Eric Mintun

HEP-AI Journal Club
May 15th, 2018

Outline

- Motivating example and definition

  Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. “Neural machine translation by jointly learning to align and translate.” In International Conference on Learning Representations, 2015. arXiv:1409.0473 [cs.CL]

- Generalizations and a little theory

  Yoon Kim, Carl Denton, Luong Hoang, and Alexander M. Rush. “Structured attention networks.” In International Conference on Learning Representations, 2017. arXiv: 1702.00887 [cs.CL]

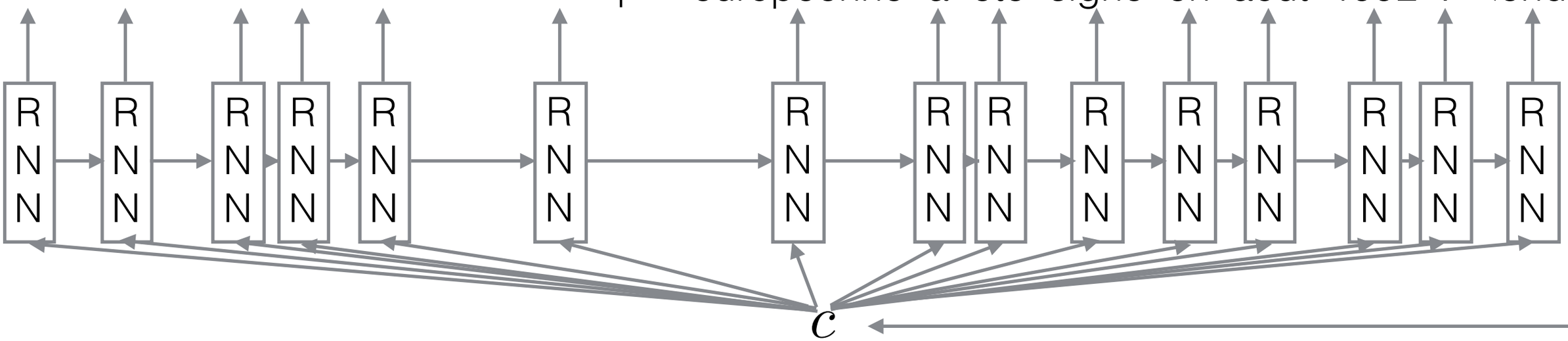
- Why attention might be better than RNNs and CNNs

  Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin. “Attention is all you need.” In 31st Conference on Neural Information Processing Systems (NIPS 2017). arXiv:1706.03762 [cs.CL]

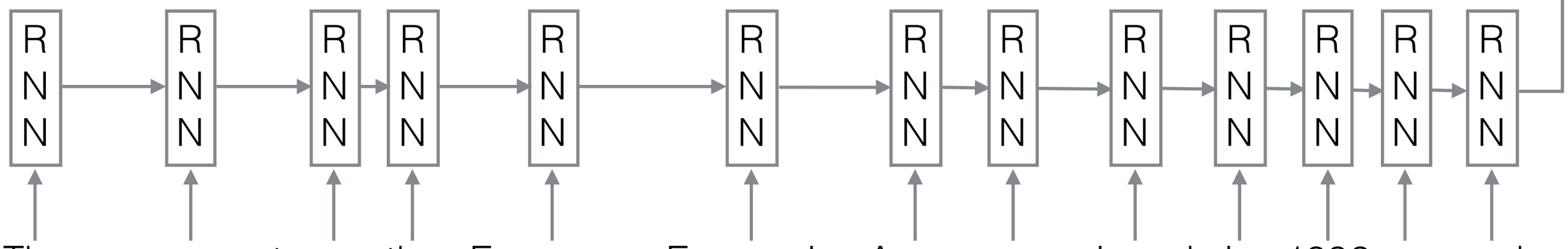
Translation

French

L’ accord sur la zone économique européenne a été signé en août 1992 . <end>



Context Vector



The agreement on the European Economic Area was signed in 1992 . <end>

English

Translation

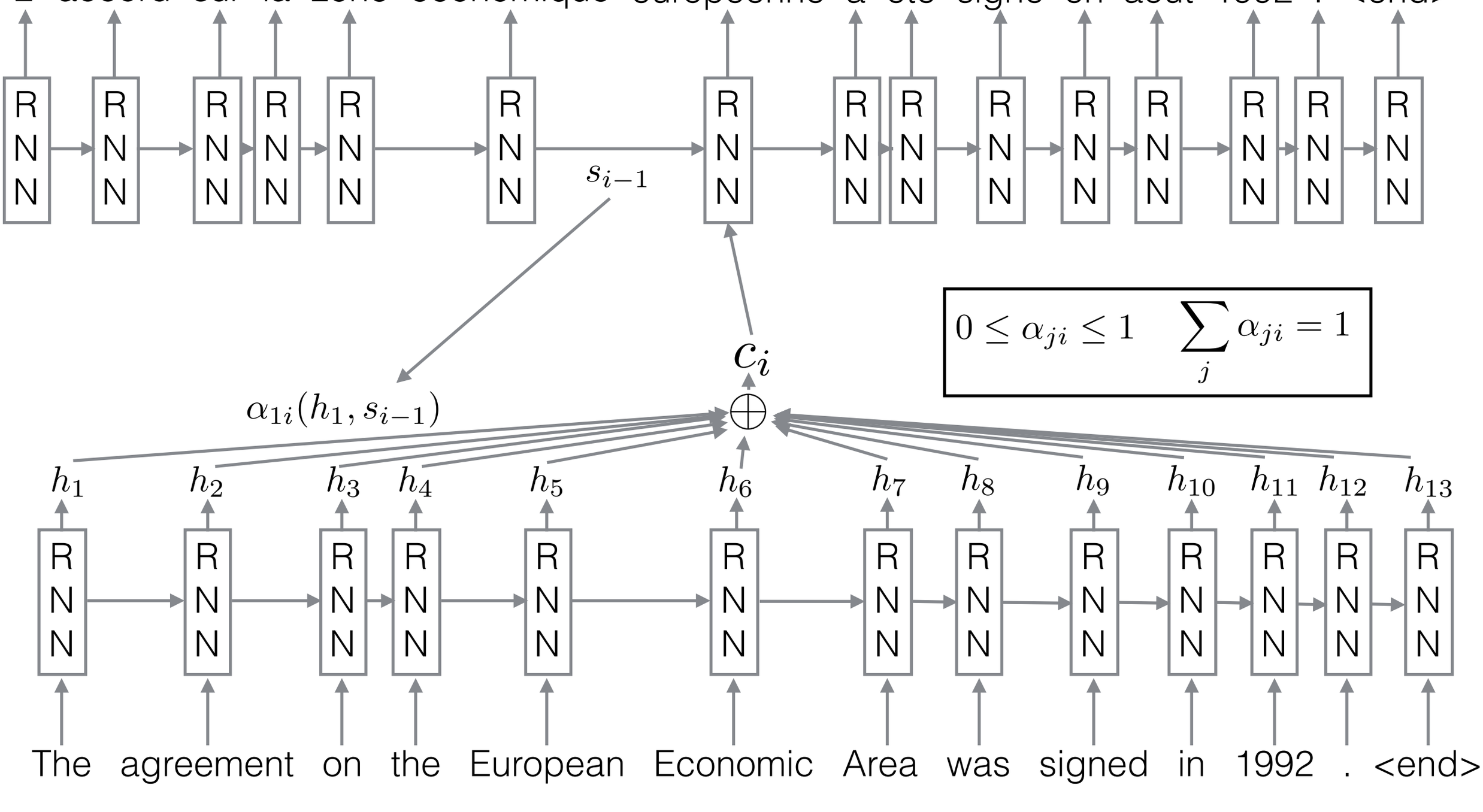
- Fixed-size context vector struggles with long sentences, fails later in sentence.

