Attention on Attention for Image Captioning

Xavier Pleimling

VIRGINIA TECH

Background and Motivation

Image Captioning

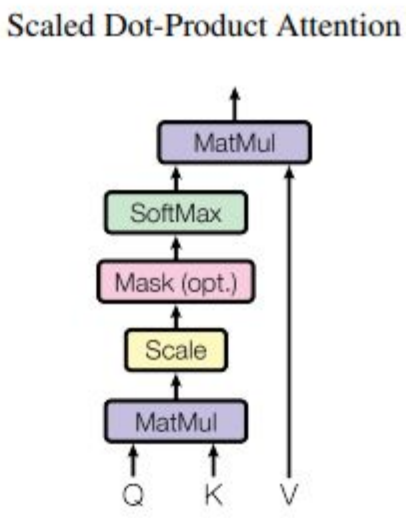
https://upload.wikimedia.org/wikipedia/commons/1/10/Tursiops_truncatus_01.jpg

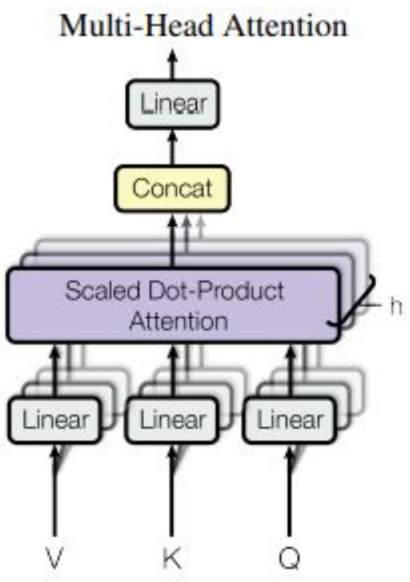
Image Captioning

A **bottlenose dolphin** riding on an **ocean wave**

https://upload.wikimedia.org/wikipedia/commons/1/10/Tursiops_truncatus_01.jpg

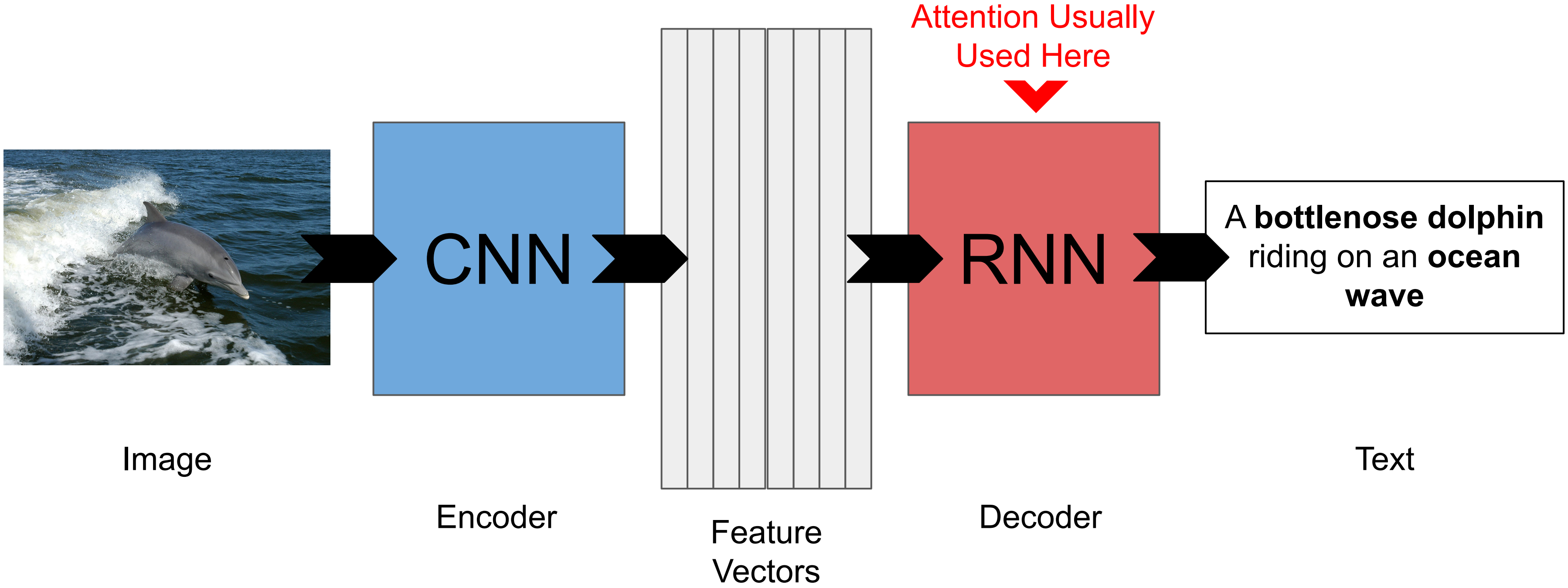
The Attention Mechanism





Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In NeurIPS, 2017.

The Attention Mechanism



Drawbacks

Not sure if attention result is related to

RNN

Drawbacks

Not sure if attention result is related to

RNN

A **seal** riding on a **cup of Sprite**

???

Drawbacks

Not sure if attention result is related to

RNN

A **seal** riding on a **cup of Sprite**

???

