Attention Mechanism in Deep Learning

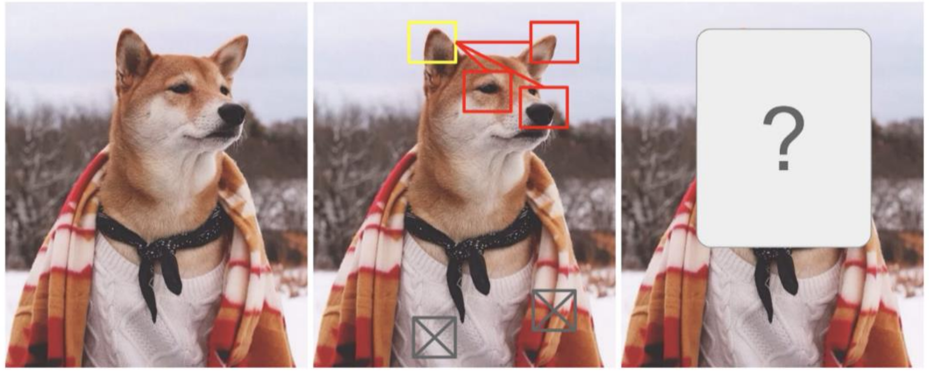
Jinjin Zhang

November 1st, 2019

Introduction

◻ Why Attention Mechanism

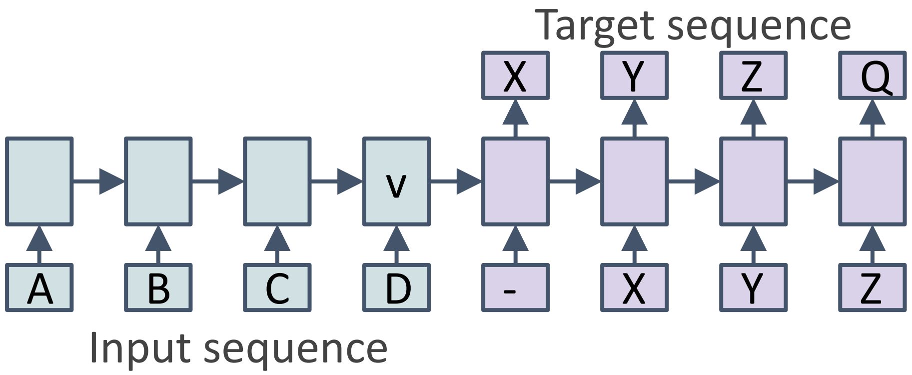
- Visual Attention
  - Focus on critical regions
- Seq2Seq[1]
  - Incapability of remembering long sentences
  - Context alignment



Source: https://lilianweng.github.io/lil-log/2018/06/24/attention-attention.html

context					target	$P(w_t \mid w_{t-1}, w_{t-2}, \ldots w_{t-5})$
the	cat	sat	on	the	mat	0.15
$w_{t\text{-}5}$	$w_{t\text{-}4}$	$w_{t\text{-}3}$	$w_{t\text{-}2}$	$w_{t\text{-}1}$	w_t	
the	cat	sat	on	the	rug	0.12
the	cat	sat	on	the	hat	0.09
the	cat	sat	on	the	dog	0.01
the	cat	sat	on	the	the	0
the	cat	sat	on	the	sat	0
the	cat	sat	on	the	robot	?
the	cat	sat	on	the	printer	?

Language Model



Seq2Seq: p(English | French)

[1] Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks." Advances in neural information processing systems. 2014.

Introduction

Attention Mechanism Timeline in Deep Learning

2014

Recurrent Model of Visual Attention(**RAM**)
Neural machine translation by jointly learning to align and translate (**Bahdanau/Additive Attention**)

2015

Show, Attend and Tell: Neural Image Caption Generation with Visual Attention
Effective Approaches to Attention-based Neural Machine Translation (**Luong/Multiplicative Attention**)

2016-2017

Long Short-Term Memory-Networks for Machine Reading(**Self-attention**)
Convolutional Sequence to Sequence Learning
Attention is all you need(**Transformer**)

2018-2019

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context

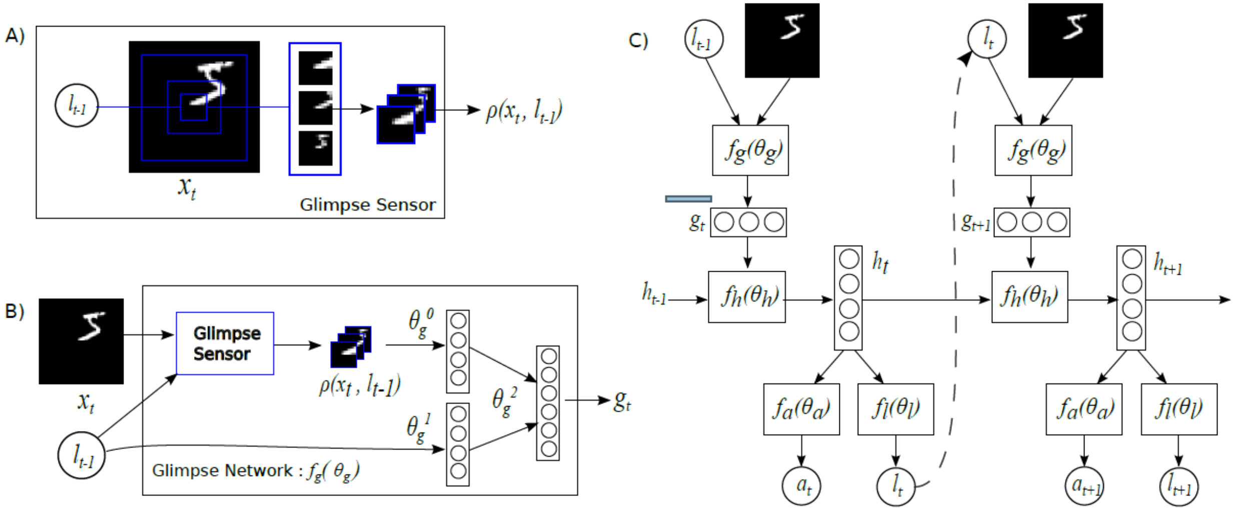
北京航空航天大学
BEIHANG UNIVERSITY

Attention Mechanism

Recurrent Model of Visual Attention

- MNIST
- Model Architecture
  - Glimpse Sensor
    - “Retina-like” representation
  - Glimpse Network
    - Contains both “what” and “where”
  - Location Network
  - Action(classification) Network

(a) Translated MNIST inputs.

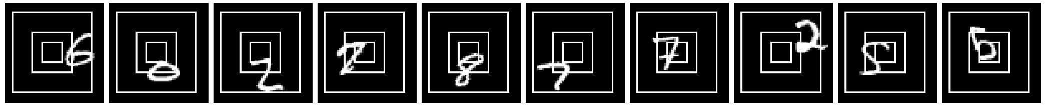


[1] Mnih, Volodymyr, Nicolas Heess, and Alex Graves. "Recurrent models of visual attention." *Advances in neural information processing systems*. 2014.

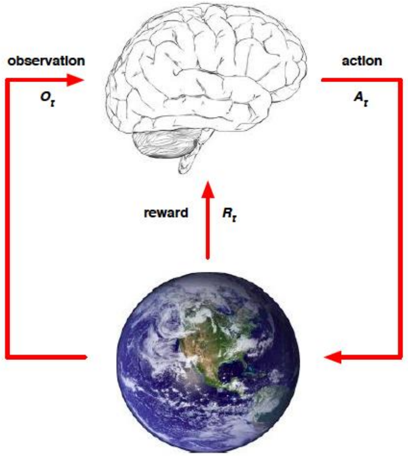
Attention Mechanism

Recurrent Model of Visual Attention

- Stochastic process
  - Gaussian distribution parameterized by location network for next location



- Reinforcement Learning
  - Partially Observable Markov Decision Process
  - Monte Carlo sampling
  - Policy gradient



$$\text{Loss} = \underbrace{\text{CrossEntropy}}_{\text{Prediction Correctness}} - \underbrace{(\text{R-b\_freeze}) * \text{loglikelihood}}_{\text{MLE}} + \underbrace{(\text{R-b})**2}_{\text{Score Function}}$$

[1] Mnih, Volodymyr, Nicolas Heess, and Alex Graves. "Recurrent models of visual attention." *Advances in neural information processing systems*. 2014.

Attention Mechanism

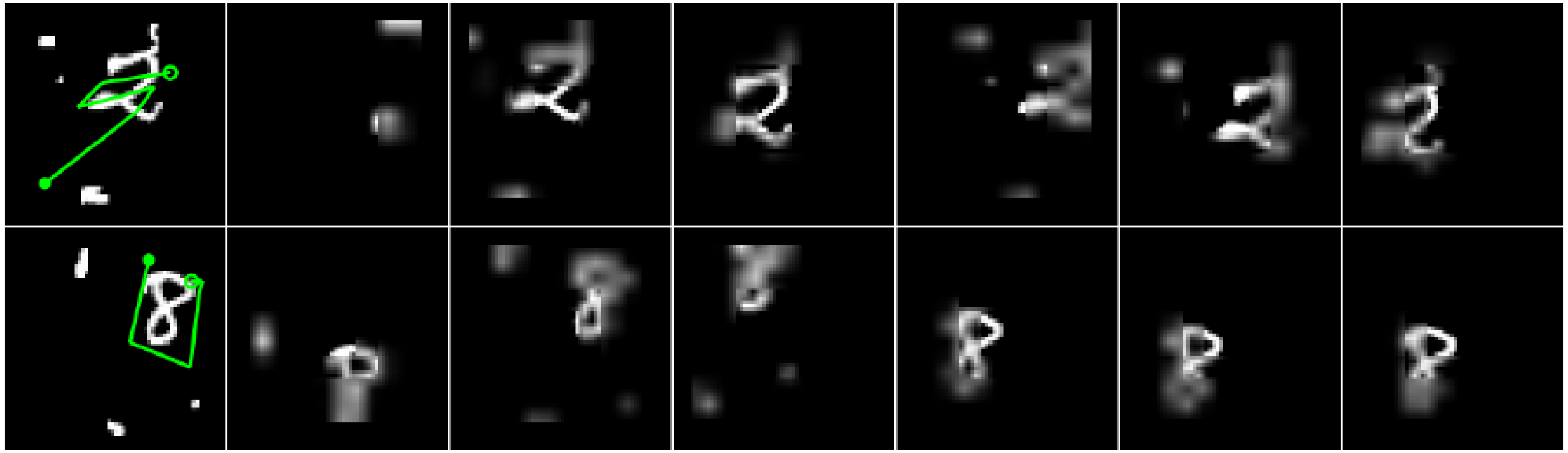
Recurrent Model of Visual Attention

(a) 60x60 Cluttered Translated MNIST

Model	Error
FC, 2 layers (64 hiddens each)	28.96%
FC, 2 layers (256 hiddens each)	13.2%
Convolutional, 2 layers	7.83%
RAM, 4 glimpses, 12×12, 3 scales	7.1%
RAM, 6 glimpses, 12×12, 3 scales	5.88%
RAM, 8 glimpses, 12×12, 3 scales	5.23%