An admitting privilege is the right of a doctor to admit a patient to a hospital or a medical centre to carry out a diagnosis or a procedure, based on his status as a health care worker at a hospital.

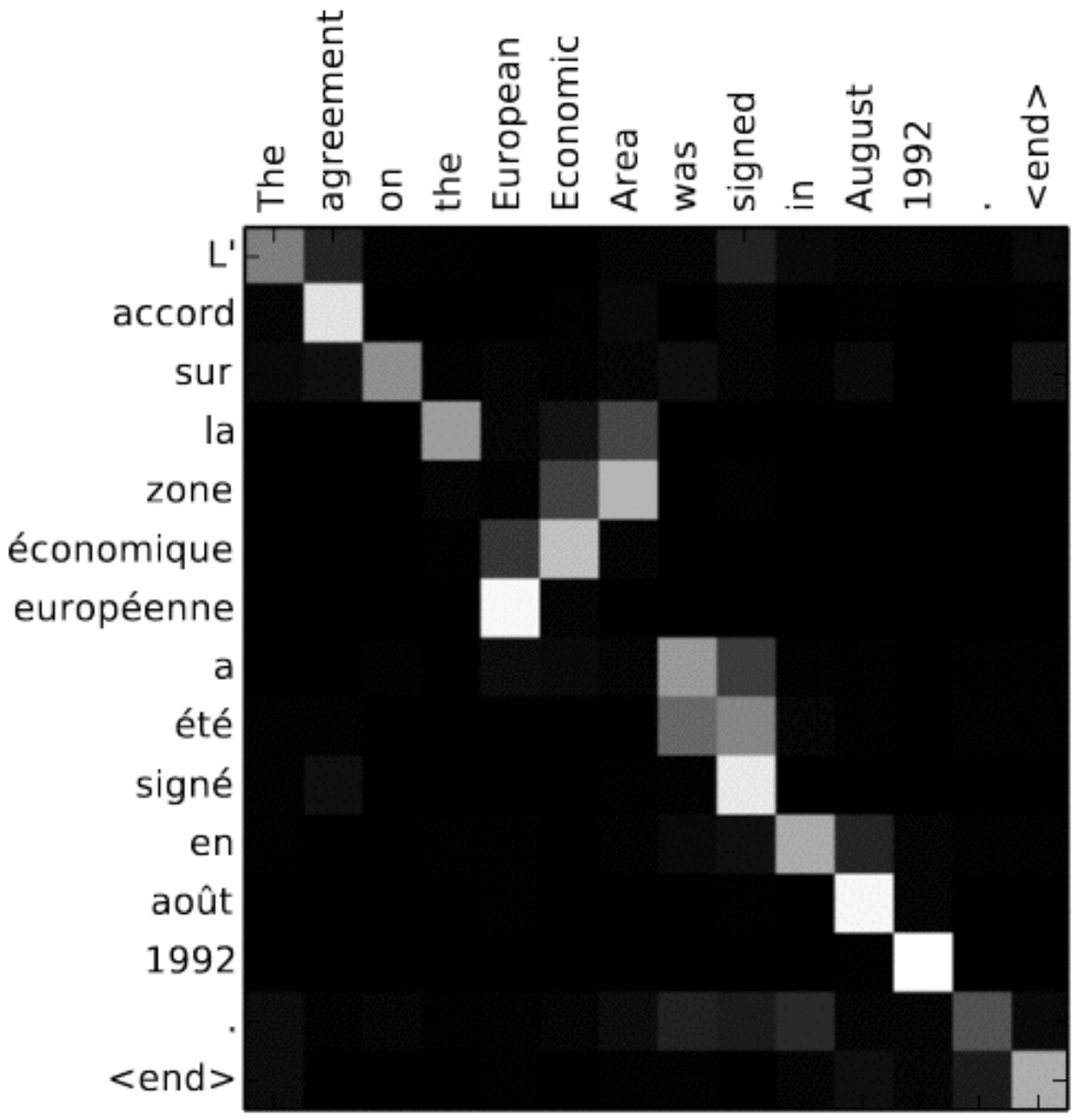
Un privilège d'admission est le droit d'un médecin de reconnaître un patient à l'hôpital ou un centre médical d'un diagnostic ou de prendre un diagnostic en fonction de son état de santé.

- Underlined portion becomes 'based on his state of health'.