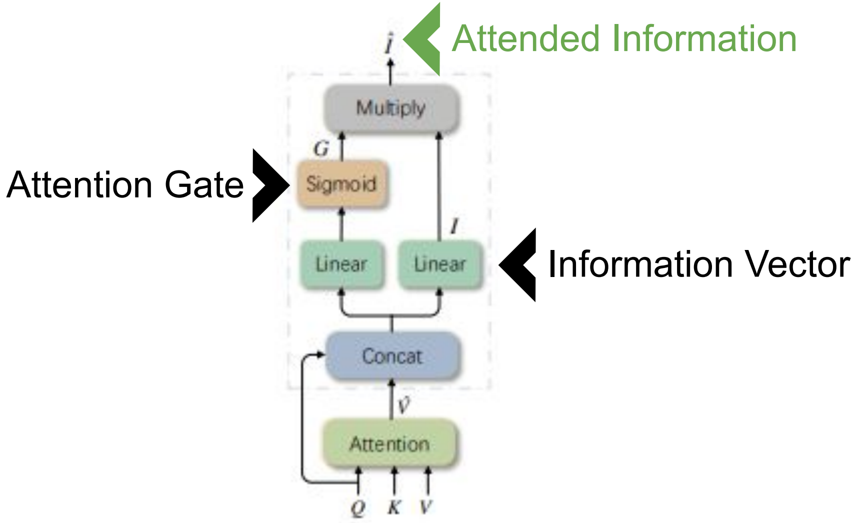
Causes

1. Attention Model does not do well
2. Vectors have no good information

Solution

Attention on Attention

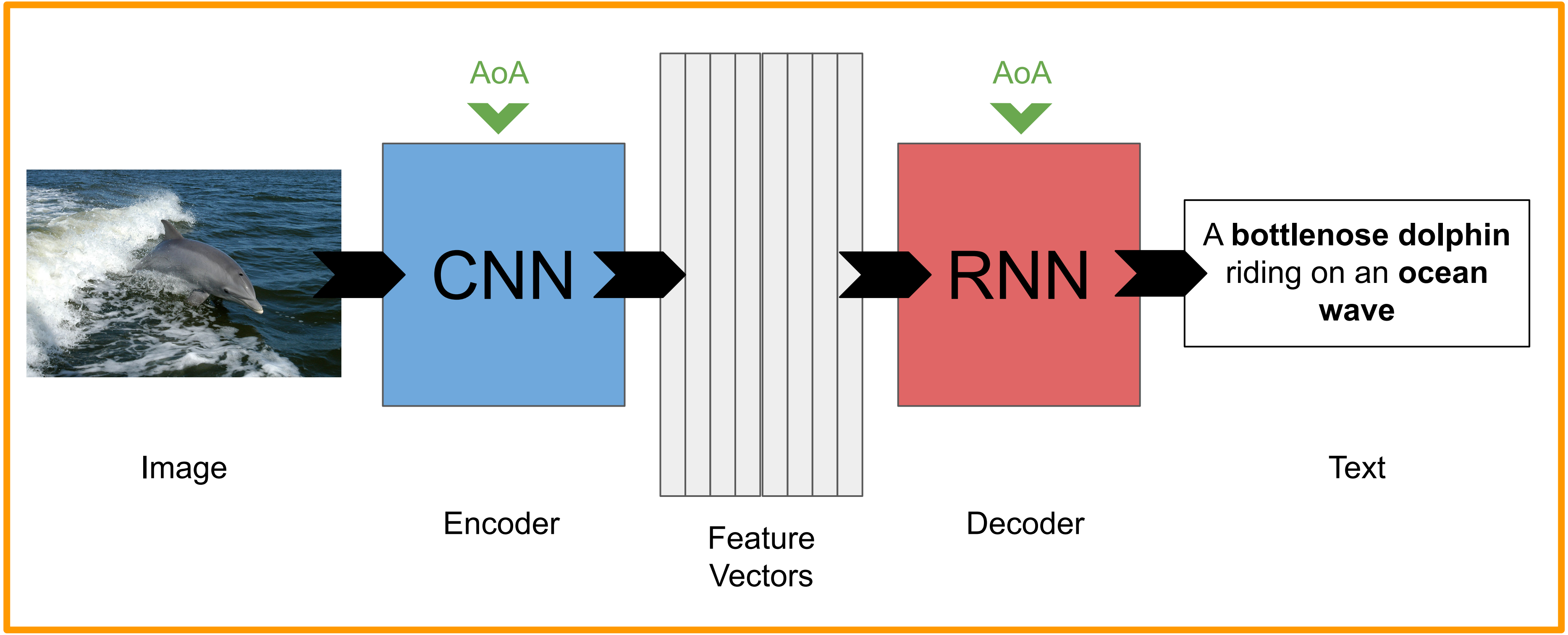
Adding another attention to the existing attention



Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. 2019. Attention on attention for image captioning. (August 2019).

Solution

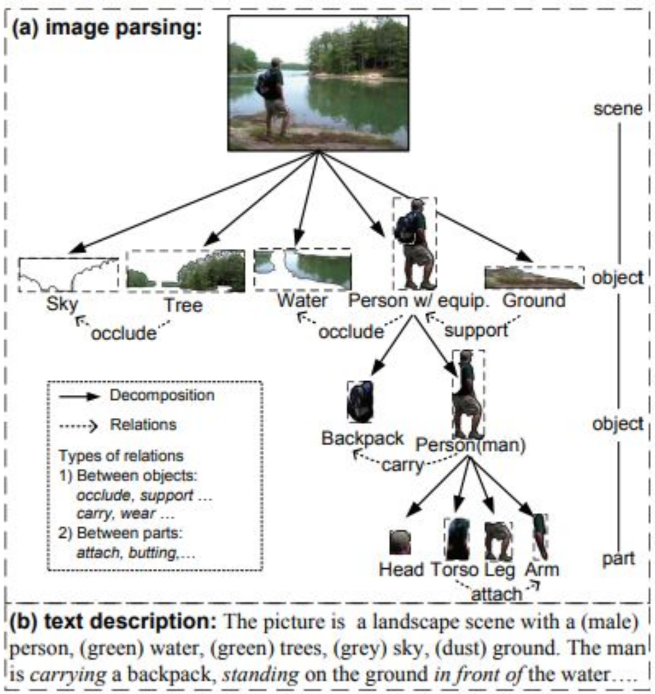
AoANet



Prior Work

Prior Work: Image Captioning

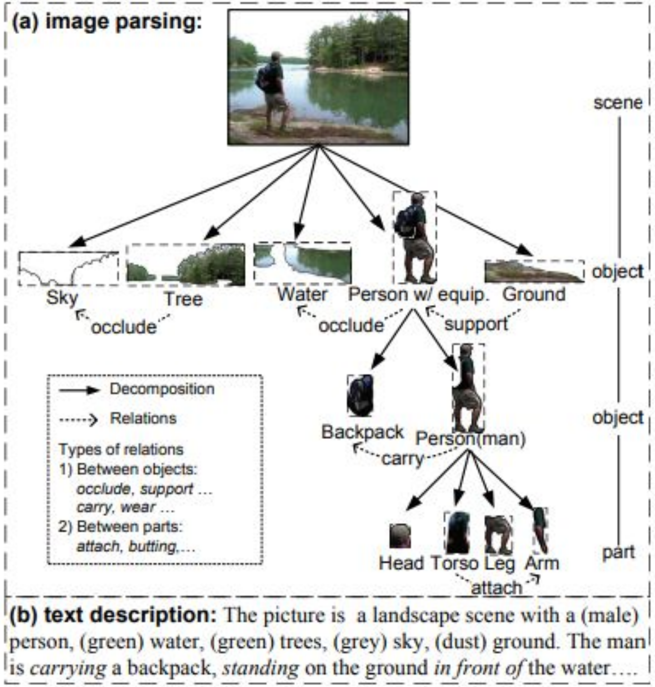
Early Approaches:



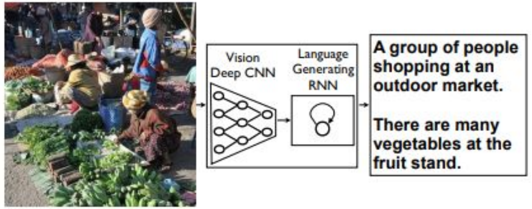
Benjamin Z Yao, Xiong Yang, Liang Lin, Mun Wai Lee, and Song-Chun Zhu. I2t: Image parsing to text description. Proceedings of the IEEE, 98(8):1485– 1508, 2010.
Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In CVPR, pages 3156–3164, 2015.
Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In ICML, 2015.
Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In CVPR, 2017.
Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In CVPR, 2018.