(b) 100x100 Cluttered Translated MNIST

Model	Error
Convolutional, 2 layers	16.51%
RAM, 4 glimpses, 12×12, 4 scales	14.95%
RAM, 6 glimpses, 12×12, 4 scales	11.58%
RAM, 8 glimpses, 12×12, 4 scales	10.83%

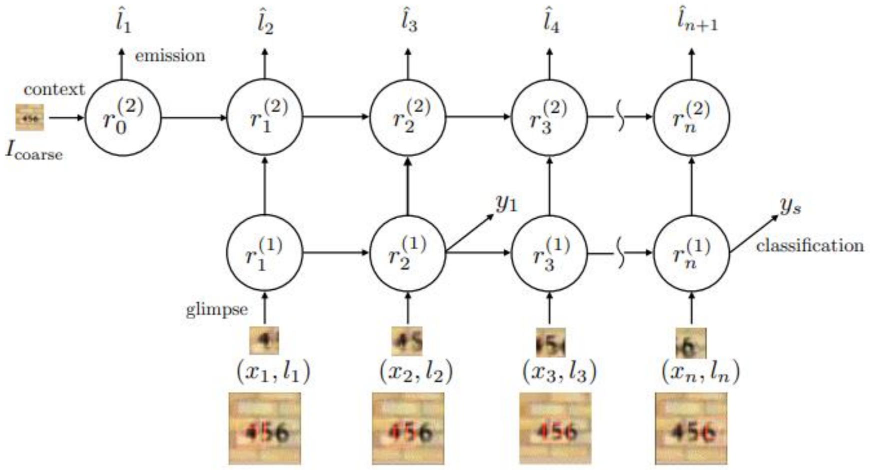


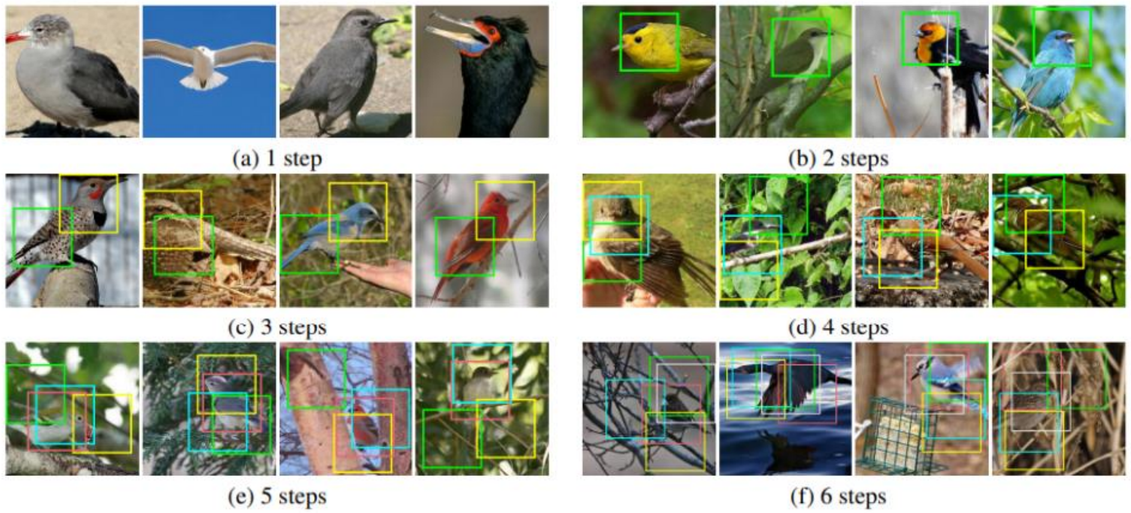
[1] Mnih, Volodymyr, Nicolas Heess, and Alex Graves. "Recurrent models of visual attention." *Advances in neural information processing systems*. 2014.

Attention Mechanism

Recurrent Model of Visual Attention

- Multi-object/sequential classification[1]
- Fine-grained recognition[2]



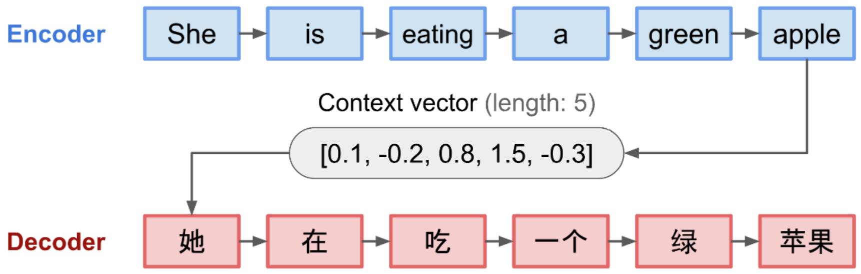


[1] Ba, Jimmy, Volodymyr Mnih, and Koray Kavukcuoglu. "Multiple object recognition with visual attention." arXiv preprint arXiv:1412.7755 (2014).
[2] Li, Zhichao, et al. "Dynamic computational time for visual attention." Proceedings of the IEEE International Conference on Computer Vision. 2017.

Attention Mechanism

Neural Machine Translation by Jointly Learning to Align and Translation

- Soft-Alignments
  - Automatically search for parts of a source sentence that are relevant to predicting a target word
- Bahdanau/Additive Attention
  - Score
  - Align
  - Context

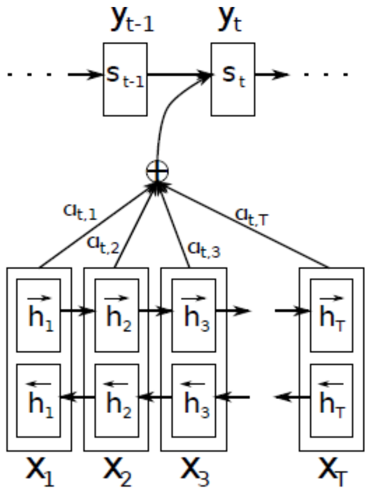


$$e_{ij} = v_a^\top \tanh\left(W_a s_{i-1} + U_a h_j\right),$$

$$\alpha_{ij} = \frac{\exp\left(e_{ij}\right)}{\sum_{k=1}^{T_x} \exp\left(e_{ik}\right)},$$

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j.$$

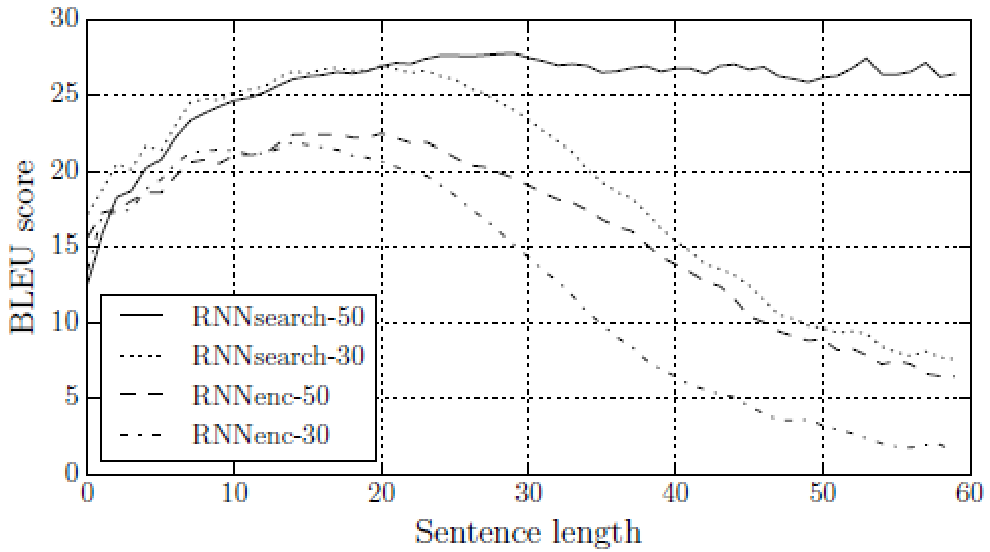
Additive Formulation

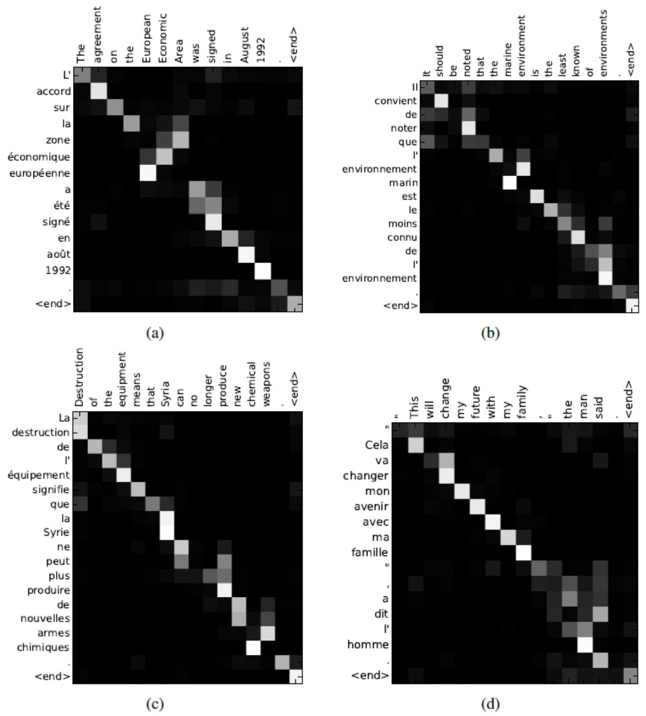


[1] Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." arXiv preprint arXiv:1409.0473 (2014).

