Visual Attention

2020. 02. 14

Data Mining & Quality Analytics Lab.

강 현 규

Contents

- **Introduction**
- **Attention**
  - Basics
  - Visual Attention
  - Conclusion
- **Self-Attention**
  - Basics
  - Visual Self-Attention
  - Conclusion

❖ Attention: Concept

정답률: 13%

26. So far as you are wholly concentrated on bringing about a certain result, clearly the quicker and easier it is brought about the better. Your resolve to secure a sufficiency of food for yourself and your family will induce you to spend weary days in tilling the ground and tending livestock; but if Nature provided food and meat in abundance ready for the table, you would thank Nature for sparing you much labor and consider yourself so much the better off. An executed purpose, in short, is a transaction in which the time and energy spent on the execution are balanced against the resulting assets, and the ideal case is one in which ______________________ . Purpose, then, justifies the efforts it exacts only conditionally, by their fruits. [3점]

① demand exceeds supply, resulting in greater returns
② life becomes fruitful with our endless pursuit of dreams
③ the time and energy are limitless and assets are abundant
④ Nature does not reward those who do not exert efforts
⑤ the former approximates to zero and the latter to infinity

Attention Mechanism

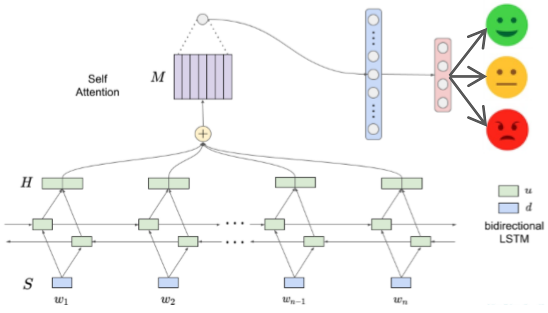
The decoder decides parts of the source sentence to pay attention to. By letting the decoder have an **attention mechanism**, we relieve the encoder from the burden of having to encode all information in the source sentence into a fixed length vector. With this new approach the information can be spread throughout the sequence of annotations, which can be selectively retrieved by the decoder accordingly.

출처: 2011년 대학수학능력시험 외국어영역 홀수형 문제지

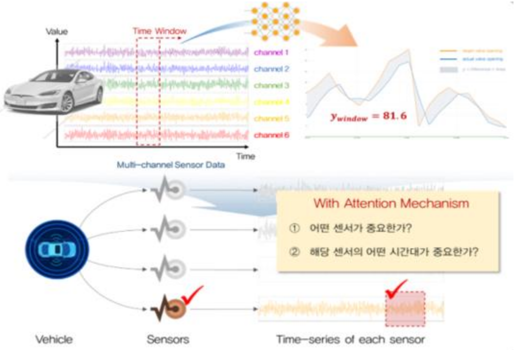
Data Mining Quality Analytics

❖ Attention: 19`s DMQA

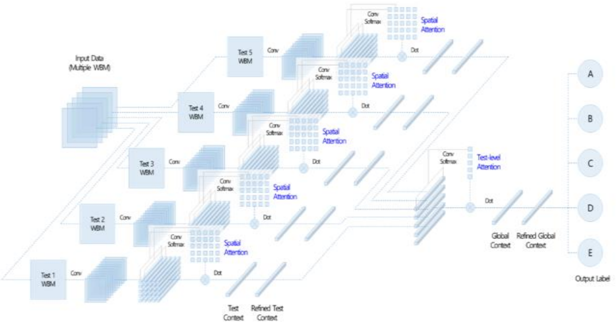
① Sentiment Analysis



② Anomaly Detection



③ Wafer Bin Map Classification



❖ Attention: Application

① Natural Language Processing

- Document Classification
- Machine Translation
- Question & Answering
- Document Summarization