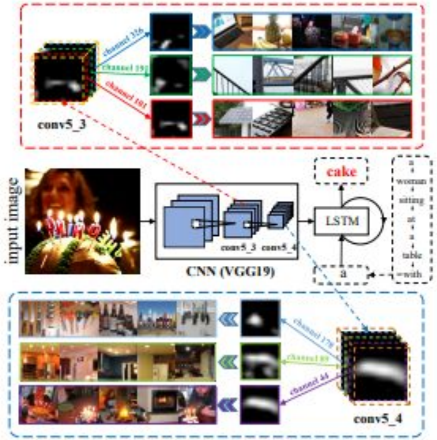
Prior Work: Image Captioning

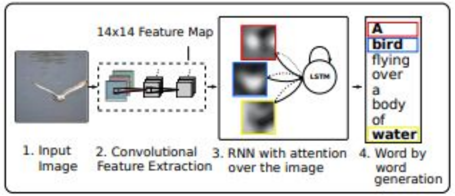
Early Approaches:

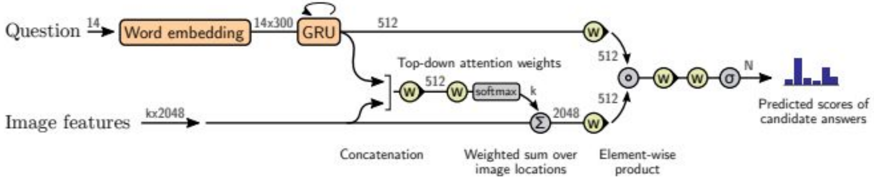


More Recent Approaches:



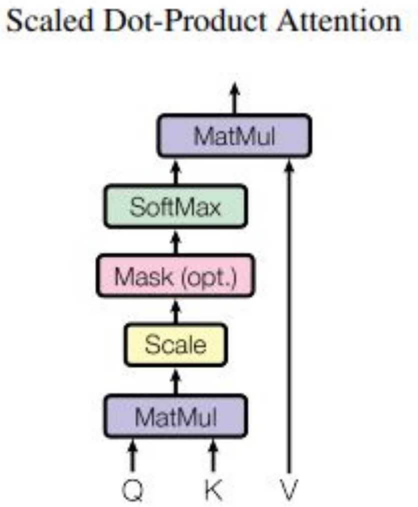


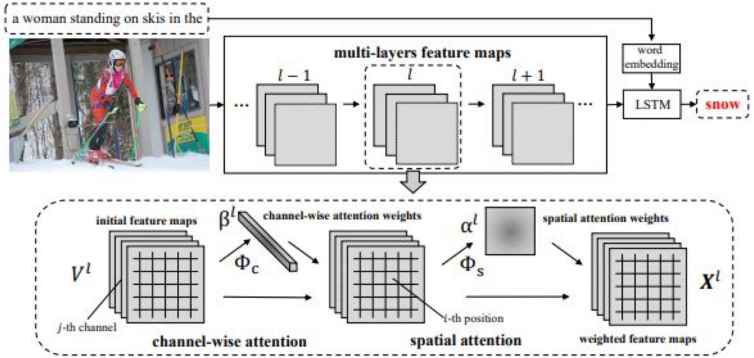


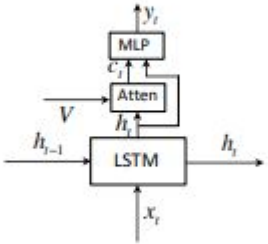


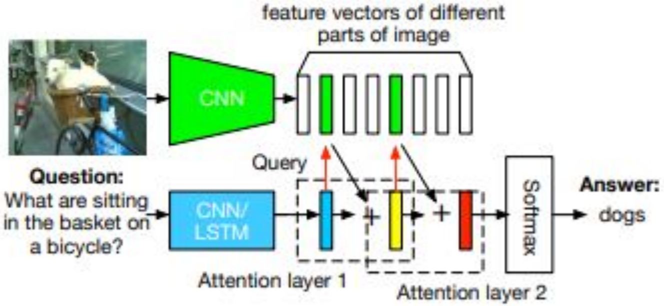
Benjamin Z Yao, Xiong Yang, Liang Lin, Mun Wai Lee, and Song-Chun Zhu. I2t: Image parsing to text description. Proceedings of the IEEE, 98(8):1485– 1508, 2010.
Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In CVPR, pages 3156–3164, 2015.
Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In ICML, 2015.
Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In CVPR, 2017.
Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In CVPR, 2018.