Attention Mechanism

- Neural Machine Translation by Jointly Learning to Align and Translation





[1] Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." arXiv preprint arXiv:1409.0473 (2014).

Attention Mechanism

Effective Approaches to Attention-based Neural Machine Translation

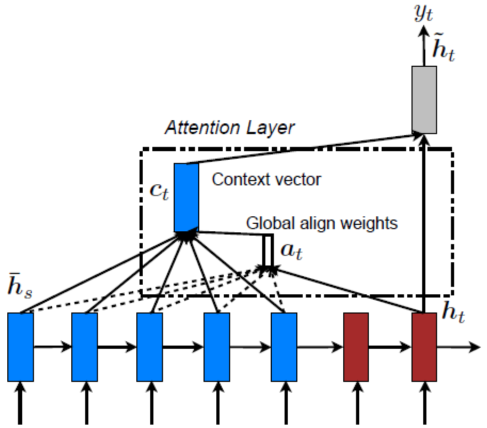
- Luong/Multiplicative Attention
  - Score(content-based) function
  - Align(location-based) function
  - Generate Context vector
- Attention-based Models
  - Global Attention
  - Local Attention

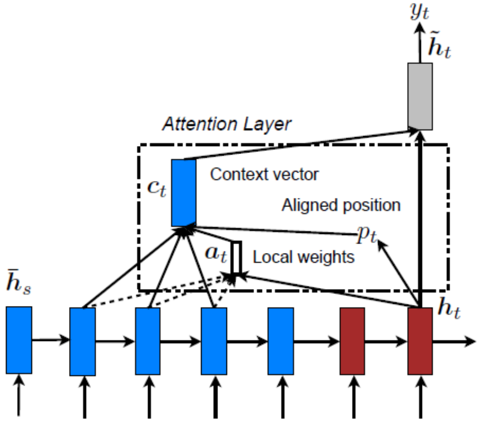
$$\mathrm{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s) = \begin{cases} \boldsymbol{h}_t^\top \bar{\boldsymbol{h}}_s & \textit{dot} \\ \boldsymbol{h}_t^\top \boldsymbol{W}_a \bar{\boldsymbol{h}}_s & \textit{general} \\ \boldsymbol{v}_a^\top \tanh\left(\boldsymbol{W}_a [\boldsymbol{h}_t; \bar{\boldsymbol{h}}_s]\right) & \textit{concat} \end{cases}$$

Multiplicative Formulation

$$\boldsymbol{a}_t = \mathrm{softmax}(\boldsymbol{W}_a \boldsymbol{h}_t) \qquad \textit{location}$$

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j .$$





[1] Luong, Minh-Thang, Hieu Pham, and Christopher D. Manning. "Effective approaches to attention-based neural machine translation." arXiv preprint arXiv:1508.04025 (2015).

Attention Mechanism

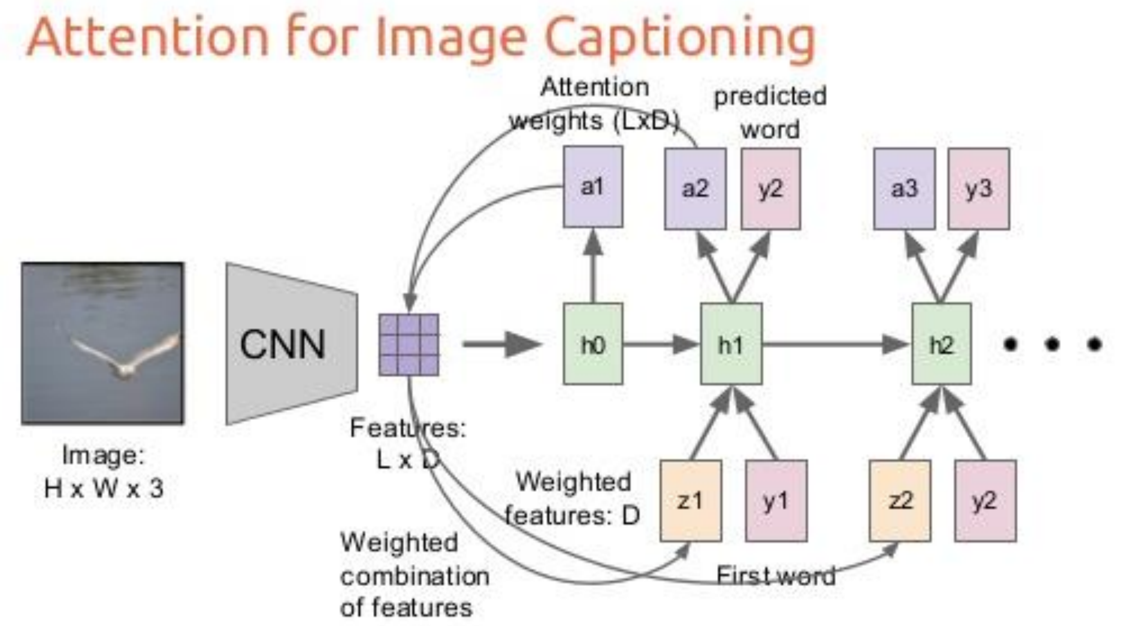
Show, Attend and Tell

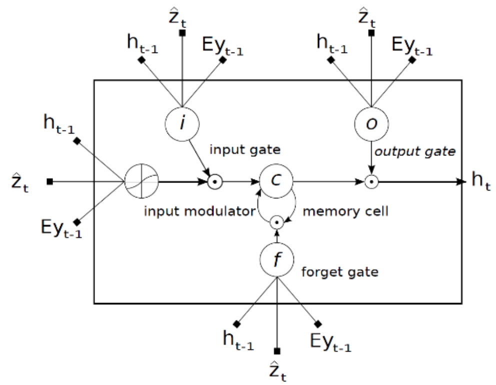
- Encoder
  - CNN features
- Decoder
  - LSTM

$$\begin{pmatrix}\mathbf{i}_t\\ \mathbf{f}_t\\ \mathbf{o}_t\\ \mathbf{g}_t\end{pmatrix} = \begin{pmatrix}\sigma\\ \sigma\\ \sigma\\ \tanh\end{pmatrix} T_{D+m+n,n}\begin{pmatrix}\mathbf{E}\mathbf{y}_{t-1}\\ \mathbf{h}_{t-1}\\ \hat{\mathbf{z}}_t\end{pmatrix}$$

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \mathbf{g}_t$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t).$$





  - Attention Mechanism

$$e_{ti} = f_{\text{att}}(\mathbf{a}_i, \mathbf{h}_{t-1}) \qquad \hat{\mathbf{z}}_t = \phi\left(\{\mathbf{a}_i\}, \{\alpha_i\}\right)$$

$$\alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{k=1}^{L}\exp(e_{tk})}. \qquad p(\mathbf{y}_t|\mathbf{a}, \mathbf{y}_1^{t-1}) \propto \exp(\mathbf{L}_o(\mathbf{E}\mathbf{y}_{t-1} + \mathbf{L}_h\mathbf{h}_t + \mathbf{L}_z\hat{\mathbf{z}}_t))$$

[1] Xu, Kelvin, et al. "Show, attend and tell: Neural image caption generation with visual attention." International conference on machine learning. 2015.

Attention Mechanism

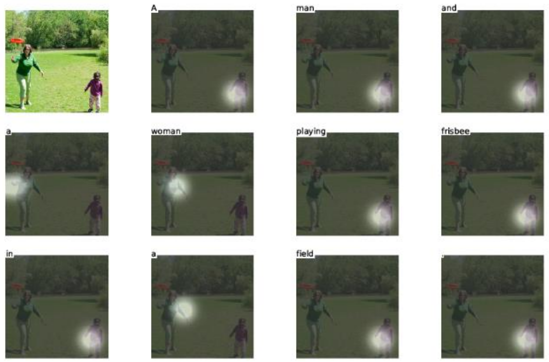
Show, Attend and Tell

- Stochastic “Hard” Attention
  - Multinoulli Distribution
  - Monte Carlo sampling
- Deterministic “Soft” Attention
  - Bahdanau/Additive Attention[2]
  - Doubly Stochastic Regularization

approximately optimizing the marginal likelihood under the attention location random variable **s**

$$L_s = \sum_s p(s \mid \mathbf{a}) \log p(\mathbf{y} \mid s, \mathbf{a})$$

$$\mathbb{E}[\mathbf{n}_t] = \mathbf{L}_o(\mathbf{E}\mathbf{y}_{t-1} + \mathbf{L}_h \mathbb{E}[\mathbf{h}_t] + \mathbf{L}_z \mathbb{E}[\hat{\mathbf{z}}_t]) \quad L_d = -\log(P(\mathbf{y}|\mathbf{x})) + \lambda \sum_i^L (1 - \sum_t^C \alpha_{ti})^2$$