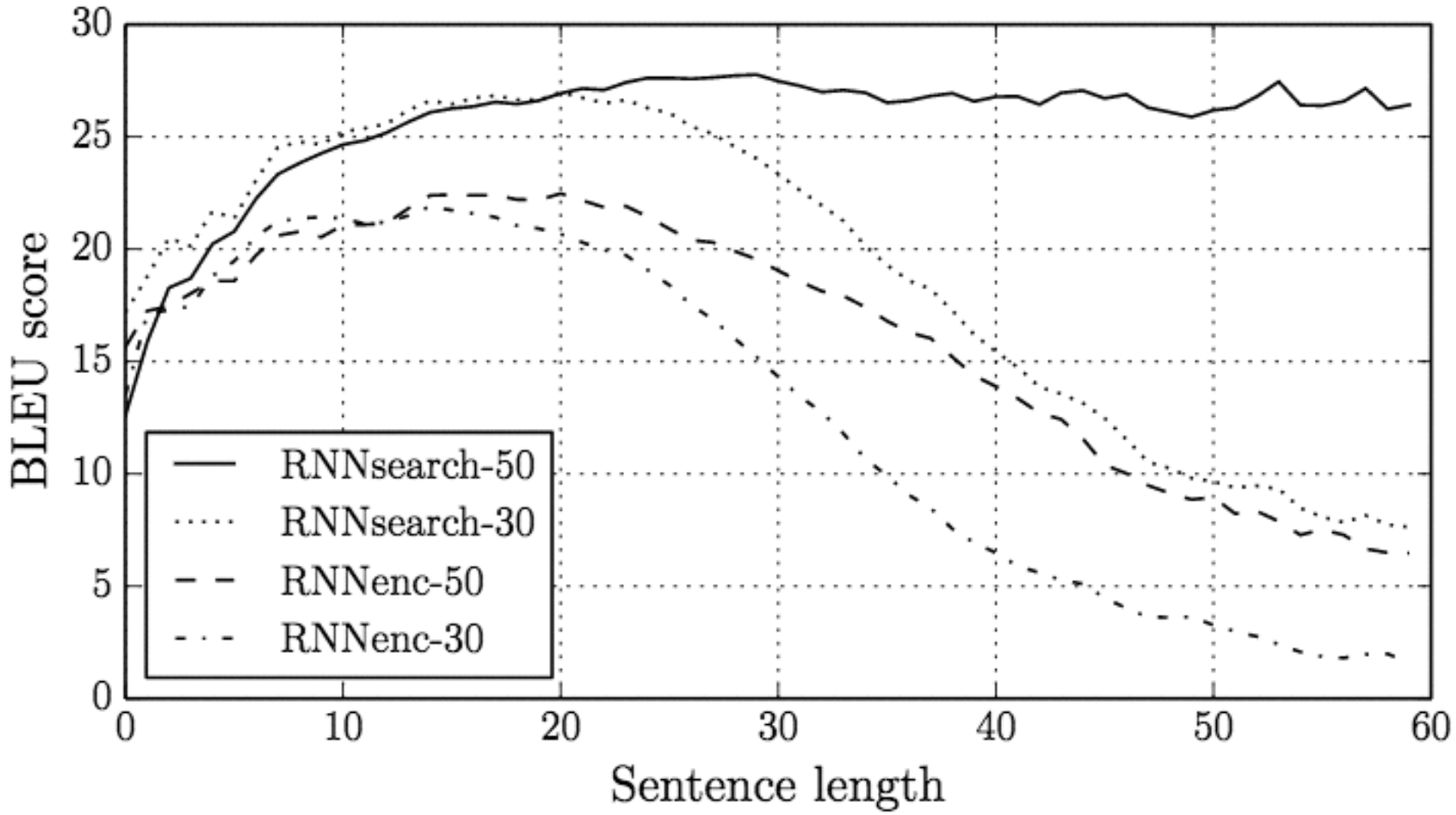
Translation w/ Attention



Translation w/ Attention

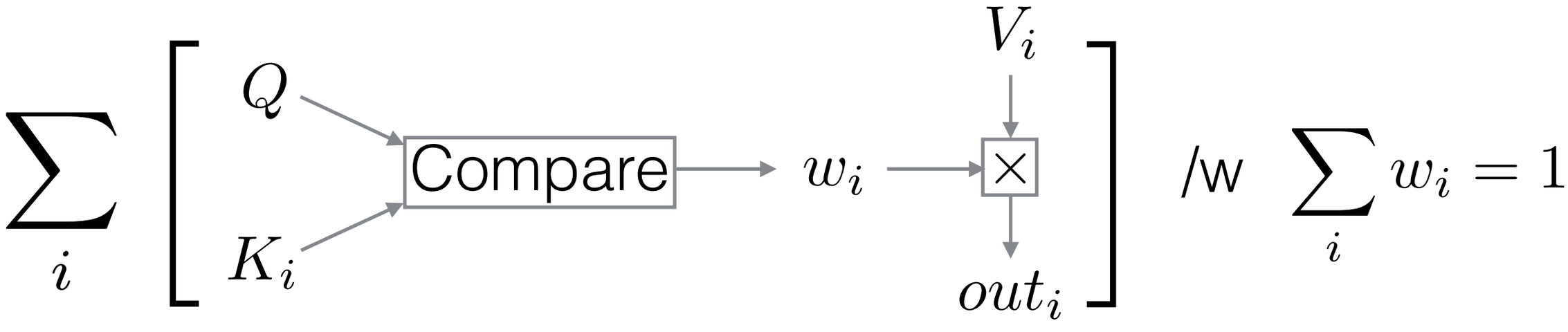


Translation w/ Attention



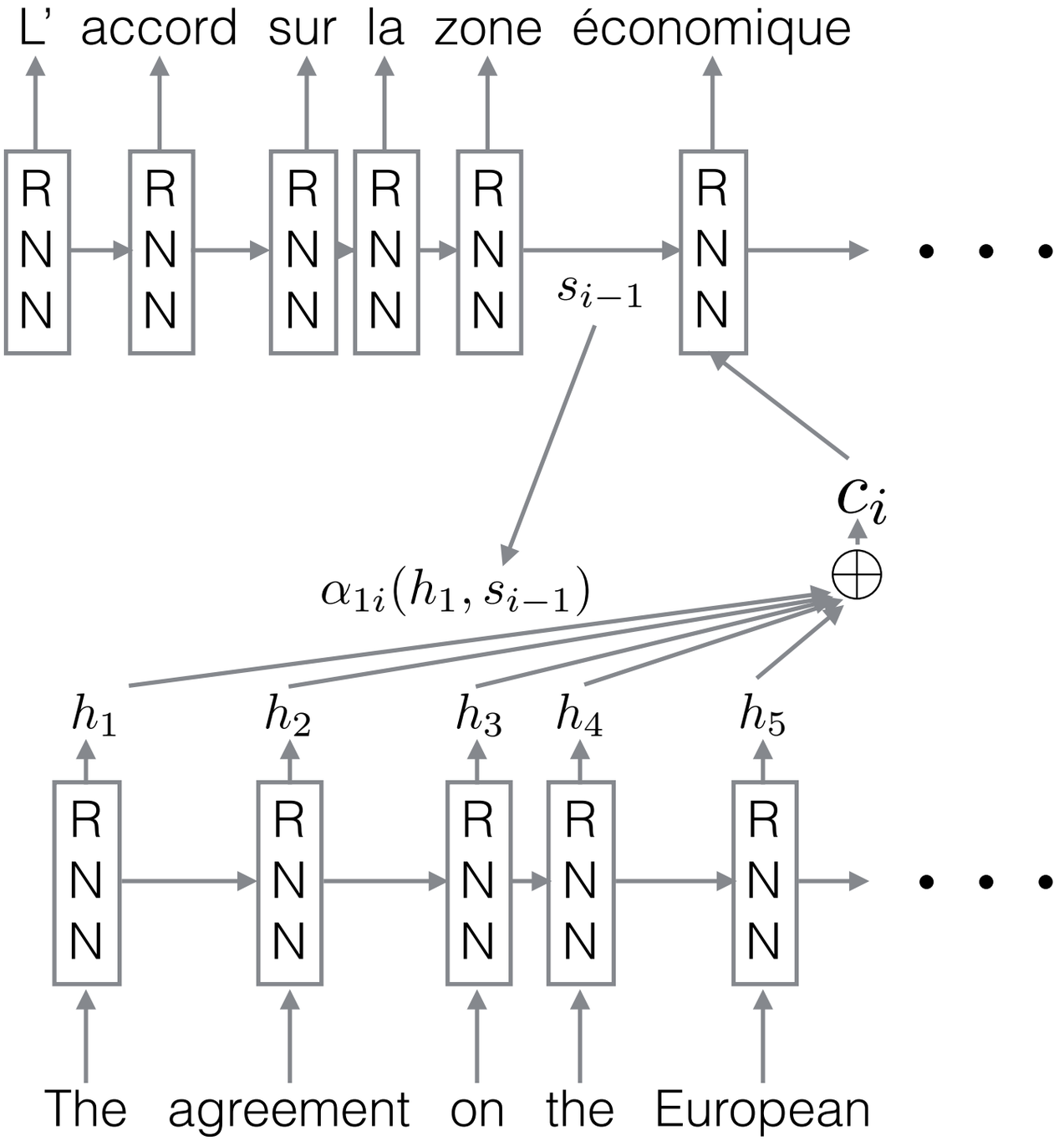
Attention

- Attention consists of learned key-value pairs.
- Input query is compared against the key. A better match lets more of the value through:



- Additive compare: Q and K fed into neural net
- Multiplicative compare: Dot-product Q and K

Keys/Values for Example



Query: s_{i-1}

Keys: h_j

Values: h_j

Compare: $\alpha_{ji}(h_j, s_{i-1})$

Additive Attention

Outline

- Motivating example and definition

  Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. “Neural machine translation by jointly learning to align and translate.” In International Conference on Learning Representations, 2015. arXiv:1409.0473 [cs.CL]

- Generalizations and a little theory

  Yoon Kim, Carl Denton, Luong Hoang, and Alexander M. Rush. “Structured attention networks.” In International Conference on Learning Representations, 2017. arXiv: 1702.00887 [cs.CL]

- Why attention might be better than RNNs and CNNs

  Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin. “Attention is all you need.” In 31st Conference on Neural Information Processing Systems (NIPS 2017). arXiv:1706.03762 [cs.CL]

Structured Attention

- What if we know trained attention should have a known structure? E.g.:
- Each output by decoder should attend to a connected subsequence in encoder (character to word conversion).
- Output sequence is organized as a tree (sentence parsing, equation input and output).

Structured Attention

- Attention weights α_i define a probability distribution. Write context vector as:

$$\mathbf{c} = \mathbb{E}_{z \sim p(z|x,q)}[f(x,z)]$$

$$f(\mathbf{x}, z) = \mathbf{x}_z$$

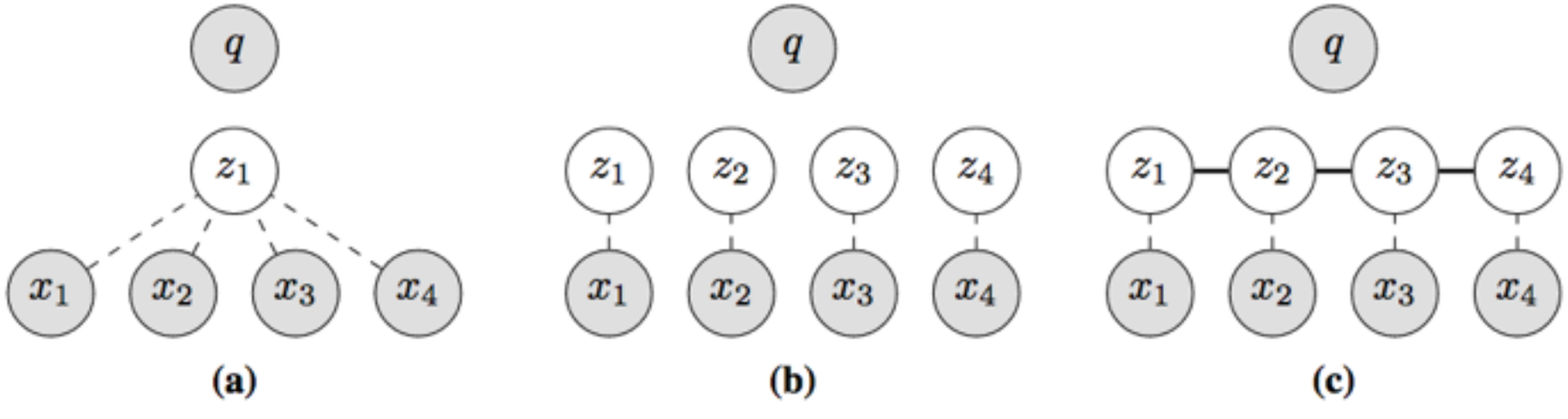
$$z \in 1, \ldots, n$$

- Generalize this by adding more latent variables, changing annotation function. Add structure by dividing into cliques:

$$\mathbf{c} = \mathbb{E}_{z \sim p(z|x,q)}[f(x,z)] = \sum_C \mathbb{E}_{z \sim p(z_C|x,q)}[f_C(x, z_C)]$$

$$p(z|x,q;\theta) = \text{softmax}\left(\sum_C \theta_C(z_C)\right)$$

Subsequence Attention



- (a) original unstructured attention network

- (b) 1 independent binary latent variable per input:

$$\mathbf{c} = \mathbb{E}_{z_1,\ldots,z_n}[f(x,z)] = \sum_{i=1}^{n} p(z_i = 1|x,q)\mathbf{x}_i \qquad f(\mathbf{x},z) = \sum_{i=1}^{n} \mathbb{1}\{z_i = 1\}\mathbf{x}_i$$

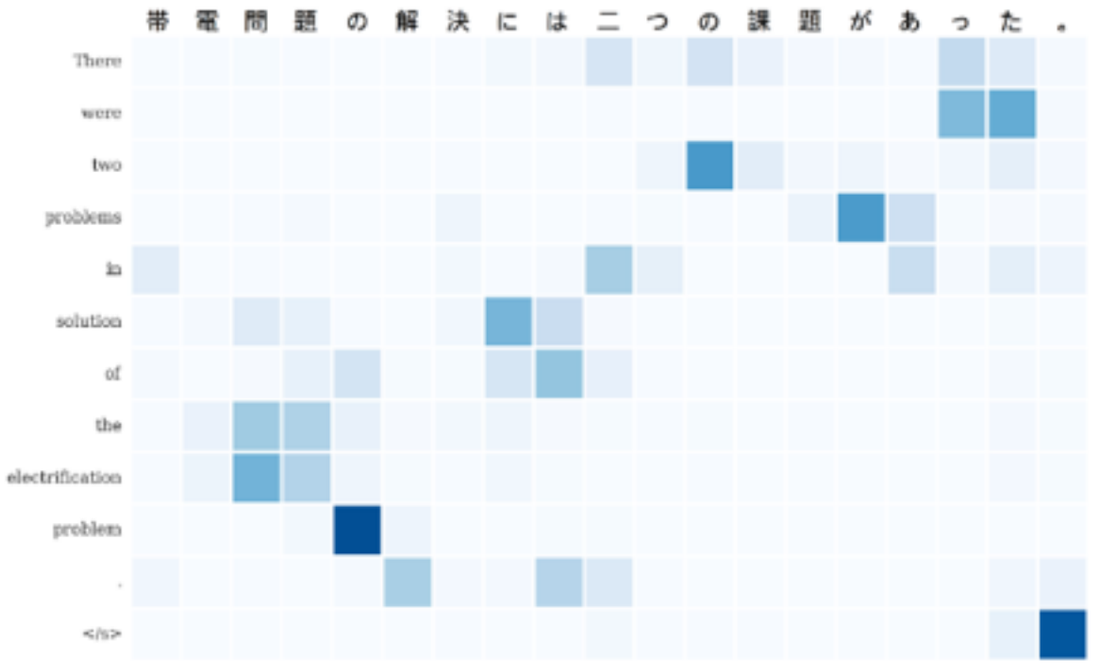
$$p(z_i = 1|x,q) = \text{sigmoid}(\theta_i) \qquad z_i \in 0,1$$

- (c) probability of each z depends on neighbors.