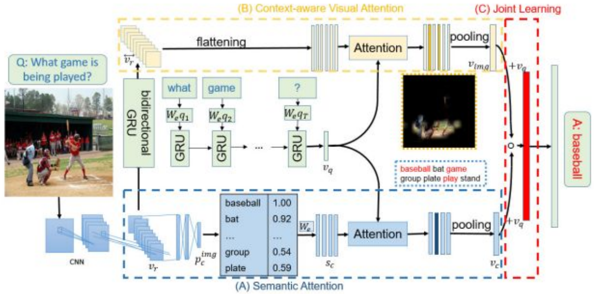
Prior Work: Attention Mechanisms





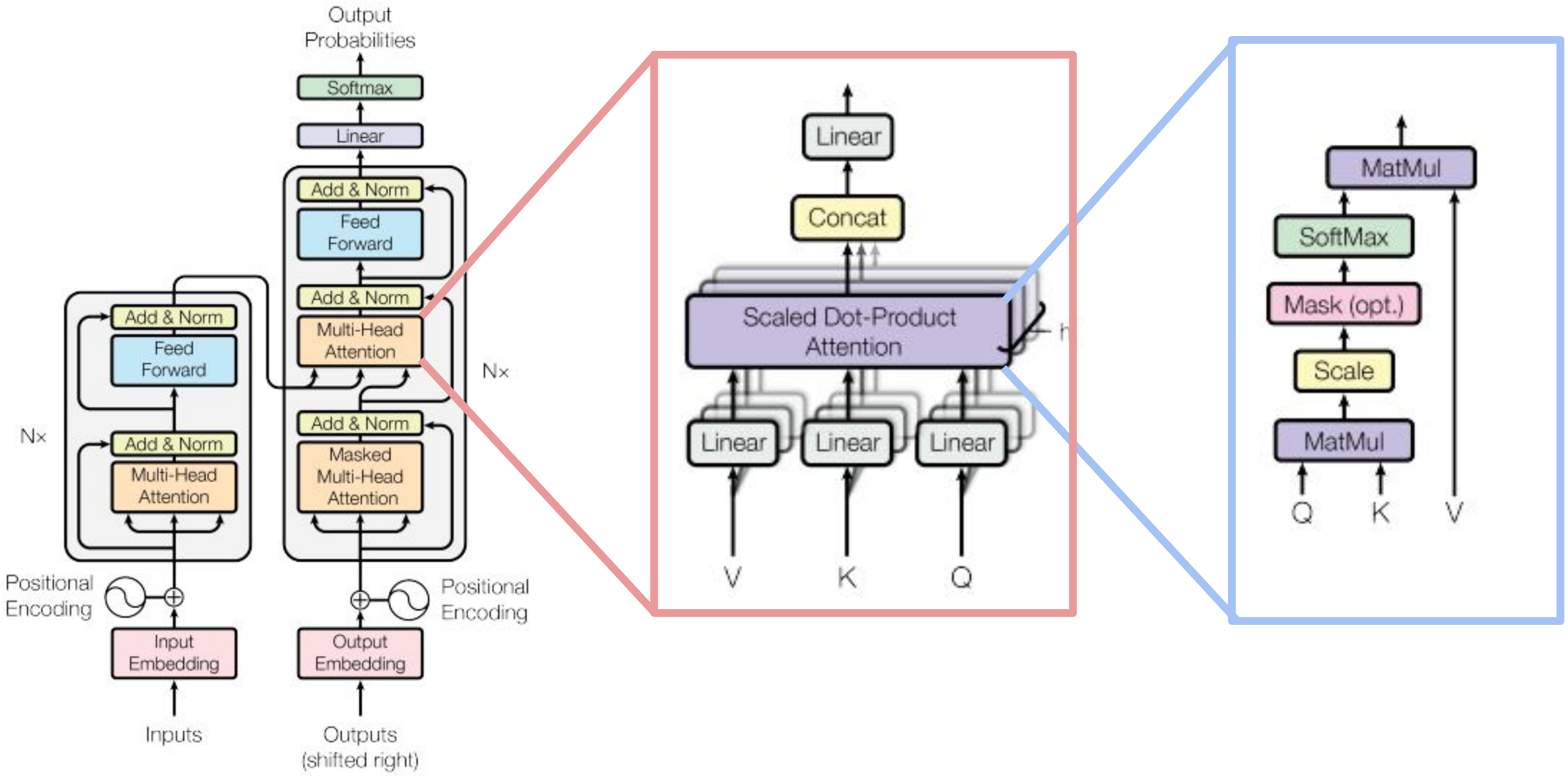






Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In CVPR, 2017.
Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention 4642 via a visual sentinel for image captioning. In CVPR, 2017.
Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In CVPR, June 2016.
Dongfei Yu, Jianlong Fu, Tao Mei, and Yong Rui. Multi-level attention networks for visual question answering. In CVPR, July 2017.
Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In NeurIPS, 2017.

Prior Work: Self-Attention



Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In NeurIPS, 2017.

Prior Work: Self-Attention

Given Q is a matrix of queries, K is a matrix of keys, and V is a matrix of values:

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

$$\mathrm{MultiHead}(Q, K, V) = \mathrm{Concat}(\mathrm{head_1}, ..., \mathrm{head_h})W^O$$
$$\text{where } \mathrm{head_i} = \mathrm{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

Uses in the Transformer model:

1. “Encoder-decoder attention” layers
2. Self-attention layers for encoder
3. Self-attention layers for decoder

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In NeurIPS, 2017.

Prior Work: Self-Attention

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In NeurIPS, 2017.

Prior Work: Self-Attention

Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [15]	23.75			
Deep-Att + PosUnk [32]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [31]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [8]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [26]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [32]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [31]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [8]	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**	**41.0**	$2.3 \cdot 10^{19}$	

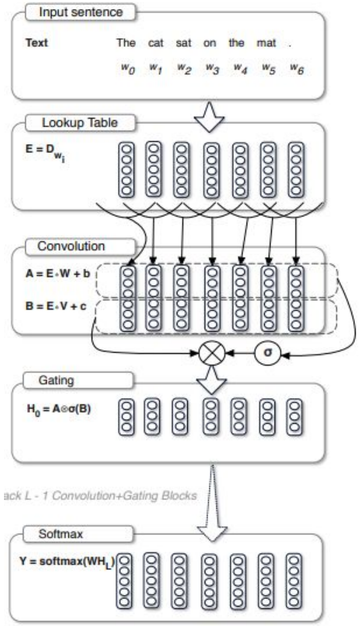
Self-attention can achieve state-of-the-art results in machine translation and computer vision

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In NeurIPS, 2017.

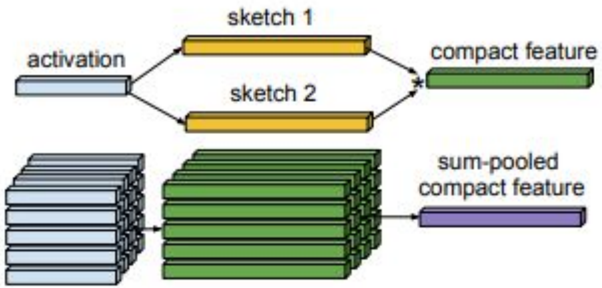
Prior Work: Attention Gates

Creation and application of attention gates are similar to:

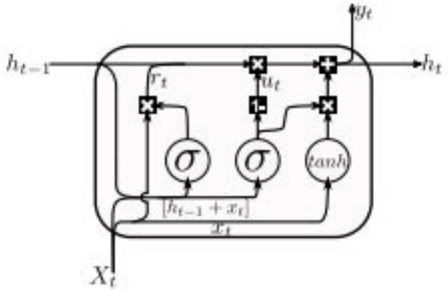
GLUs:

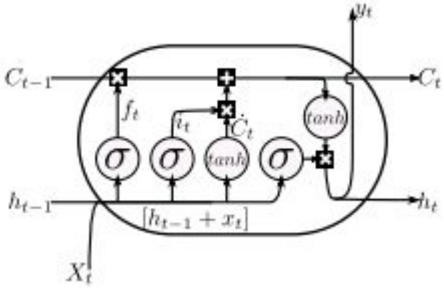


Multi-modal fusion:



GRUs/LSTMs:





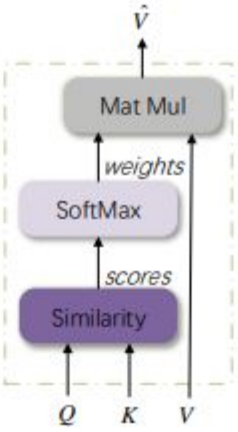
Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In ICLR, 2016.
Y. Gao, O. Beijbom, N. Zhang, and T. Darrell, “Compact bilinear pooling,” arXiv.org, 12-Apr-2016. [Online]. Available: https://arxiv.org/abs/1511.06062. [Accessed: 12-Feb-2023].
R. Cahuantzi, X. Chen, and S. Güttel, “A comparison of LSTM and GRU networks for learning symbolic sequences,” arXiv.org, 04-Jan-2023. [Online]. Available: https://arxiv.org/abs/2107.02248. [Accessed: 12-Feb-2023].