Hard attention

Soft attention

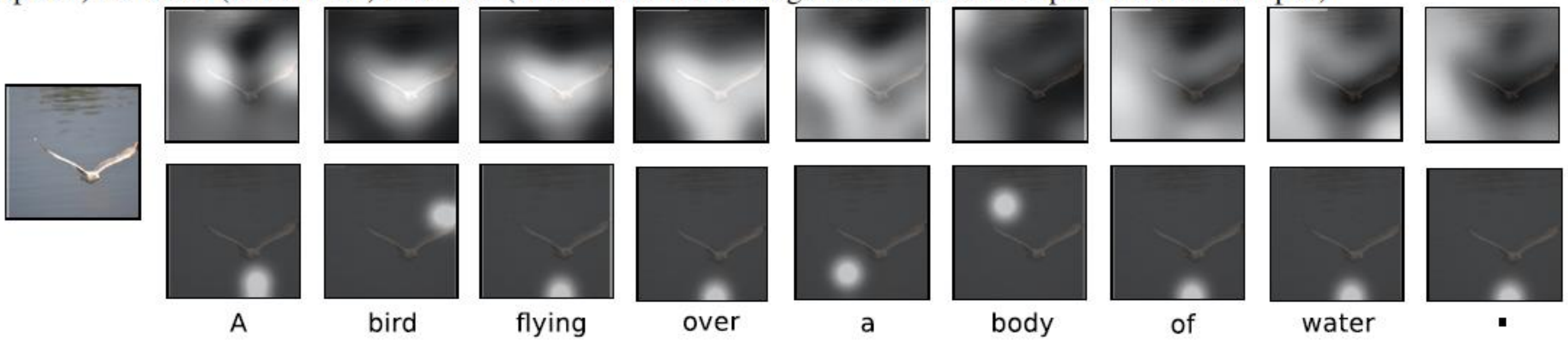
Dataset	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
Flickr8k	Google NIC(Vinyals et al., 2014)$^{\dagger\Sigma}$	63	41	27	—	—
	Log Bilinear (Kiros et al., 2014a)$^{\circ}$	65.6	42.4	27.7	17.7	17.31
	Soft-Attention	**67**	44.8	29.9	19.5	18.93
	Hard-Attention	**67**	**45.7**	**31.4**	**21.3**	**20.30**
Flickr30k	Google NIC$^{\dagger\circ\Sigma}$	66.3	42.3	27.7	18.3	—
	Log Bilinear	60.0	38	25.4	17.1	16.88
	Soft-Attention	66.7	43.4	28.8	19.1	**18.49**
	Hard-Attention	**66.9**	**43.9**	**29.6**	**19.9**	18.46
COCO	CMU/MS Research (Chen & Zitnick, 2014)^{a}	—	—	—	—	20.41
	MS Research (Fang et al., 2014)$^{\dagger a}$	—	—	—	—	20.71
	BRNN (Karpathy & Li, 2014)$^{\circ}$	64.2	45.1	30.4	20.3	—
	Google NIC$^{\dagger\circ\Sigma}$	66.6	46.1	32.9	24.6	—
	Log Bilinear$^{\circ}$	70.8	48.9	34.4	24.3	20.03
	Soft-Attention	70.7	49.2	34.4	24.3	**23.90**
	Hard-Attention	**71.8**	**50.4**	**35.7**	**25.0**	23.04

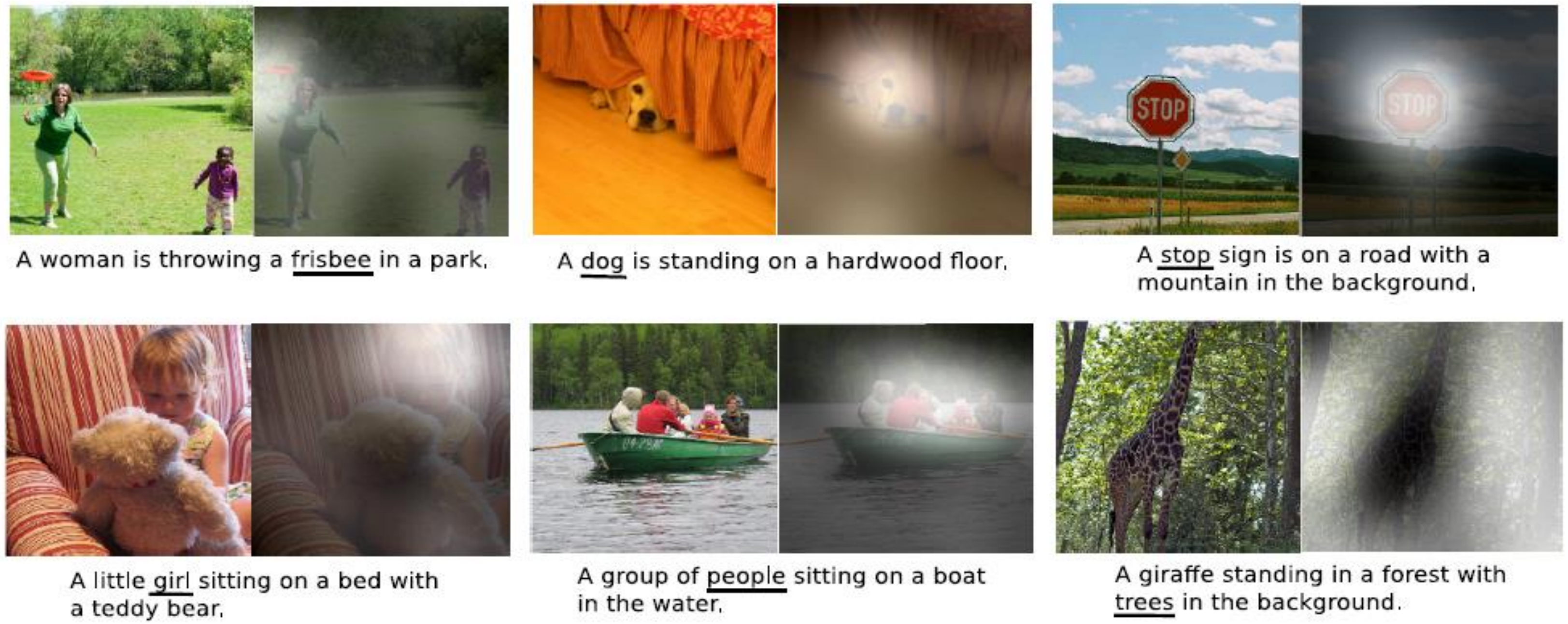
[1] Xu, Kelvin, et al. "Show, attend and tell: Neural image caption generation with visual attention." International conference on machine learning. 2015.
[2] Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." arXiv preprint arXiv:1409.0473 (2014).

北京航空航天大学 BEIHANG UNIVERSITY

Attention Mechanism

Figure 2. Attention over time. As the model generates each word, its attention changes to reflect the relevant parts of the image. "soft" (top row) vs "hard" (bottom row) attention. (Note that both models generated the same captions in this example.)

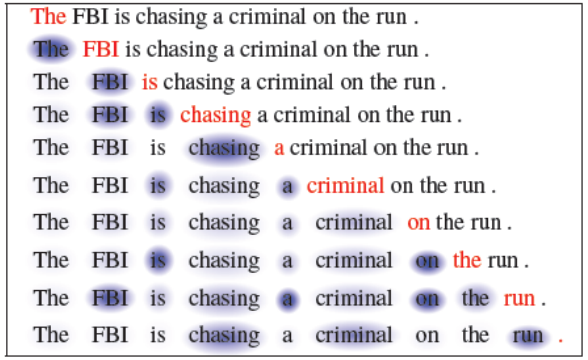


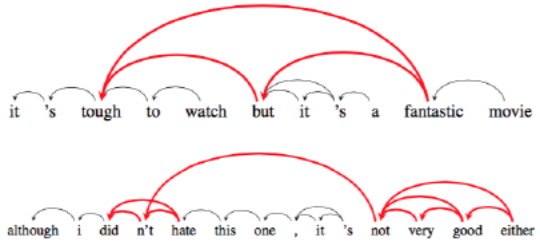


[1] Xu, Kelvin, et al. "Show, attend and tell: Neural image caption generation with visual attention." International conference on machine learning. 2015.

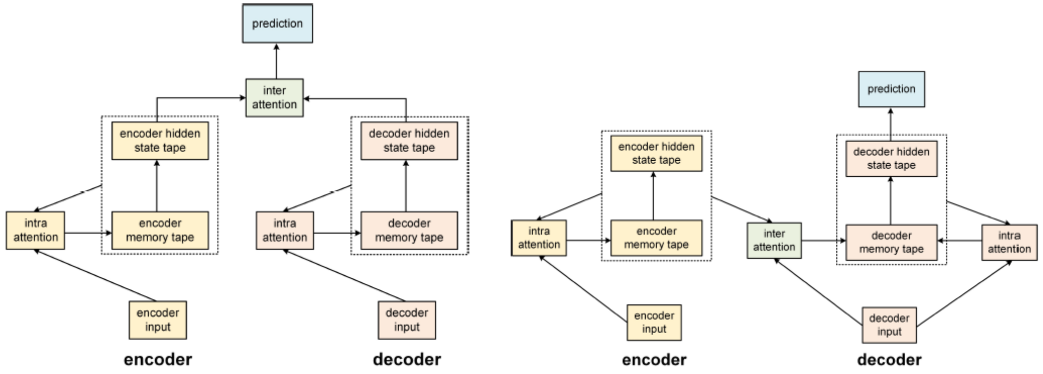
Attention Mechanism

- Long Short-Term Memory-Networks for Machine Reading
  - Long Short-Term Memory-Networks
    - Encoder
      - Self-attention/intra-attention





    - Decoder
      - Both intra- and inter-attention



(a) Decoder with shallow attention fusion. (b) Decoder with deep attention fusion.

[1] Cheng, Jianpeng, Li Dong, and Mirella Lapata. "Long short-term memory-networks for machine reading." arXiv preprint arXiv:1601.06733 (2016).