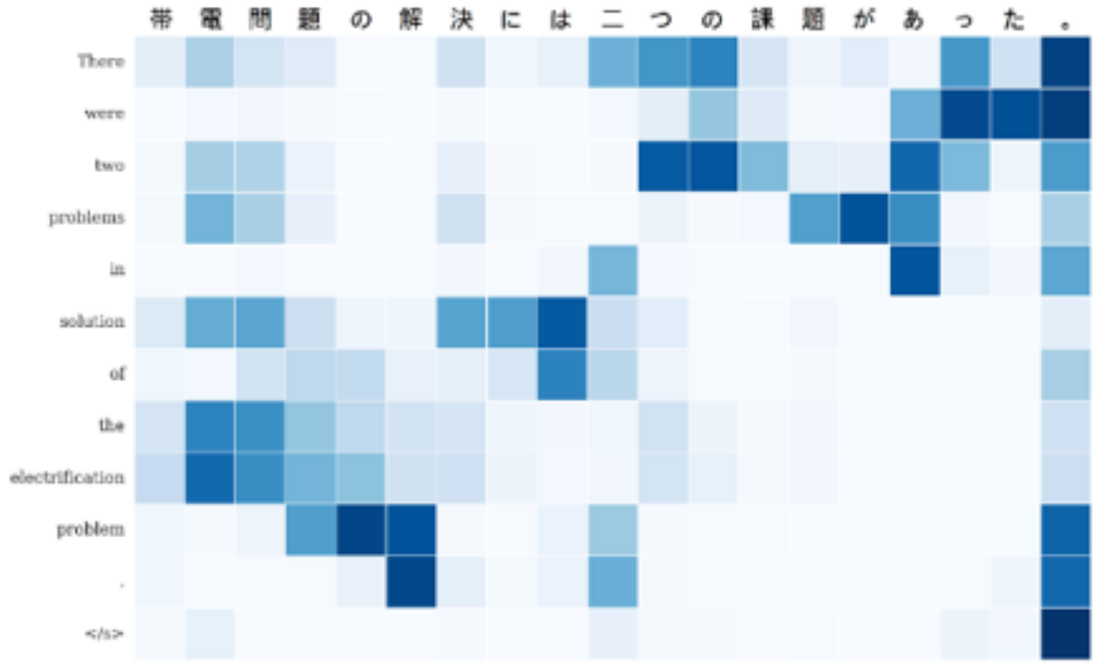
$$p(z_1,\ldots,z_n) = \text{softmax}\left(\sum_{i=1}^{n-1} \theta_{i,i+1}(z_i, z_{i+1})\right)$$

Subsequence Attention

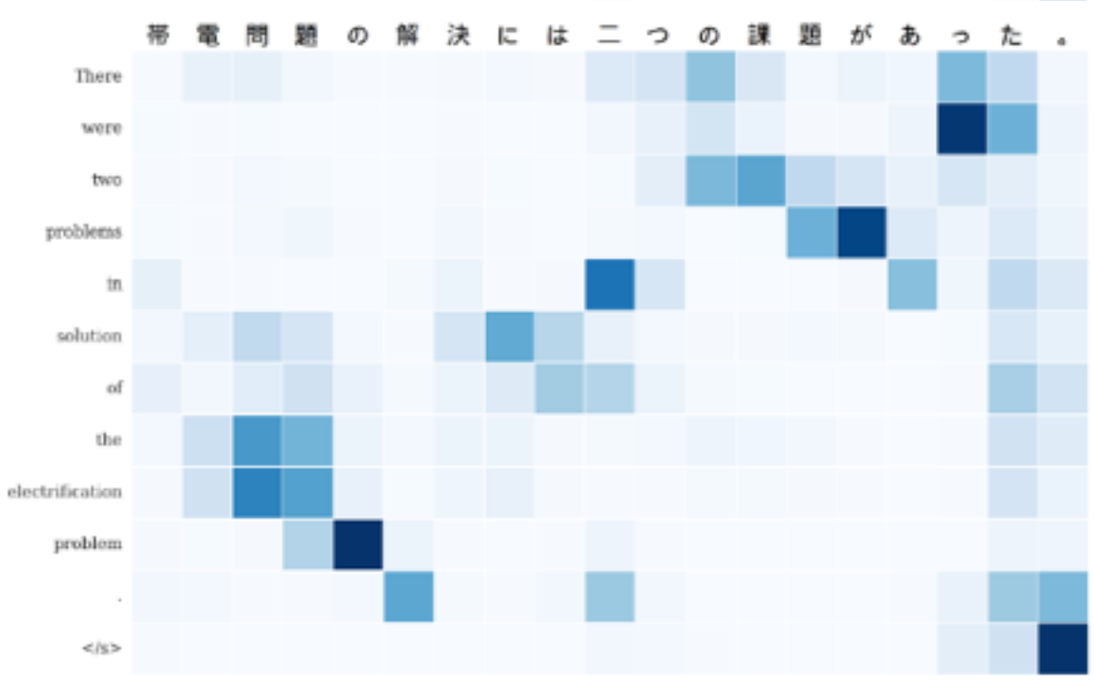
(a)