Prior Work

Overall, Attention on Attention is an extension of the existing attention mechanisms and can be applied to any of them.

Proposed Approach

Attention



Suppose the following:

Q is the set of queries, **K** is the set of keys, and **V** is the set of values

$\mathbf{f_{sim}(q_i,k_j)}$ is an arbitrary similarity model, with inputs $\mathbf{q_i}$ and $\mathbf{k_j}$, respectively, being the **i**th query in **Q** and **j**th key in **K**

$\mathbf{v_j}$ is the **j**th value in **V** corresponding to $\mathbf{k_j}$

Then:

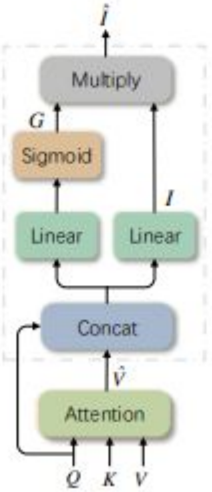
the attended vector $\mathbf{\hat{v}_i}$ for query $\mathbf{q_i}$ can be described as $\mathbf{\hat{v}_i = \Sigma_j f_{sim}(q_i,k_j)v_j}$

This method will be denoted as $\mathbf{f_{att}(Q,K,V)}$

$\mathbf{f_{att}(Q,K,V) = \hat{V}}$ with $\mathbf{\hat{V}}$ as the resulting weighted average vectors over **V**

Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. 2019. Attention on attention for image captioning. (August 2019).

Attention on Attention



Define **i** to be the information vector and **g** to be the attention gate

Given a query **q** and the attention result **v̂** from $\mathbf{f_{att}(Q,K,V)}$:

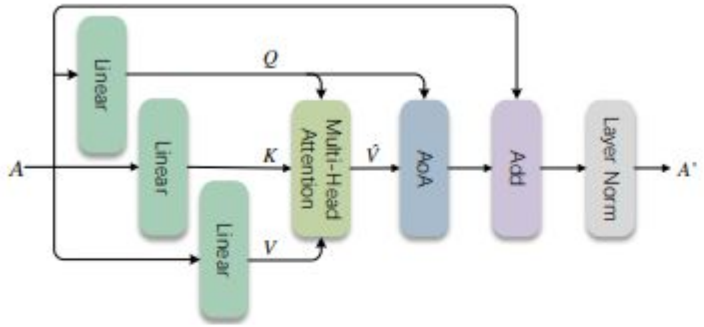
$\mathbf{i = W_q^i q + W_v^i \hat{v} + b^i}$ and $\mathbf{g} = \boldsymbol{sigmoid}(\mathbf{W_q^g q + W_v^g \hat{v} + b^g})$
where **W, b** are associated linear projection constants

g is then element-wise multiplied with **i** to obtain attended information **î**

$$\mathrm{AoA}(f_{att}, \boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \sigma(W_q^g \boldsymbol{Q} + W_v^g f_{att}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) + b^g) \odot (W_q^i \boldsymbol{Q} + W_v^i f_{att}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) + b^i) \quad (6)$$

Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. 2019. Attention on attention for image captioning. (August 2019).

AoANet: Encoder (Refining Module)



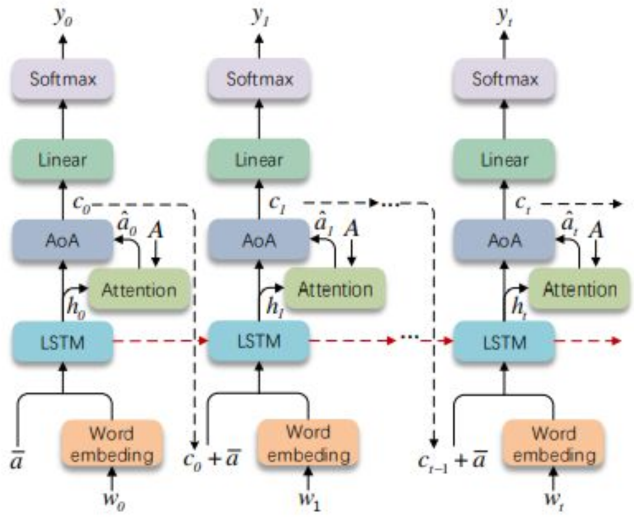
A is a set of feature vectors from the CNN encoder network

A' = *LayerNorm*(A + AoA(*MultiHeadAttention*, W^QA, W^KA, W^VA))

Similar to the Transformer structure but with the feed-forward layer removed

Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. 2019. Attention on attention for image captioning. (August 2019).

AoANet: Decoder



$\mathbf{c_t}$ **= AoA(*MultiHeadAttention*, $\mathbf{W^Q h_t}$, $\mathbf{W^K A}$, $\mathbf{W^V A}$))** with $\mathbf{h_t}$ being the LSTM output

$$\boldsymbol{x}_t = [W_e \Pi_t, \bar{\boldsymbol{a}} + \boldsymbol{c}_{t-1}]$$

$$\boldsymbol{h}_t, \boldsymbol{m}_t = \mathrm{LSTM}(\boldsymbol{x}_t, \boldsymbol{h}_{t-1}, \boldsymbol{m}_{t-1})$$

Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. 2019. Attention on attention for image captioning. (August 2019).

Loss and Optimization

Cross Entropy Loss:

$$L_{XE}(\theta) = -\sum_{t=1}^{T} \log(p_\theta(\boldsymbol{y}_t^* \mid \boldsymbol{y}_{1:t-1}^*))$$

CIDEr-D Score Optimization:

$$L_{RL}(\theta) = -\mathbf{E}_{\boldsymbol{y}_{1:T} \sim p_\theta}[r(\boldsymbol{y}_{1:T})]$$

Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. 2019. Attention on attention for image captioning. (August 2019).