北京航空航天大學 BEIHANG UNIVERSITY

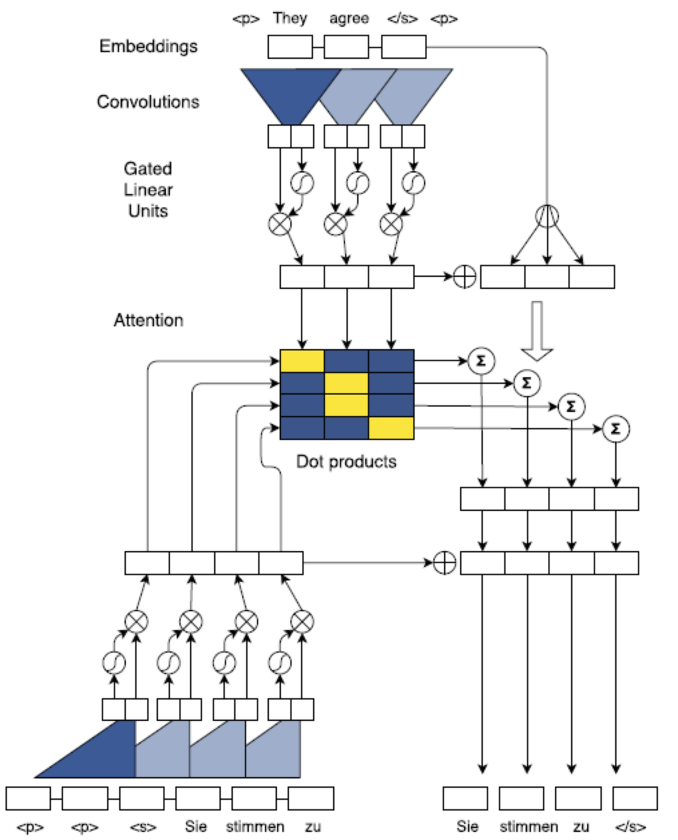
Attention Mechanism

Convolutional Sequence to Sequence Learning

- Problems with LSTM
  - Sequential computation inhibits parallelization
- CNN: Wavenet[1], ByteNet[2], ConvS2S[3]
- ConvS2S Model Architecture
  - Convolutional block
  - Position embeddings
  - Multi-step attention(for each decoder layer)

$$d_i^l = W_d^l h_i^l + b_d^l + g_i$$

$$a_{ij}^l = \frac{\exp\left(d_i^l \cdot z_j^u\right)}{\sum_{t=1}^{m} \exp\left(d_i^l \cdot z_t^u\right)} \qquad c_i^l = \sum_{j=1}^{m} a_{ij}^l \left(z_j^u + e_j\right)$$



[1] Oord, Aaron van den, et al. "Wavenet: A generative model for raw audio." arXiv preprint arXiv:1609.03499 (2016).
[2] Kalchbrenner, Nal, et al. "Neural machine translation in linear time." arXiv preprint arXiv:1610.10099 (2016).
[3] Gehring, Jonas, et al. "Convolutional sequence to sequence learning." Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, 2017.

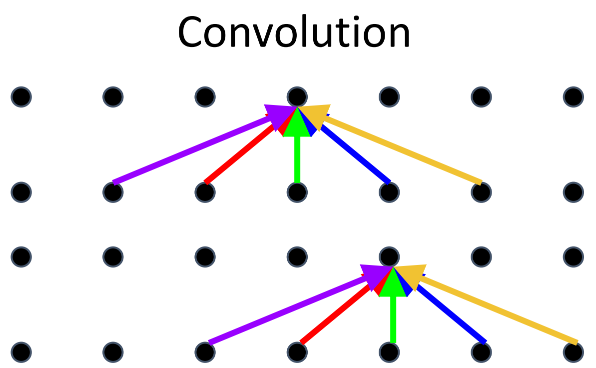
Attention Mechanism

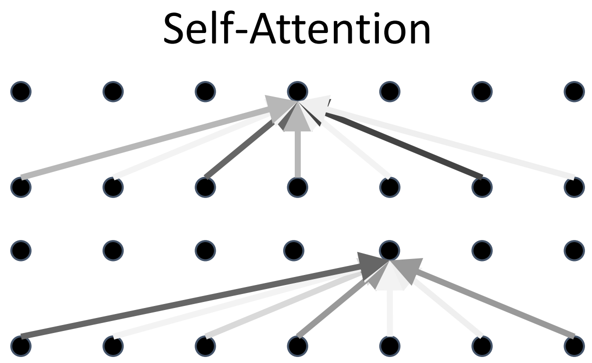
Attention Is All You Need

- Comparison

	Advantages	Problems
RNN based	variable-length representations core of seq2seq (with attention)	the sequentiality prohibits parallelization fixed-size context and hard to model hierarchical-alike domains
CNN based	trivial to parallelize (per layer) exploits local dependencies	left-padding for text and fixed width kernels difficult to learn dependencies between distant positions

- Transformer
  - Self-Attention



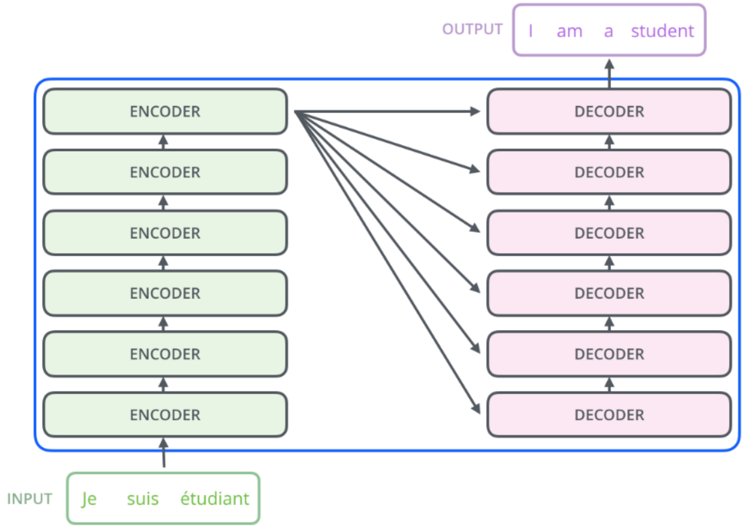


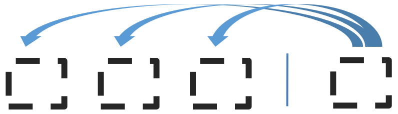
[1] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems. 2017.

Attention Mechanism

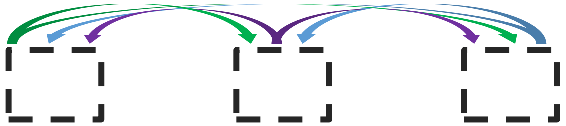
Attention Is All You Need

- Model Architecture
  - Multi-Head Attention
  - Position-wise Feed-Forward Networks
  - Positional Encoding

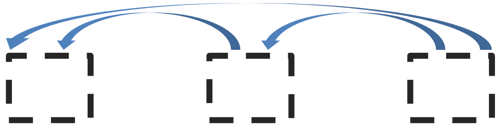




Encoder-Decoder Attention



Encoder Self-Attention



MaskedDecoder Self-Attention

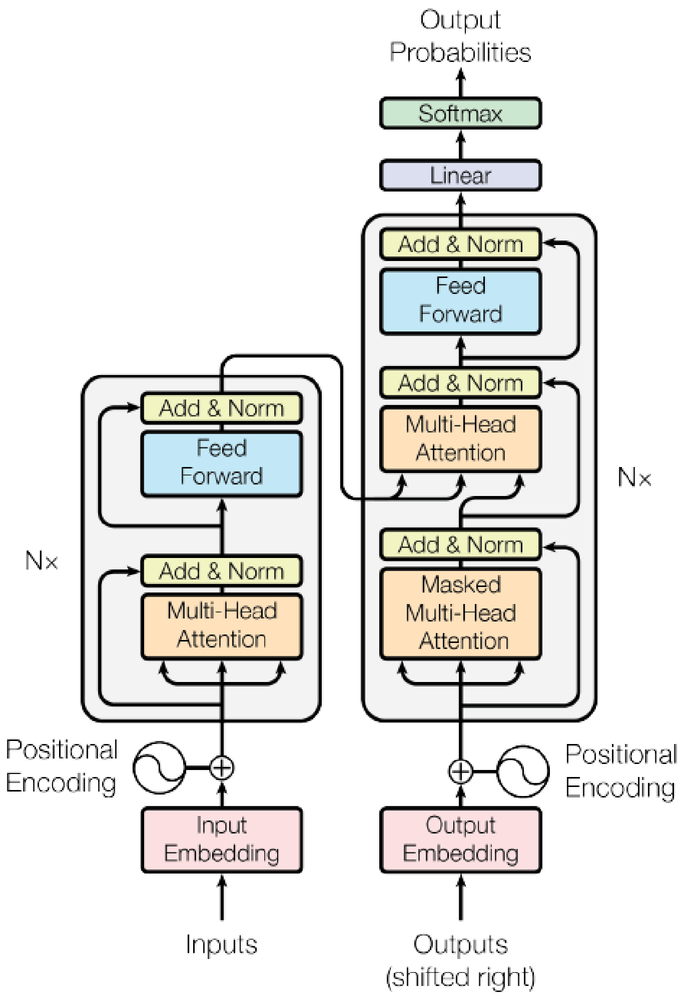


Figure 1: The Transformer - model architecture.

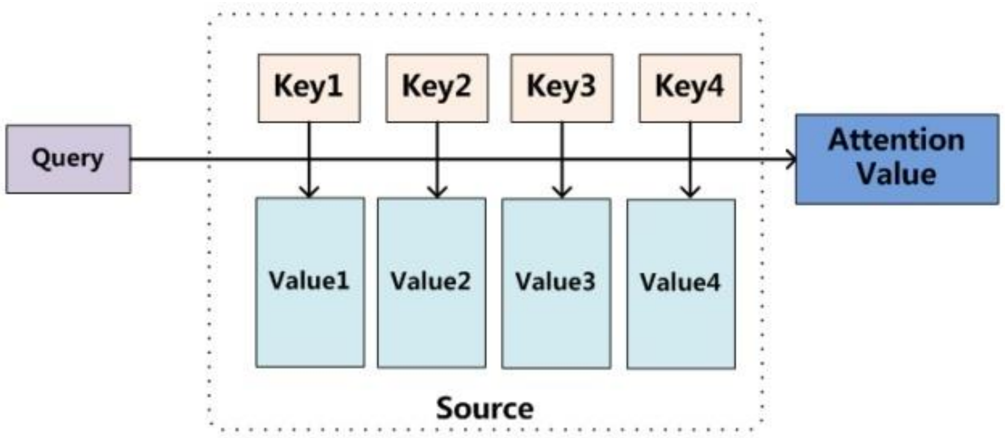
[1] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems. 2017.

