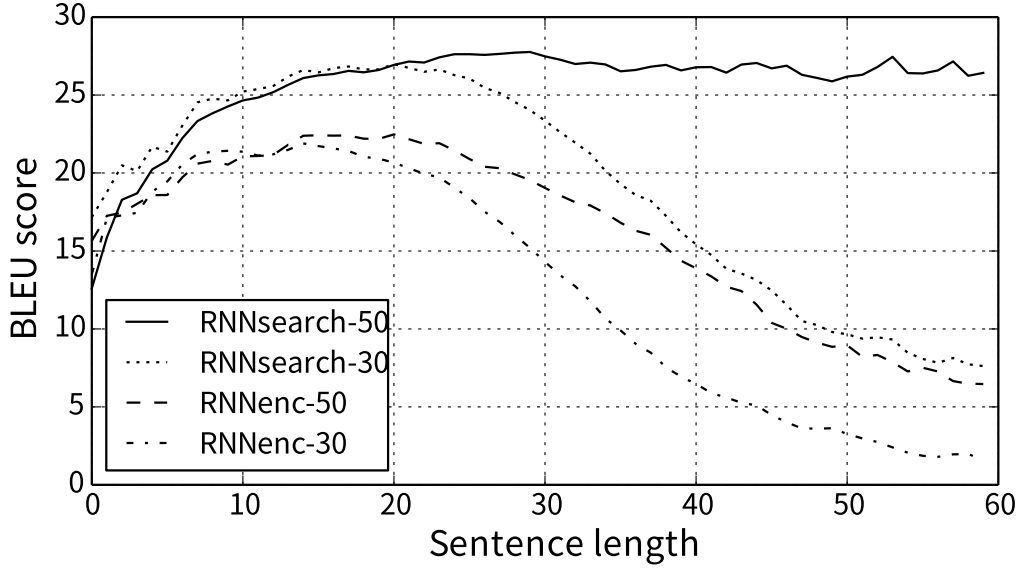


Figure 2: The BLEU score
of the generated translat
on the testset with respe
to the lengthsof the sen
tences. The results are o
the full test set which in
cludes sentences having
known words to the mod

Bilingual evaluation understudy (BLEU)

A way to evaluate the quality of machine-translated text from one natural language to another.

A modified form of precision to compare a translation candidate with multiple references

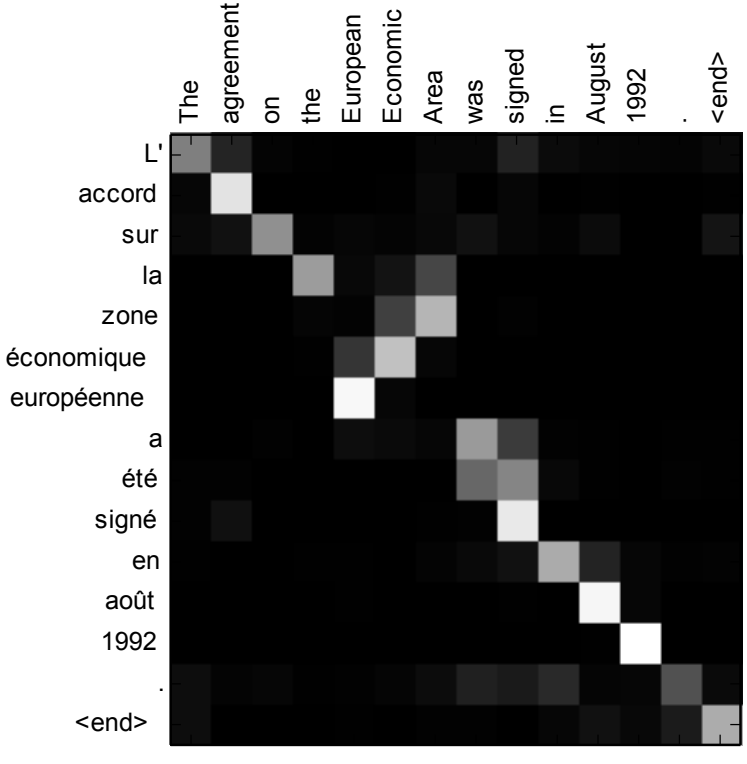
- Candidate: the the the the the the the
- Reference 1: the cat is on the mat
- Reference 2: there is a cat on the mat

$$BLEU = \frac{\text{matched number of words in candidate}}{\text{total number of words in candidate}} = \frac{7}{7}$$