Implementation

Encoder: Faster-RCNN pre-trained on ImageNet and Visual Genome to retrieve 2048 dimensional vectors

Decoder: LSTM with hidden size 1024

Training:

Batch Size: 10
Epochs: 30 for $L_{\mathbf{XE}}$ then 15 for L_{RL}
Learning Rate: 2e-4 annealed by 0.8 every 3 epochs for L_{XE}, 2e-5 annealed by 0.5 if score does not improve for L_{RL}
Optimizer: ADAM for L_{XE}, SCST for L_{RL}

Evaluation

Dataset and Metrics

Dataset:

MS COCO - 123,287 images with 5 captions each

“Kaparthy” split used for offline training

Metrics: BLEU, METEOR, ROUGE-L, CIDEr-D, SPICE

Tsung Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and ´ C. Lawrence Zitnick. Microsoft coco: Common objects in context. In ECCV, 2014.

Quantitative Evaluation

Baselines:

LSTM, SCST, Up-Down, RFNet, GCN-LSTM, SGAE

All trained under XE loss and then optimized with RL loss

Offline Evaluation: Tested on the “Kaparthy” training split

Online Evaluation: Tested on the online COCO test server

Qualitative Evaluation also performed

Offline Quantitative Evaluation

Model	Cross-Entropy Loss						CIDEr-D Score Optimization					
Metric	B@1	B@4	M	R	C	S	B@1	B@4	M	R	C	S
							Single Model					
LSTM [37]	-	29.6	25.2	52.6	94.0	-	-	31.9	25.5	54.3	106.3	-
SCST [31]	-	30.0	25.9	53.4	99.4	-	-	34.2	26.7	55.7	114.0	-
LSTM-A [50]	75.4	35.2	26.9	55.8	108.8	20.0	78.6	35.5	27.3	56.8	118.3	20.8
Up-Down [2]	77.2	36.2	27.0	56.4	113.5	20.3	79.8	36.3	27.7	56.9	120.1	21.4
RFNet [20]	76.4	35.8	27.4	56.8	112.5	20.5	79.1	36.5	27.7	57.3	121.9	21.2
GCN-LSTM [49]	77.3	36.8	27.9	57.0	116.3	20.9	80.5	38.2	28.5	58.3	127.6	22.0
SGAE [44]	-	-	-	-	-	-	**80.8**	38.4	28.4	58.6	127.8	22.1
AoANet (Ours)	**77.4**	**37.2**	**28.4**	**57.5**	**119.8**	**21.3**	80.2	**38.9**	**29.2**	**58.8**	**129.8**	**22.4**
							Ensemble/Fusion					
SCST [31]$^{\Sigma}$	-	32.8	26.7	55.1	106.5	-	-	35.4	27.1	56.6	117.5	-
RFNet [20]$^{\Sigma}$	77.4	37.0	27.9	57.3	116.3	20.8	80.4	37.9	28.3	58.3	125.7	21.7
GCN-LSTM [49]$^{\Sigma}$	77.4	37.1	28.1	57.2	117.1	21.1	80.9	38.3	28.6	58.5	128.7	22.1
SGAE [44]$^{\Sigma}$	-	-	-	-	-	-	81.0	39.0	28.4	58.9	129.1	22.2
AoANet (Ours)$^{\Sigma}$	**78.7**	**38.1**	**28.5**	**58.2**	**122.7**	**21.7**	**81.6**	**40.2**	**29.3**	**59.4**	**132.0**	**22.8**

Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. 2019. Attention on attention for image captioning. (August 2019).

Online Quantitative Evaluation

Model	BLEU-1		BLEU-2		BLEU-3		BLEU-4		METEOR		ROUGE-L		CIDEr-D	
Metric	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
SCST [31]	78.1	93.7	61.9	86.0	47.0	75.9	35.2	64.5	27.0	35.5	56.3	70.7	114.7	116.0
LSTM-A [50]	78.7	93.7	62.7	86.7	47.6	76.5	35.6	65.2	27.0	35.4	56.4	70.5	116.0	118.0
Up-Down [2]	80.2	95.2	64.1	88.8	49.1	79.4	36.9	68.5	27.6	36.7	57.1	72.4	117.9	120.5
RFNet [20]	80.4	95.0	64.9	89.3	50.1	80.1	38.0	69.2	28.2	37.2	58.2	73.1	122.9	125.1
GCN-LSTM [49]	-	-	65.5	89.3	50.8	80.3	38.7	69.7	28.5	37.6	58.5	73.4	125.3	126.5
SGAE [44]	**81.0**	**95.3**	65.6	89.5	50.7	80.4	38.5	69.7	28.2	37.2	58.6	73.6	123.8	126.5
AoANet (Ours)	**81.0**	95.0	**65.8**	**89.6**	**51.4**	**81.3**	**39.4**	**71.2**	**29.1**	**38.5**	**58.9**	**74.5**	**126.9**	**129.6**

Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. 2019. Attention on attention for image captioning. (August 2019).

Qualitative Evaluation

Comparison Baseline:

Up-Down with the settings of AoANet

Observations:

1. AoANet counts objects of the same kind more accurately

2. AoANet properly determines the interactions of objects

Image	Captions
	**AoANet**: Two birds sitting on top of a giraffe. **Baseline**: A bird sitting on top of a tree. **GT1**. Two birds going up the back of a giraffe. **GT2**. A large giraffe that is walking by some trees. **GT3**. Two birds are sitting on a wall near the bushes.
	**AoANet**: Two cats laying on top of a bed. **Baseline**: A black and white cat laying on top of a bed. **GT1**. A couple of cats laying on top of a bed. **GT2**. Two cats laying on a big bed and looking at the camera. **GT3**. A couple of cats on a mattress laying down.
	**AoANet**: A cat looking at its reflection in a mirror. **Baseline**: A cat is looking out of a window. **GT1**. A cat looking at his reflection in the mirror. **GT2**. A cat that is looking in a mirror. **GT3**. A cat looking at itself in a mirror.
	**AoANet**: A young boy hitting a tennis ball with a tennis racket. **Baseline**: A young man holding a tennis ball on a court. **GT1**. A guy in a maroon shirt is holding a tennis racket out to hit a tennis ball. **GT2**. A man on a tennis court that has a racquet. **GT3**. A boy hitting a tennis ball on the tennis court.

Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. 2019. Attention on attention for image captioning. (August 2019).

Ablation Studies

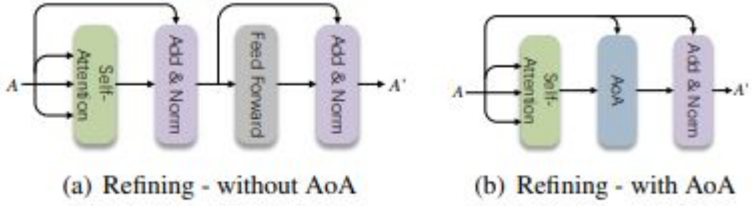


Figure 6: Refining modules w/o and w/ AoA.