Attention Mechanism

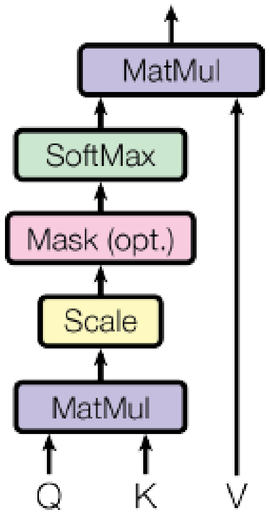
Attention Is All You Need

- Scaled Dot-Product Attention
  - Memory based[2]

$$\mathrm{Attention}(Q,K,V)=\mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$



Scaled Dot-Product Attention



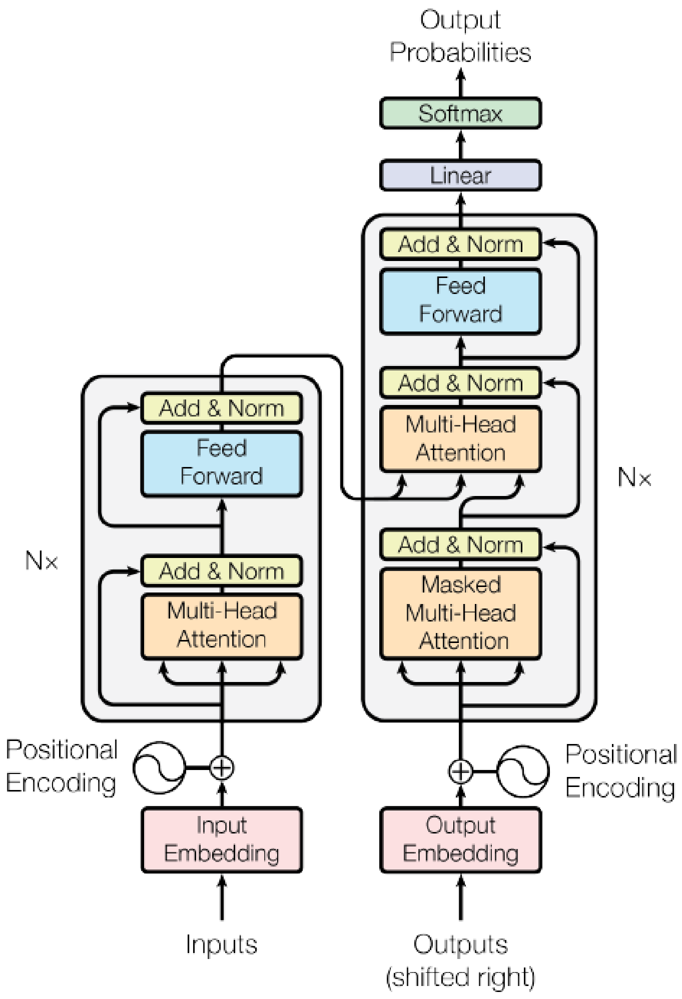


Figure 1: The Transformer - model architecture.

[1] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems. 2017.
[2] Sukhbaatar, Sainbayar, Jason Weston, and Rob Fergus. "End-to-end memory networks." Advances in neural information processing systems. 2015.

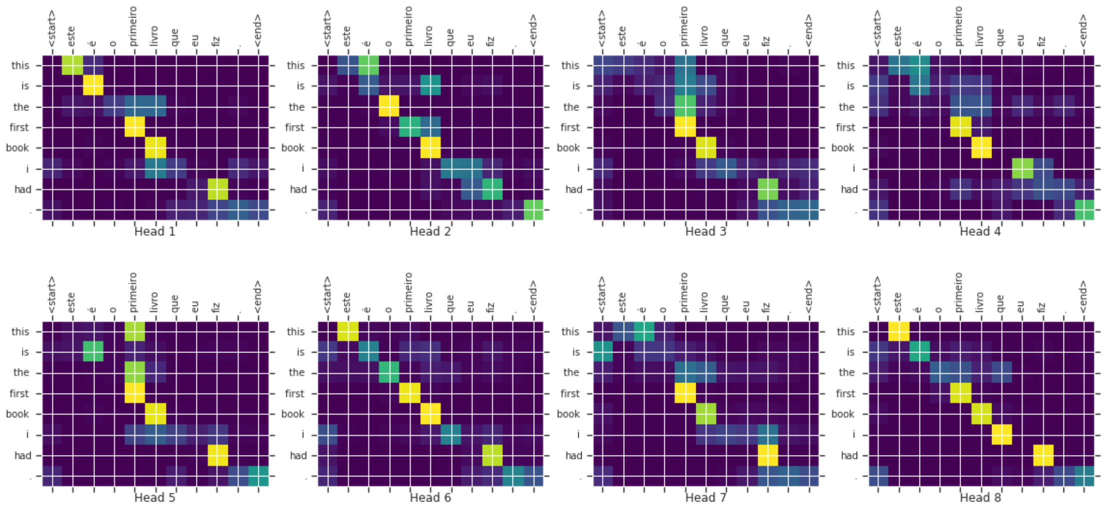
Attention Mechanism

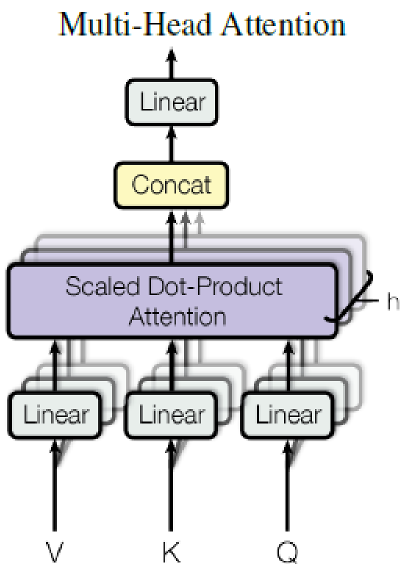
Attention Is All You Need

- Multi-Head Attention
  - Based on Scaled Dot-Product Attention

$$\mathrm{MultiHead}(Q, K, V) = \mathrm{Concat}(\mathrm{head}_1, ..., \mathrm{head_h})W^O$$

$$\text{where } \mathrm{head_i} = \mathrm{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$





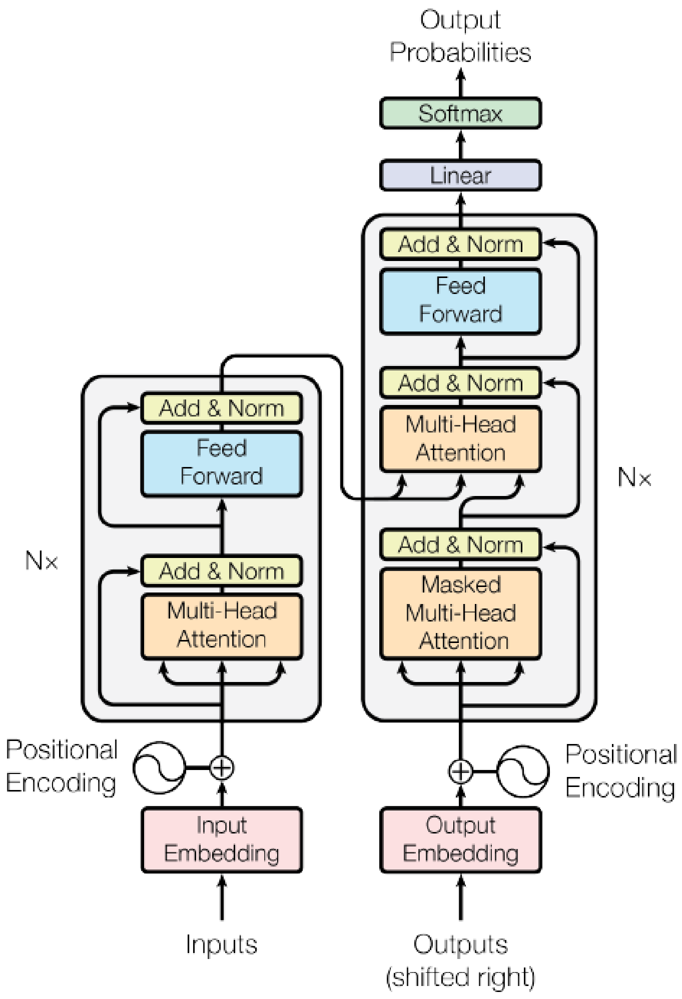


Figure 1: The Transformer - model architecture.

[1] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems. 2017.