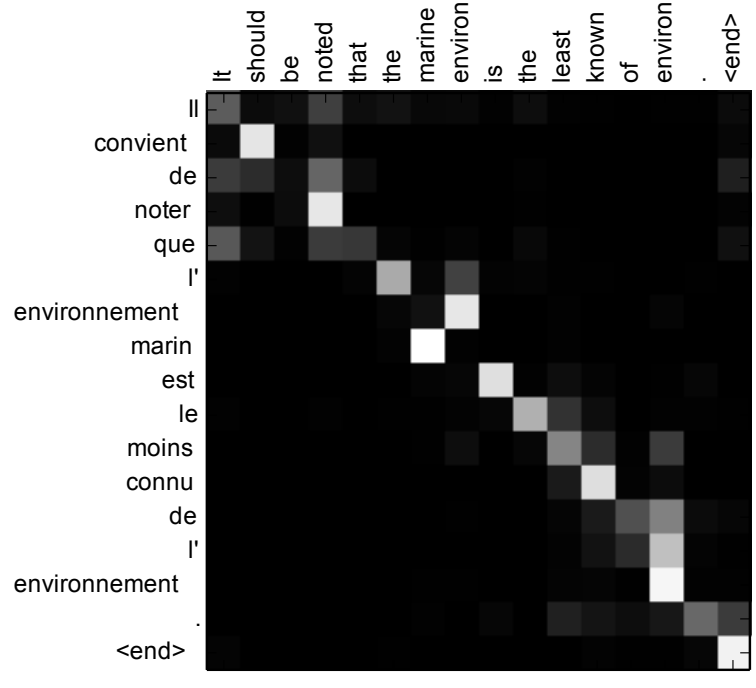
Quantitative comparison

Model	All	No UNK$^{\circ}$
RNNencdec-30	13.93	24.19
RNNsearch-30	21.50	31.44
RNNencdec-50	17.82	26.71
RNNsearch-50	26.75	34.16
RNNsearch-50*	28.45	36.15
Moses	33.30	35.63

Translation

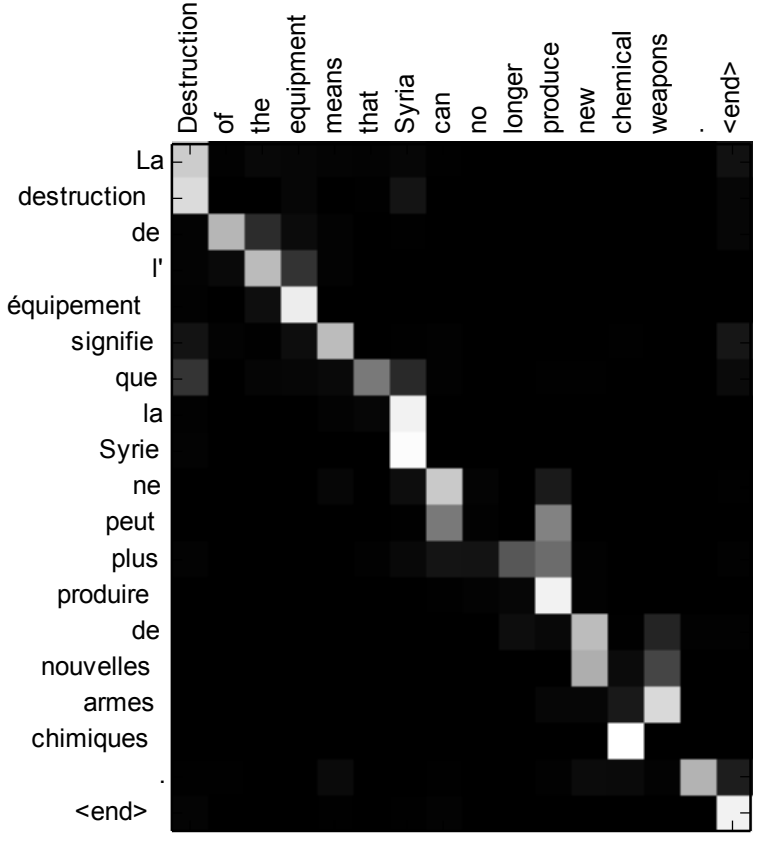


(a)