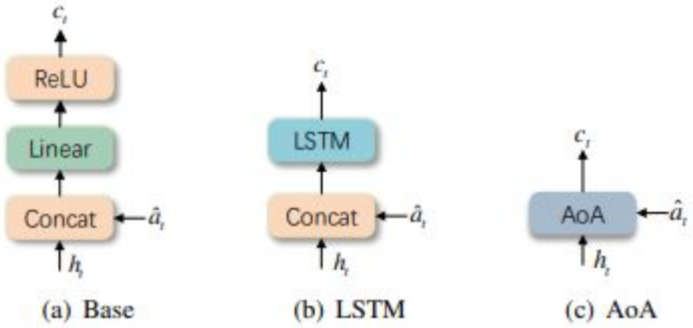


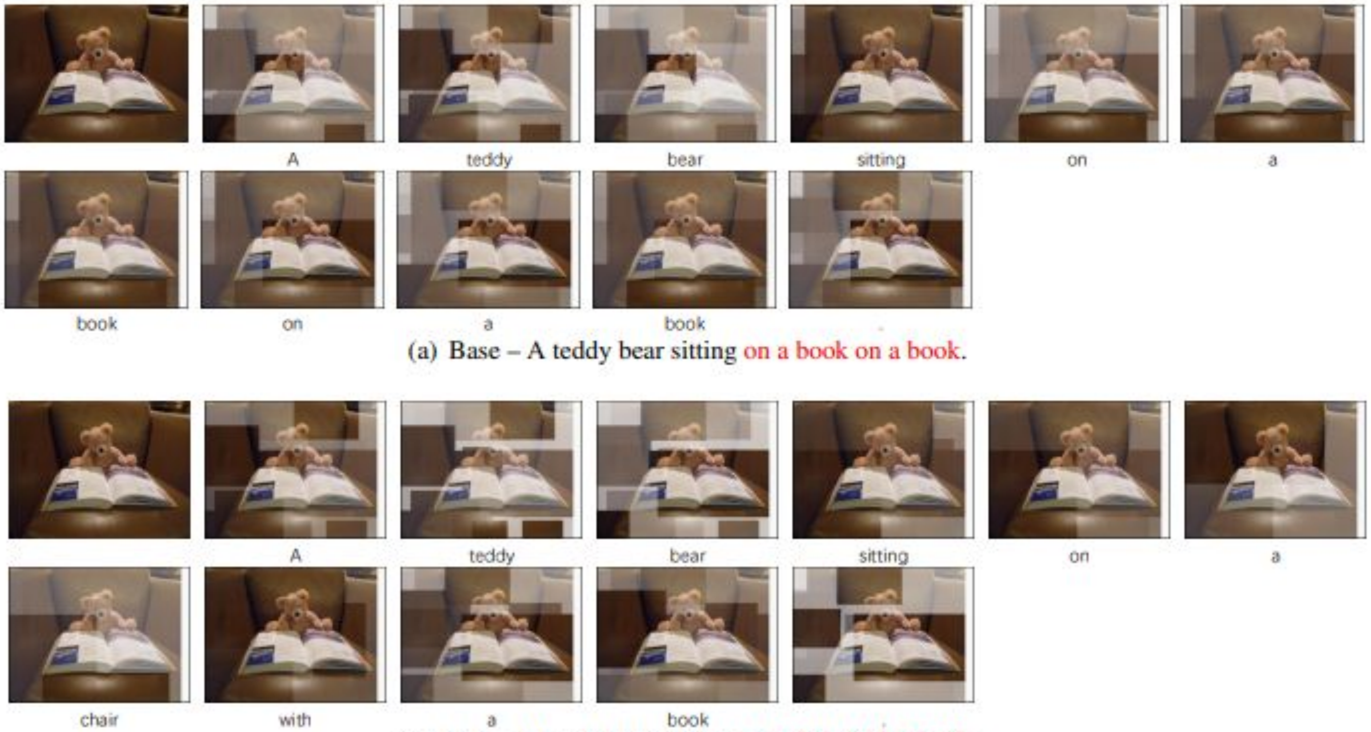
Figure 7: Different schemes for decoders to model c_t.

Model	B@1	B@4	R	C
Base	75.7	34.9	56.0	109.5
+ Enc: Refine (w/o AoA)	**77.0**	35.6	56.4	112.5
+ Enc: Refine (w/ AoA)	76.7	**36.1**	**56.7**	**114.5**
+ Dec: LSTM	76.8	35.9	56.6	113.5
+ Dec: AoA	76.6	35.8	56.6	113.8
+ Dec: LSTM + AoA	*unstable training process*			
+ Dec: MH-Att	75.8	34.8	56.0	109.6
+ Dec: MH-Att, LSTM	76.6	35.8	**56.7**	113.8
+ Dec: MH-Att, AoA	**76.9**	**36.1**	56.6	**114.3**
Full: AoANet	**77.4**	**37.2**	**57.5**	**119.8**

Comparatively, AoA requires less computation than LSTM

Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. 2019. Attention on attention for image captioning. (August 2019).

Ablation Studies



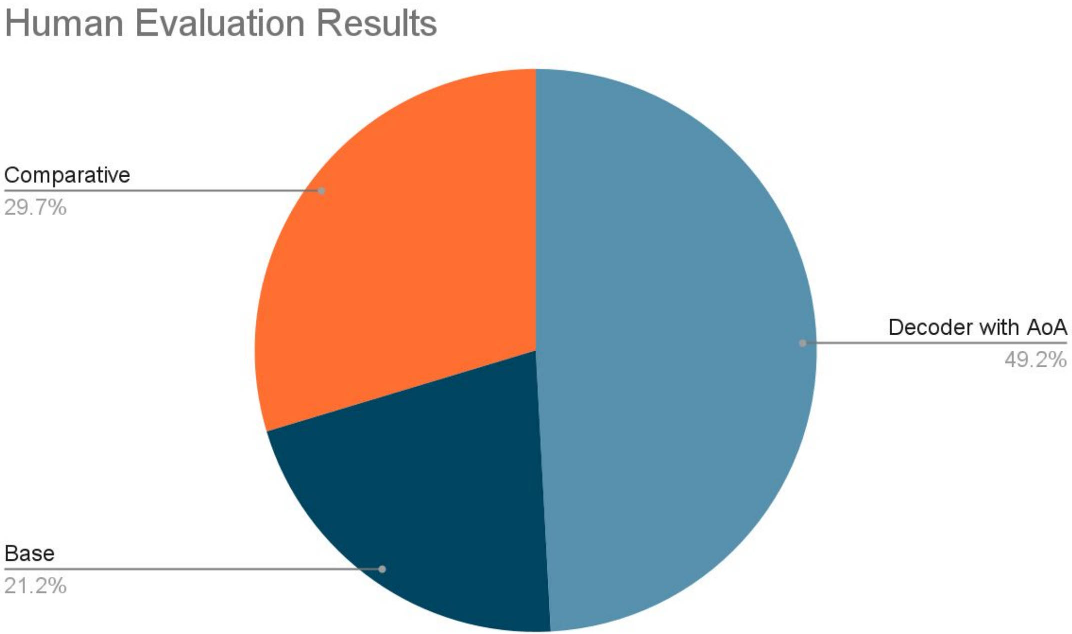
(a) Base – A teddy bear sitting on a book on a book.