Attention Mechanism

Attention Is All You Need

- Position-wise Feed-Forward Networks

$$\mathrm{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$$

- Residuals and Layer-Norm
- Position Encoding
  - Representing The Order of The Sequence
  - Linear function (PE_{pos+k} and PE_{pos})

$$\sin(\alpha+\beta) = \sin\alpha\cos\beta + \cos\alpha\sin\beta$$
$$\cos(\alpha+\beta) = \cos\alpha\cos\beta - \sin\alpha\sin\beta$$

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

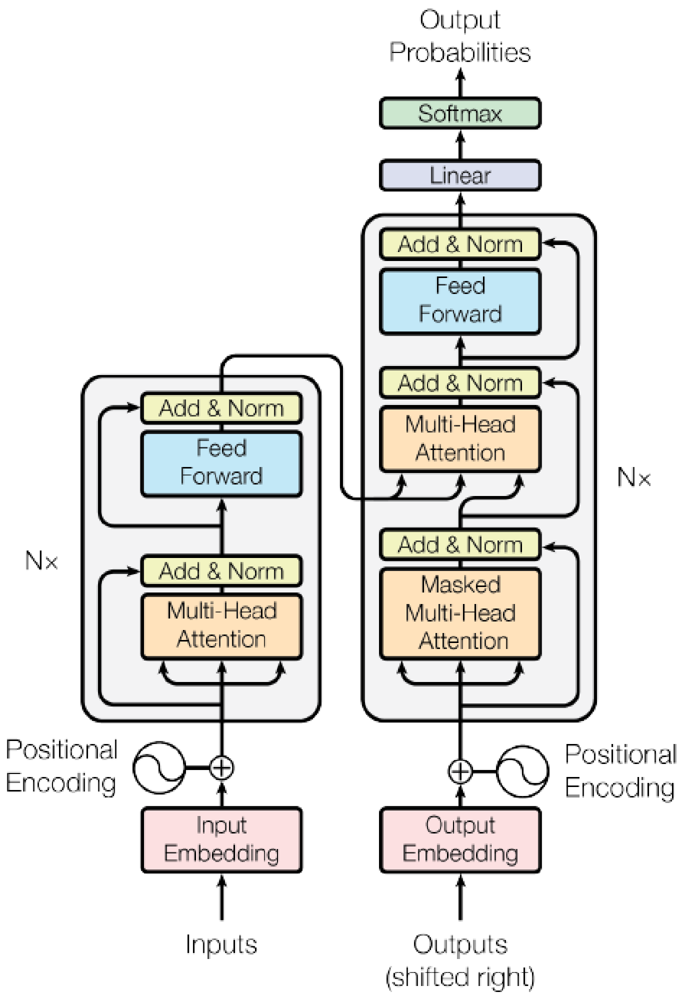


Figure 1: The Transformer - model architecture.

[1] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems. 2017.

Attention Mechanism

Attention Is All You Need

- Conclusion

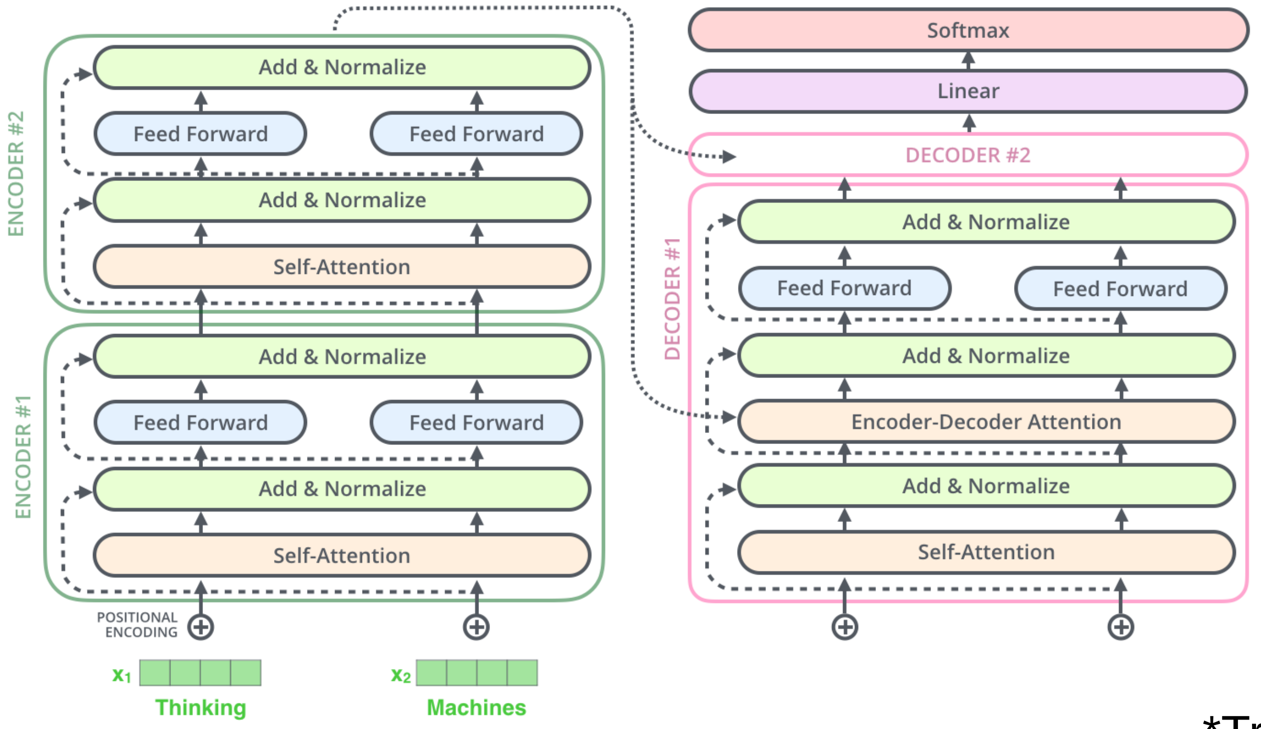


Table 1: Maximum path lengths, per-layer complexity and minimum number of sequential operations for different layer types. n is the sequence length, d is the representation dimension, k is the kernel size of convolutions and r the size of the neighborhood in restricted self-attention.

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)	$O(r \cdot n \cdot d)$	$O(1)$	$O(n/r)$

Table 2: The Transformer achieves better BLEU scores than previous state-of-the-art models on the English-to-German and English-to-French newstest2014 tests at a fraction of the training cost.

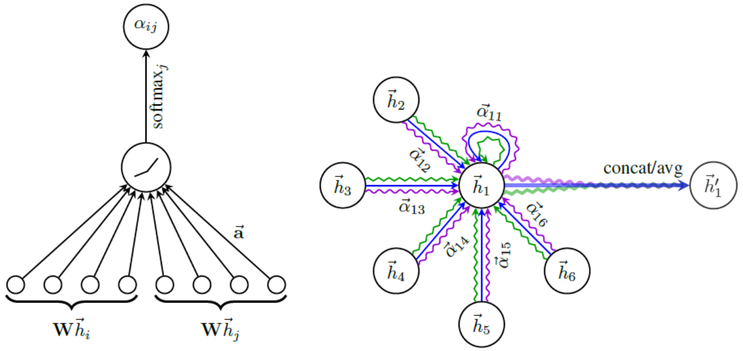
Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [18]	23.75			
Deep-Att + PosUnk [39]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [38]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [9]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [32]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
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	$\mathbf{3.3 \cdot 10^{18}}$	
Transformer (big)	**28.4**	**41.8**	$2.3 \cdot 10^{19}$	

*Transformer models trained >3x faster than the others.

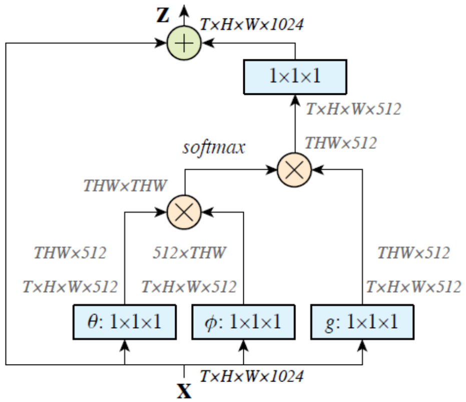
[1] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems. 2017.

Attention Mechanism

Attention Is All You Need



Graph Attention Networks[1]



Non-local Neural Networks[2]

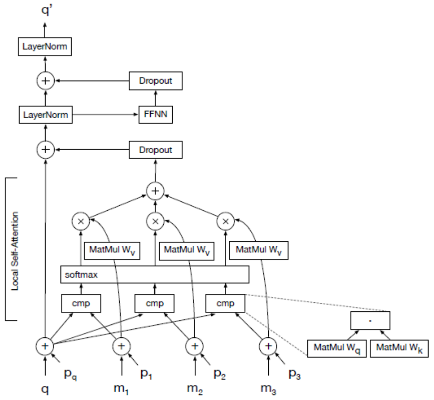
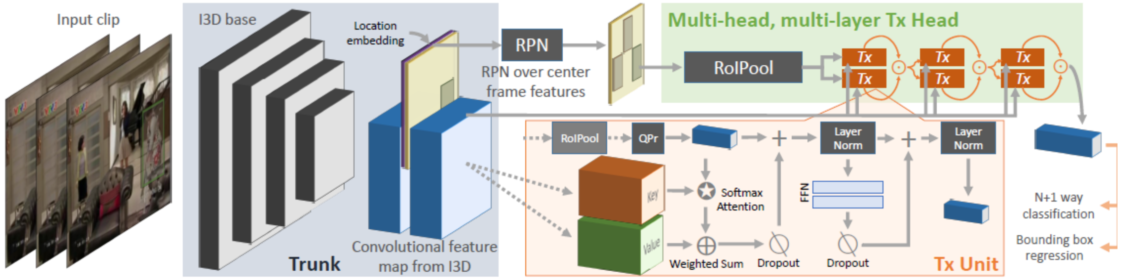
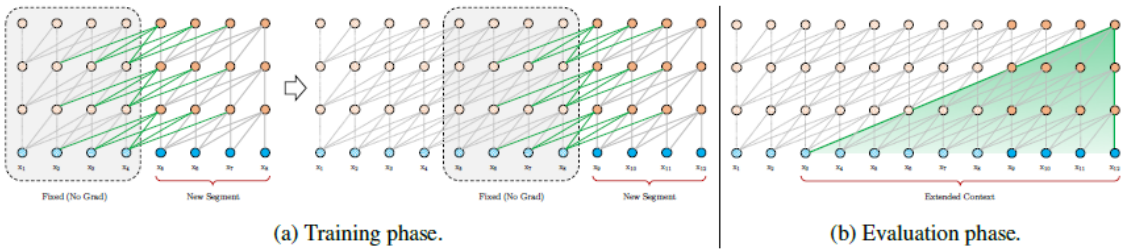