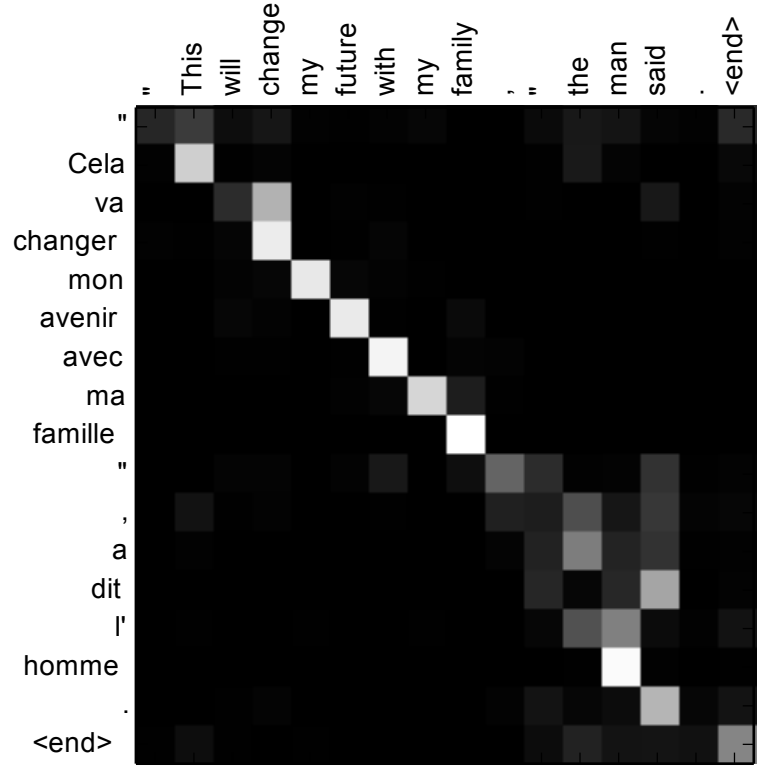


(b)

Translation



(c)



(d)

A Neural Attention Model for Abstractive Sentence Summarization

Facebook

EMNLP 2015

Summarization

To shorten the sentense while keep relevant or important information.

But...what's so different from translation operationally?

What is summarization?

We usually execute some series of operations to summarize a sentense/article.

These operations are Deletion(刪除), Generalization (廣義化) and Paraphrase (改寫).

Types of summarization

Compressive

Summarize original sentnese by **deletion-only**

Extractive

Summarize original sentnese by **deletion and reordering**

Abstractive

Summarize original sentnese by **arbitary transformation**

Proposed models

Bag-of-Words Encoder

$$
\begin{aligned}
enc_1(\mathbf{x}, \mathbf{y}_c) &= \mathbf{p}^T \tilde{\mathbf{x}} \\
\mathbf{p} &= [1/M, \cdots, 1/M] \\
\tilde{\mathbf{x}} &= F[x_1, \cdots, x_M]
\end{aligned}
$$

Ignoring properties of the original order or relationships between neighboring words

Convolutional Encoder

$$
enc_2(\mathbf{x}, \mathbf{y}_c) = (\text{temporal convolution layer} \rightarrow \text{max pooling layer})^L
$$

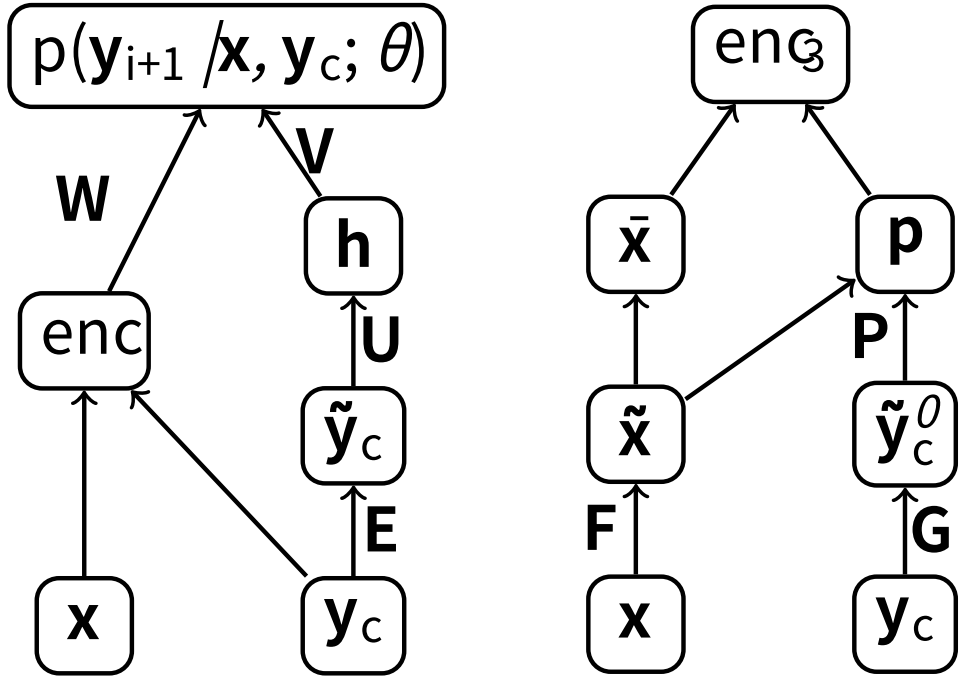
Allowing local interactions between words while also not requiring the context y_c while encoding the input.