(b) AoA – A teddy bear sitting on a chair with a book.

Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. 2019. Attention on attention for image captioning. (August 2019).

Human Evaluation

30 evaluators were invited to evaluate 100 images and asked to choose which of the two captions were better:



Generalization

MSR-VTT Dataset

1. A black and white horse runs around.
2. A horse galloping through an open field.
3. A horse is running around in green lush grass.
4. There is a horse running on the grassland.
5. A horse is riding in the grass.

	BLEU-4	CIDEr-D	ROUGE-L
base	33.53	38.83	56.90
decoder with AoA	37.22	42.44	58.32

https://production-media.paperswithcode.com/datasets/Screen_Shot_2021-01-28_at_9.51.08_PM.png

Strengths and Weaknesses

Strength #1 - Quite Efficient Compared To Normal LSTM

Less calculations are needed due to less hidden states

Strength #1 - Quite Efficient Compared To Normal LSTM

Less calculations are needed due to less hidden states

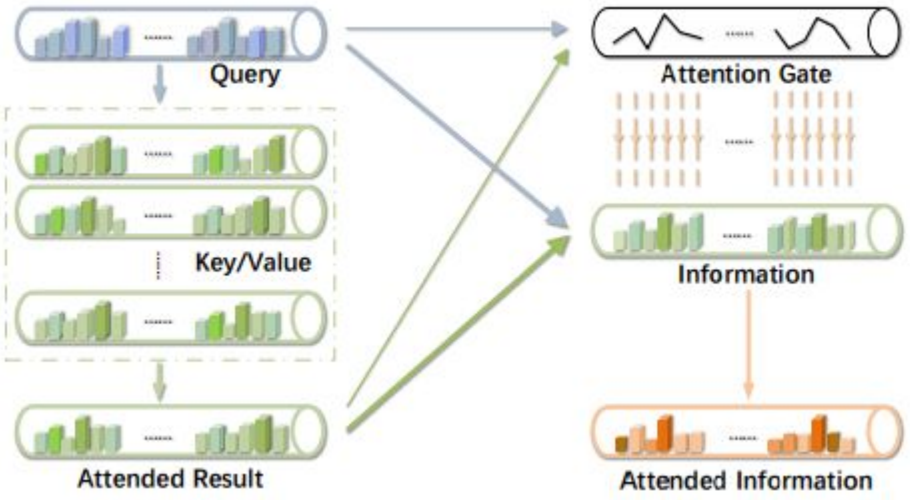
Model	B@1	B@4	R	C
Base	75.7	34.9	56.0	109.5
+ Enc: Refine (w/o AoA)	**77.0**	35.6	56.4	112.5
+ Enc: Refine (w/ AoA)	76.7	**36.1**	**56.7**	**114.5**
+ Dec: LSTM	76.8	35.9	56.6	113.5
+ Dec: AoA	76.6	35.8	56.6	113.8
+ Dec: LSTM + AoA	*unstable training process*			
+ Dec: MH-Att	75.8	34.8	56.0	109.6
+ Dec: MH-Att, LSTM	76.6	35.8	**56.7**	113.8
+ Dec: MH-Att, AoA	**76.9**	**36.1**	56.6	**114.3**
Full: AoANet	**77.4**	**37.2**	**57.5**	**119.8**

Will help AoA stand out when performance is comparable

Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. 2019. Attention on attention for image captioning. (August 2019).

Strength #2 - Novel Compared To Previous Work

Very **unique and novel** concept



Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. 2019. Attention on attention for image captioning. (August 2019).

Strength #3 - Very Thorough Evaluation

- Multiple baselines and metrics
- Multiple angles of evaluation (quantitative, qualitative, ablations, etc.)
- High amount of evidence that AoA is effective:

Helps make a convincing argument for AoA being the new state-of-the-art for image captioning and perhaps other applications as well

Weakness #1 - Underdeveloped Training for AoANet

- Standard loss functions, especially XE loss
- Implementation details without much justification
- Why is CIDEr-D optimization needed?
- How XE and RL losses are handled together is not very clear

Weakness #2 - Lack of Accuracy Increase for Decoder

Model	B@1	B@4	R	C
Base	75.7	34.9	56.0	109.5
+ Enc: Refine (w/o AoA)	**77.0**	35.6	56.4	112.5
+ Enc: Refine (w/ AoA)	76.7	**36.1**	**56.7**	**114.5**
+ Dec: LSTM	76.8	35.9	56.6	113.5
+ Dec: AoA	76.6	35.8	56.6	113.8
+ Dec: LSTM + AoA	*unstable training process*			
+ Dec: MH-Att	75.8	34.8	56.0	109.6
+ Dec: MH-Att, LSTM	76.6	35.8	**56.7**	113.8
+ Dec: MH-Att, AoA	**76.9**	**36.1**	56.6	**114.3**
Full: AoANet	**77.4**	**37.2**	**57.5**	**119.8**

Is it necessary to even have AoA when other attention methods can work just as well?

Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. 2019. Attention on attention for image captioning. (August 2019).

Weakness #3 - Methods/Evaluation Could Be Expanded?

Considering the AoA formula, try different attention mechanisms for f_{att}?

$$\begin{aligned}\mathrm{AoA}(f_{att}, \boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \sigma(W_q^g \boldsymbol{Q} + W_v^g f_{att}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) + b^g) \\ \odot (W_q^i \boldsymbol{Q} + W_v^i f_{att}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) + b^i) \quad (6)\end{aligned}$$

Use different encoders other than Faster-RCNN?

Use different decoders other than LSTM?

Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. 2019. Attention on attention for image captioning. (August 2019).

Potential Ideas for Future Work

Future Work Ideas

- Bring AoA to other machine learning tasks, such as machine translation?
- Designing a better loss function?
- Designing a better decoder that utilizes AoA more effectively?
- Designing a better attention mechanism for the decoder step of image captioning?