Image Transformer[3]



Video Action Transformer Network[4]



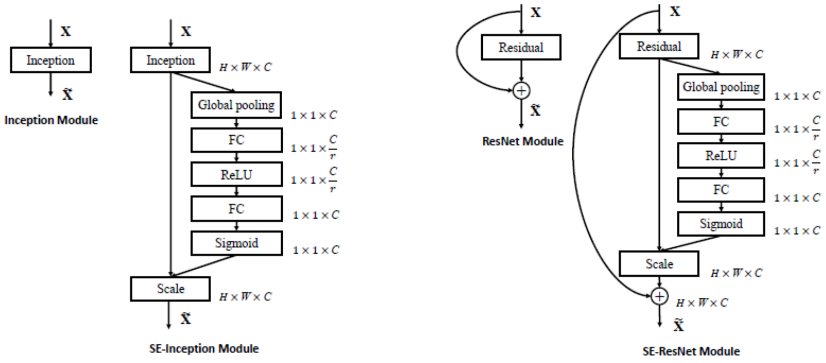
Transformer-XL[5]

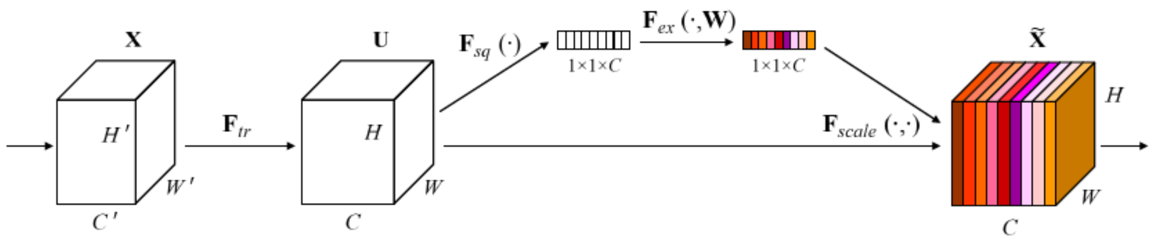
[1] Veličković, Petar, et al. "Graph attention networks." arXiv preprint arXiv:1710.10903 (2017).
[2] Wang, Xiaolong, et al. "Non-local neural networks." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.
[3] Parmar, Niki, et al. "Image transformer." arXiv preprint arXiv:1802.05751 (2018).
[4] Girdhar, Rohit, et al. "Video action transformer network." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019.
[5] Dai, Zihang, et al. "Transformer-xl: Attentive language models beyond a fixed-length context." arXiv preprint arXiv:1901.02860 (2019).

Attention Mechanism

Squeeze-and-Excitation Networks

- Squeeze-and-Excitation Block(Channel-wise Attention)



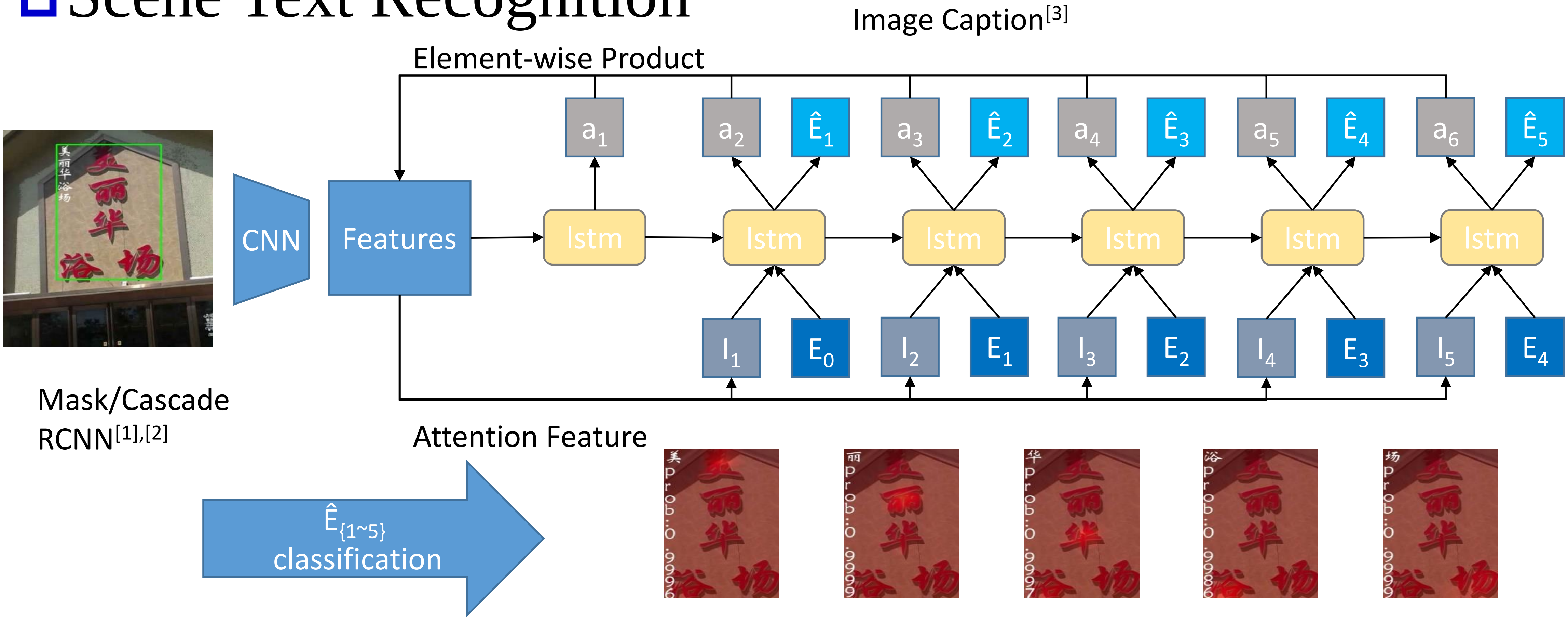


	original		re-implementation			SENet		
	top-1 err.	top-5 err.	top-1 err.	top-5 err.	GFLOPs	top-1 err.	top-5 err.	GFLOPs
ResNet-50 [13]	24.7	7.8	24.80	7.48	3.86	$23.29_{(1.51)}$	$6.62_{(0.86)}$	3.87
ResNet-101 [13]	23.6	7.1	23.17	6.52	7.58	$22.38_{(0.79)}$	$6.07_{(0.45)}$	7.60
ResNet-152 [13]	23.0	6.7	22.42	6.34	11.30	$21.57_{(0.85)}$	$5.73_{(0.61)}$	11.32
ResNeXt-50 [19]	22.2	-	22.11	5.90	4.24	$21.10_{(1.01)}$	$5.49_{(0.41)}$	4.25
ResNeXt-101 [19]	21.2	5.6	21.18	5.57	7.99	$20.70_{(0.48)}$	$5.01_{(0.56)}$	8.00
VGG-16 [11]	-	-	27.02	8.81	15.47	$25.22_{(1.80)}$	$7.70_{(1.11)}$	15.48
BN-Inception [6]	25.2	7.82	25.38	7.89	2.03	$24.23_{(1.15)}$	$7.14_{(0.75)}$	2.04
Inception-ResNet-v2 [21]	$19.9^{\dagger}$	$4.9^{\dagger}$	20.37	5.21	11.75	$19.80_{(0.57)}$	$4.79_{(0.42)}$	11.76

[1] Hu, Jie, Li Shen, and Gang Sun. "Squeeze-and-excitation networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

Application

Scene Text Recognition



[1] He, Kaiming, et al. "Mask r-cnn." Proceedings of the IEEE international conference on computer vision. 2017.
[2] Cai, Zhaowei, and Nuno Vasconcelos. "Cascade r-cnn: Delving into high quality object detection." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.
[3] Xu, Kelvin, et al. "Show, attend and tell: Neural image caption generation with visual attention." International conference on machine learning. 2015.

Conclusion and Discussion

Conclusion

- Attention Mechanism Framework
  - Attention Mechanism
    - Soft attention(End2end)
    - Hard attention(Reinforcement Learning)
    - Self attention
  - Context Learning
    - Score function: measure similarity
    - Alignment function: select relevant context(semantics and position)
    - Context Vector

Name	Alignment score function	Citation
Content-base attention	$\text{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \text{cosine}[\boldsymbol{s}_t, \boldsymbol{h}_i]$	Graves2014
Additive(*)	$\text{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \mathbf{v}_a^\top \tanh(\mathbf{W}_a[\boldsymbol{s}_t; \boldsymbol{h}_i])$	Bahdanau2015
Location-Base	$\alpha_{t,i} = \text{softmax}(\mathbf{W}_a \boldsymbol{s}_t)$ Note: This simplifies the softmax alignment to only depend on the target position.	Luong2015
General	$\text{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \boldsymbol{s}_t^\top \mathbf{W}_a \boldsymbol{h}_i$ where $\mathbf{W}_a$ is a trainable weight matrix in the attention layer.	Luong2015
Dot-Product	$\text{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \boldsymbol{s}_t^\top \boldsymbol{h}_i$	Luong2015
Scaled Dot-Product(^)	$\text{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \frac{\boldsymbol{s}_t^\top \boldsymbol{h}_i}{\sqrt{n}}$ Note: very similar to the dot-product attention except for a scaling factor; where n is the dimension of the source hidden state.	Vaswani2017

(*) Referred to as “concat” in Luong, et al., 2015 and as “additive attention” in Vaswani, et al., 2017.
(^) It adds a scaling factor $1/\sqrt{n}$, motivated by the concern when the input is large, the softmax function may have an extremely small gradient, hard for efficient learning.

Source： https://lilianweng.github.io/lil-log/2018/06/24/attention-attention.html

Conclusion and Discussion

Conclusion

- Attention Mechanism in Deep Learning
  - Weighted summation
  - Unsupervised or Semi-supervised Context Learning

- Application
  - Sequence Transduction
  - Generative Models
  - Image Caption, Action Recognition
  - Relation Reasoning