Proposed models

Attention-Based Encoder

$$
\begin{aligned}
enc_3(\mathbf{x}, \mathbf{y}_c) &= \mathbf{p}^T \bar{\mathbf{x}} \\
\mathbf{p} &= exp(\tilde{\mathbf{x}} P \mathbf{y}'_c) \\
\tilde{\mathbf{x}} &= F[x_1, \cdots, x_M] \\
\mathbf{y}'_c &= G[y_{i-C+1}, \cdots, y_i] \\
\bar{x}_i &= \sum_{q=i-Q}^{i+Q} \tilde{\mathbf{x}}_i / Q
\end{aligned}
$$

Neural network language model (NNLM)



Perplexity 困惑度

A method to evaluate a language model. A language model describes the probability distribution over whole sentnese.

Perplexity of discrete probability distribution

$$2^{H(p)} = 2^{-\sum p(x) \log 2p(x)}$$

Language model is a probability distribution

If each word is specified in a sentense, the meaning of a sentense is clear. We evaluate the occurrence (probability) of words in sentense. Lower entropy means precise meaning. Small perplexity is better.

Evaluation

Model	Encoder	Perplexity
KN-Smoothed 5-Gram	none	183.2
Feed-Forward NNLM	none	145.9
Bag-of-Word	enc_1	43.6
Convolutional (TDNN)	enc_2	35.9
Attention-Based (ABS)	enc_3	27.1

Recall-Oriented Understudy for Gisting Evaluation (ROUGE)

A method to evaluate machine translation and machine abstraction. It evaluates generated result and a reference (usually written by human), and further calculate the similarity between them.

$$\text{ROUGE-N} = \frac{\sum_{S \in reference} \sum_{gram_n \in S} count_{match}(gram_n}{\sum_{S \in reference} \sum_{gram_n \in S} count(gram_n)}$$

- $count_{match}(gram_n)$: Maximum number of n-grams co-occuring in candidate and reference.

ROUGE: A Package for Automatic Evaluation of Summaries (https://www.aclweb.org/anthology/W04-1013/)

ROUGE-N

- Candidate: the cat was found under the bed
- Reference: the cat was under the bed

1-gram

- Candidate: the, cat, was, found, under, bed
- Reference: the, cat, was, under, bed
- ROUGE-1: $5/5 = 1.0$