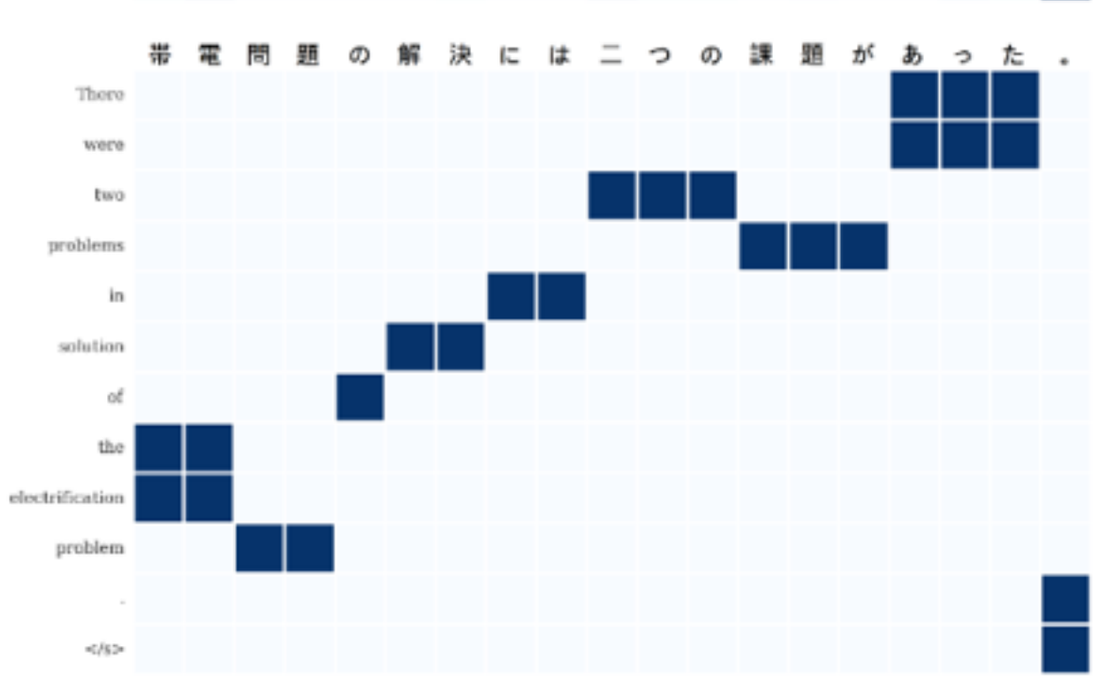
(b)



(c)



Truth



	Simple	Sigmoid	Structured
CHAR	12.6	13.1	14.6
WORD	14.1	13.8	14.3

Tree Attention

- Task:

$$(* (+ (+ 15\ 7) 1\ 8) (+ 19\ 0\ 11)) \Rightarrow ((15 + 7) + 1 + 8) * (19 + 0 + 11)$$

- Latent variables $z_{ij} = 1$ if symbol j has parent i:

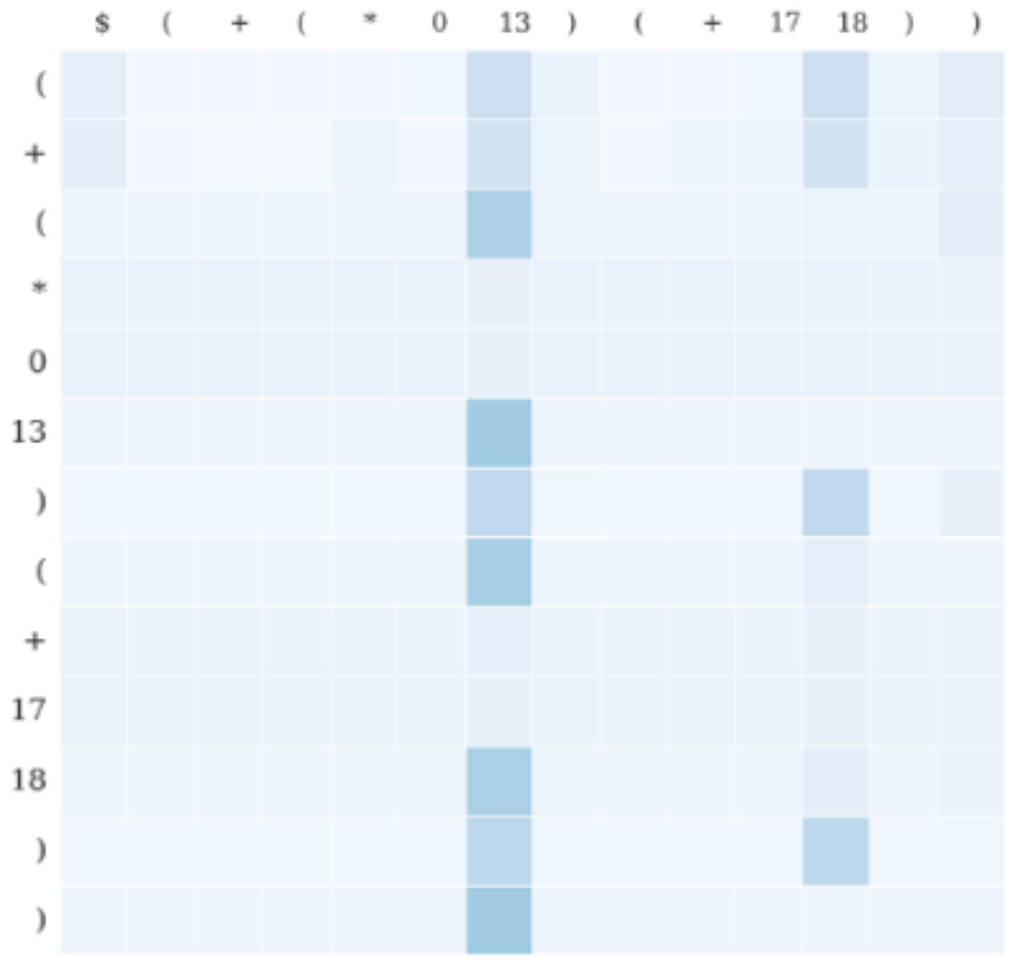
$$p(z|x,q) = \text{softmax}\left(\mathbb{1}\{z \text{ is valid}\} \sum_{i \neq j} \mathbb{1}\{z_{ij} = 1\}\theta_{ij}\right)$$

- Context vector per symbol that attends to its parent in the tree:

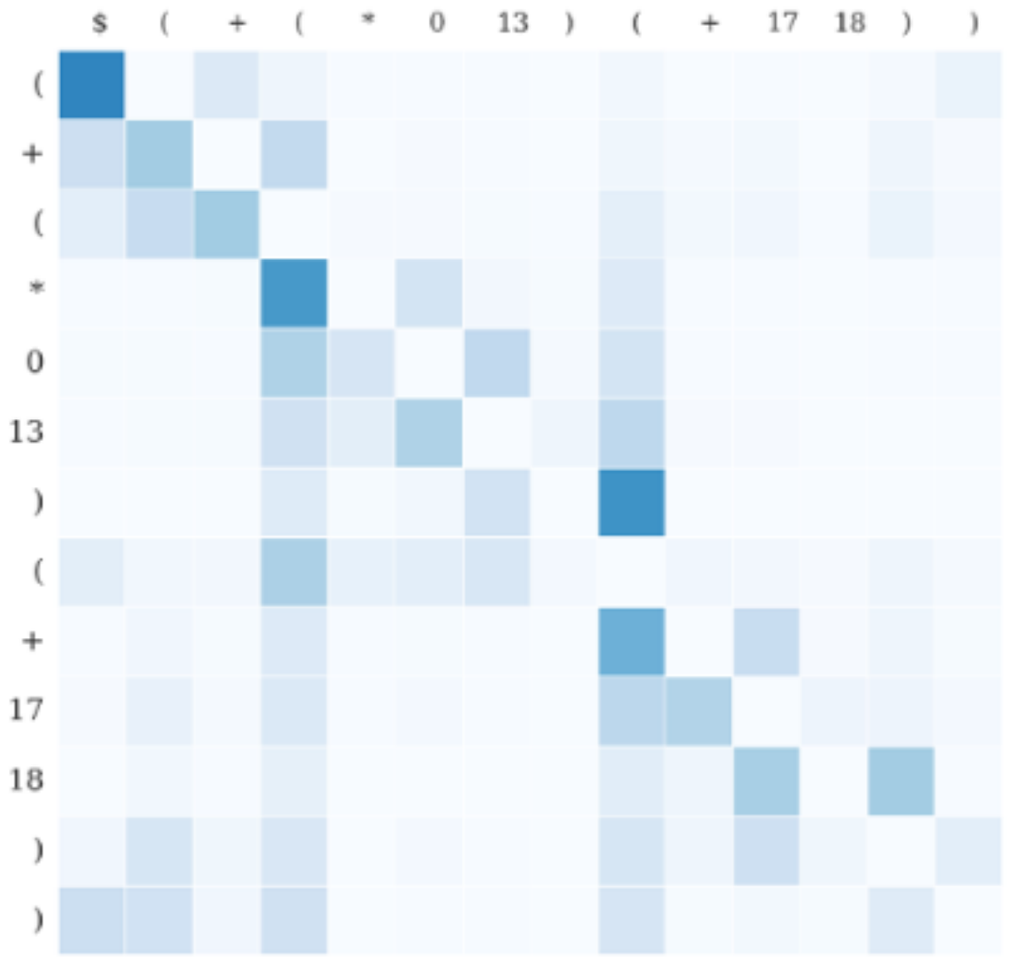
$$\mathbf{c}_j = \sum_{i=1}^{n} p(z_{ij} = 1|x,q)\mathbf{x}_i$$

- No input query in this case, since a symbol's parent doesn't depend on decoder's location.

Tree Attention



Simple



Structured

Depth	No Atten	Simple	Structured
2	7.6	87.4	99.2
3	4.1	49.6	87.0
4	2.8	23.3	64.5
5	2.1	15.0	30.8
6	1.5	8.5	18.2

Outline

- Motivating example and definition

  Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. “Neural machine translation by jointly learning to align and translate.” In International Conference on Learning Representations, 2015. arXiv:1409.0473 [cs.CL]

- Generalizations and a little theory

  Yoon Kim, Carl Denton, Luong Hoang, and Alexander M. Rush. “Structured attention networks.” In International Conference on Learning Representations, 2017. arXiv: 1702.00887 [cs.CL]

- Why attention might be better than RNNs and CNNs

  Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin. “Attention is all you need.” In 31st Conference on Neural Information Processing Systems (NIPS 2017). arXiv:1706.03762 [cs.CL]

Attention Is All You Need

- Can we replace CNNs and RNNs with attention for sequential tasks?

- Self attention: the sequence is the query, key, and value.

- Stack attention layers: output of attention layer is a sequence which is fed into the next layer.

- Attention loses positional information; must insert as additional input.

Attention Is All You Need

Regular attention: keys and values from encoder, query from decoder.

Outputs probabilities for just the next word.

All linear layers applied per position with weight sharing.

Stacked a fixed N number of times

Masked to prevent attending to words that were written later.

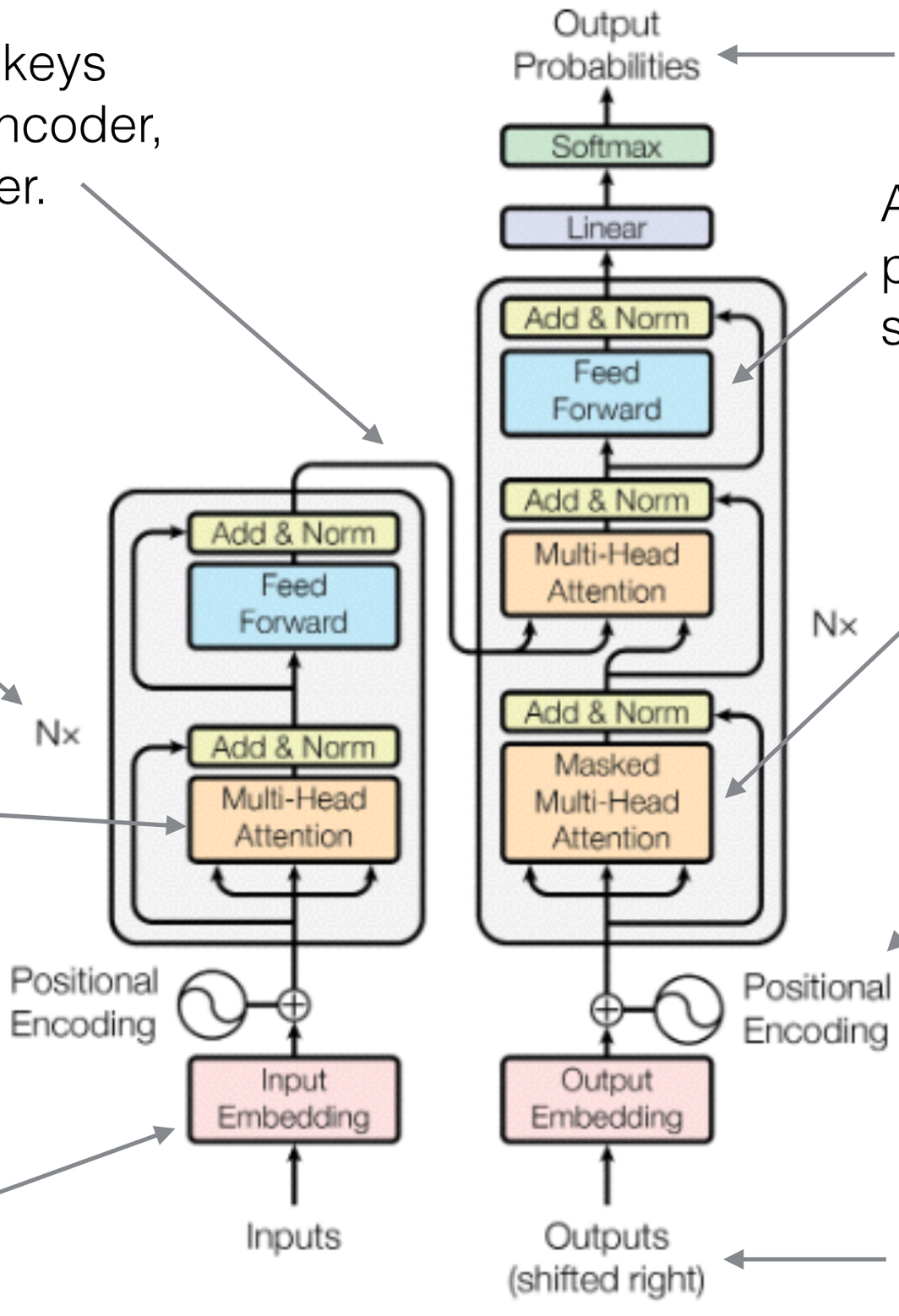
Self attention: keys, values, queries all from previous layer

Pointwise add sinusoids of different frequencies to the input features.