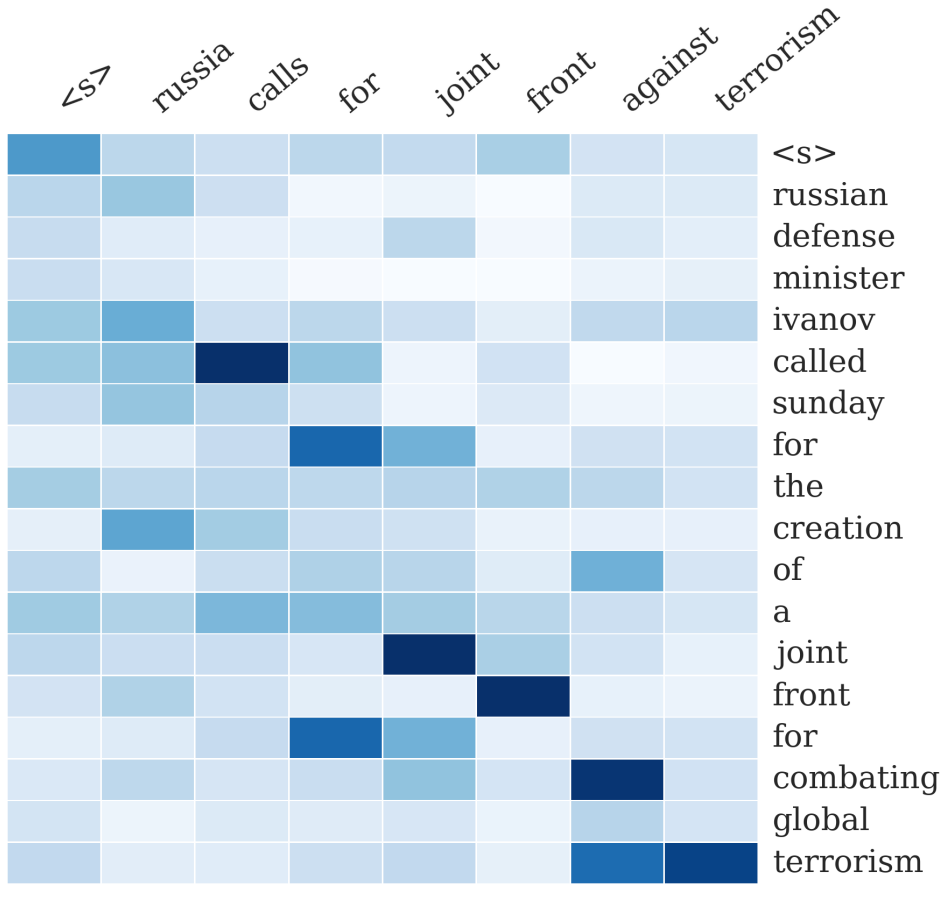
2-gram

- Candidate: the cat, cat was, was found, found under, under the, the bed
- Reference: the cat, cat was, was under, under the, the bed
- ROUGE-2: $4/5 = 0.8$

Evaluation

Decoder	Model	Cons.	R-1	R-2	R-L
Greedy	ABS+	Abs	26.67	6.72	21.70
Beam	BOW	Abs	22.15	4.60	18.23
Beam	ABS+	Ext	27.89	7.56	22.84
Beam	ABS+	Abs	28.48	8.91	23.97

Summarization



Summarization

I(1): a detained iranian-american academic accused of acting against national security has been released from a tehran prison after a hefty bail was posted , a to p judiciary official said tuesday .
G: iranian-american academic held in tehran released on bail
A: detained iranian-american academic released from jail after posting bail
A+: detained iranian-american academic released from prison after hefty bail

Summarization

I(2): ministers from the european union and its mediterranean neighbors gathered here under heavy security on monday for an unprecedented conference on economic and political cooperation .
G: european mediterranean ministers gather for landmark conference by julie bradford
A: mediterranean neighbors gather for unprecedented conference on heavy security
A+: mediterranean neighbors gather under heavy security for unprecedented conference

Summarization

I(3): the death toll from a school collapse in a haitian shanty-town rose to ## after rescue workers uncovered a classroom with ## dead students and their teacher , officials said saturday .
G: toll rises to ## in haiti school unk : official
A: death toll in haiti school accident rises to ##
A+: death toll in haiti school to ## dead students

Summarization

I(4): australian foreign minister stephen smith sunday congratulated new zealand ’s new prime minister-elect john key as he praised ousted leader helen clark as a “ gutsy ” and respected politician .
G: time caught up with nz ’s gutsy clark says australian fm
A: australian foreign minister congratulates new nz pm after election
A+: australian foreign minister congratulates smith new zealand as leader

Summarization

I(5): two drunken south african fans hurled racist abuse at the country 's rugby sevens coach after the team were eliminated from the weekend 's hong kong tournament , reports said tuesday .
G: rugby union : racist taunts mar hong kong sevens : report
A: south african fans hurl racist taunts at rugby sevens
A+: south african fans racist abuse at rugby sevens tournament

Summarization

I(6): christian conservatives – kingmakers in the last two us presidential elections – may have less success in getting their pick elected in #### , political observers say .
G: christian conservatives power diminished ahead of #### vote
A: christian conservatives may have less success in #### election
A+: christian conservatives in the last two us presidential elections

Thank you for attention.

References

- Attention? Attention! (https://lilianweng.github.io/lil-log/2018/06/24/attention-attention.html)
- 放棄幻想，全面擁抱 Transformer：自然語言處理三大特徵抽取器（CNN/RNN/TF）比較 (http://bangqu.com/f7t7X5.html)

Papers