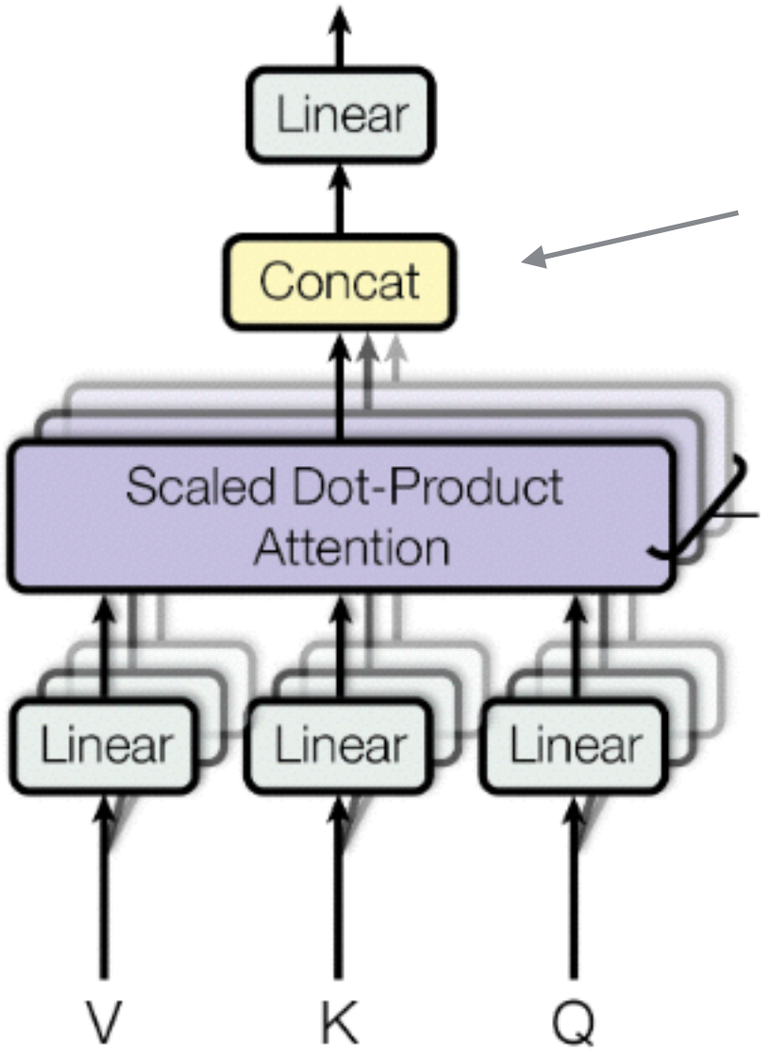
$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{\text{model}}})$$
$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{\text{model}}})$$

Input entire sequence, size is n x d_{model}

Input sequence generated so far



Multi-Head Attention



After concat, dimension is d_{model} again.

Learn linear projections into h separate d_{model}/h size vectors

Run h separate multiplicative attention steps. Scale the dot-product by $(d_{model}/h)^{1/2}$

Self Attention

- Why? Self-attention improves long-range correlations and parallelization, and sometimes complexity.

Layer Type	Complexity per Layer	Sequential Operations	Maximum Path Length
Self-Attention	$O(n^2 \cdot d)$	$O(1)$	$O(1)$
Recurrent	$O(n \cdot d^2)$	$O(n)$	$O(n)$
Convolutional	$O(k \cdot n \cdot d^2)$	$O(1)$	$O(log_k(n))$
Self-Attention (restricted)	$O(r \cdot n \cdot d)$	$O(1)$	$O(n/r)$

n: sequence length
d: representation length
k: kernel size
r: restriction size

RNNs and CNNs need a d x d matrix of weights, attention uses length d dot product.

Whole sequence attends to every position

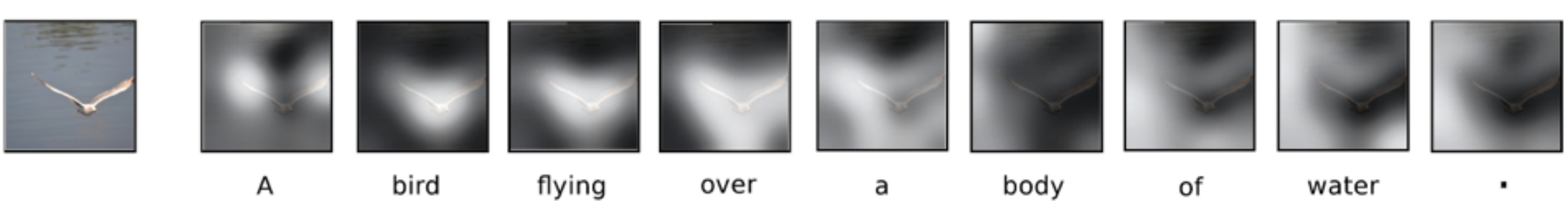
Using dilated convolutions, otherwise O(n/k)

Attention is All You Need

Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [18]	23.75			
Deep-Att + PosUnk [39]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [38]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [9]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [32]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [39]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [38]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [9]	26.36	**41.29**	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	**$3.3 \cdot 10^{18}$**	
Transformer (big)	**28.4**	**41.8**	$2.3 \cdot 10^{19}$	

Other Cool Things

- Image captioning: like translation but replace encoder with CNN. Can see where network is ‘looking’.



Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, Yoshua Bengio. “Show, attend, and tell: neural image caption generation with visual attention.” In International Conference on Machine Learning, 2015. arXiv:1502.03044 [cs.LG]

- Hard attention: sample from probability distribution instead of taking expectation value. No longer differentiable, so train as RL algorithm where choosing attention target is an action.

Summary

- Attention is an architecture-level construct for sequence analysis.

- It is essentially learned, differentiable dictionary look-up.

- More generally, it is an input-dependent, learned probability distribution for latent variables that annotate output values.

- Better long range correlation and parallelization than RNNs, often less complex.

- Produces human interpretable intermediate data.

References

- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. “Neural machine translation by jointly learning to align and translate.” In International Conference on Learning Representations, 2015. arXiv:1409.0473 [cs.CL]

- Yoon Kim, Carl Denton, Luong Hoang, and Alexander M. Rush. “Structured attention networks.” In International Conference on Learning Representations, 2017. arXiv:1702.00887 [cs.CL]

- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin. “Attention is all you need.” In 31st Conference on Neural Information Processing Systems (NIPS 2017). arXiv:1706.03762 [cs.CL]

- Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, Yoshua Bengio. “Show, attend, and tell: neural image caption generation with visual attention.” In International Conference on Machine Learning, 2015. arXiv:1502.03044 [cs.LG]

- Hard attention example: Jimmy Lei Ba, Volodymyr Mnih, Koray Kavukcuoglu. “Multiple object recognition with visual attention.” In International Conference on Learning Representations, 2015. arXiv:1412.7755 [cs.LG]

- Title page picture: https://eurovisionireland.net/2014/02/24/lithuania-which-version-of-attention-should-go-to-eurovision-2014/