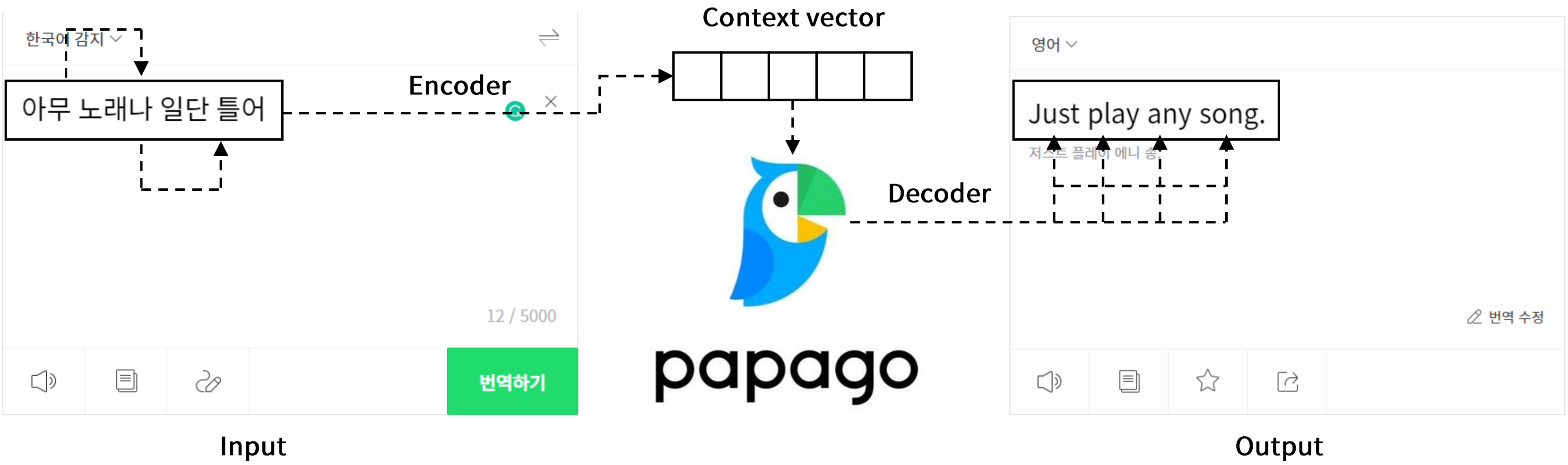
② Multivariate Time Series

- Anomaly Detection
- Predictive Modeling

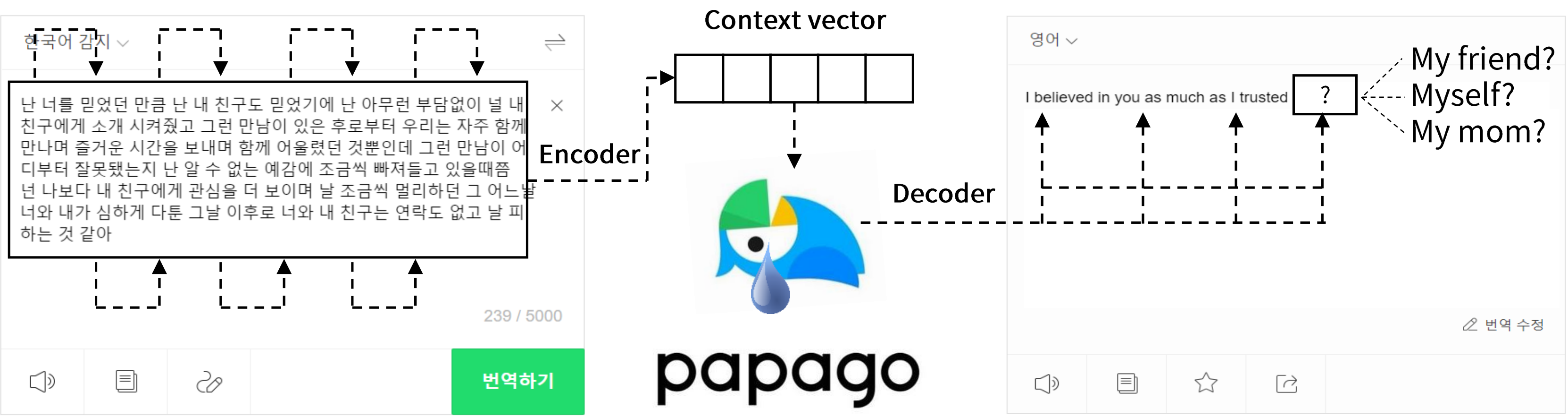
③ Computer Vision

- Image Classification
- Image Captioning
- Object Detection
- Visual QA

❖ Machine Translation: Seq2seq



❖ Machine Translation: Seq2seq

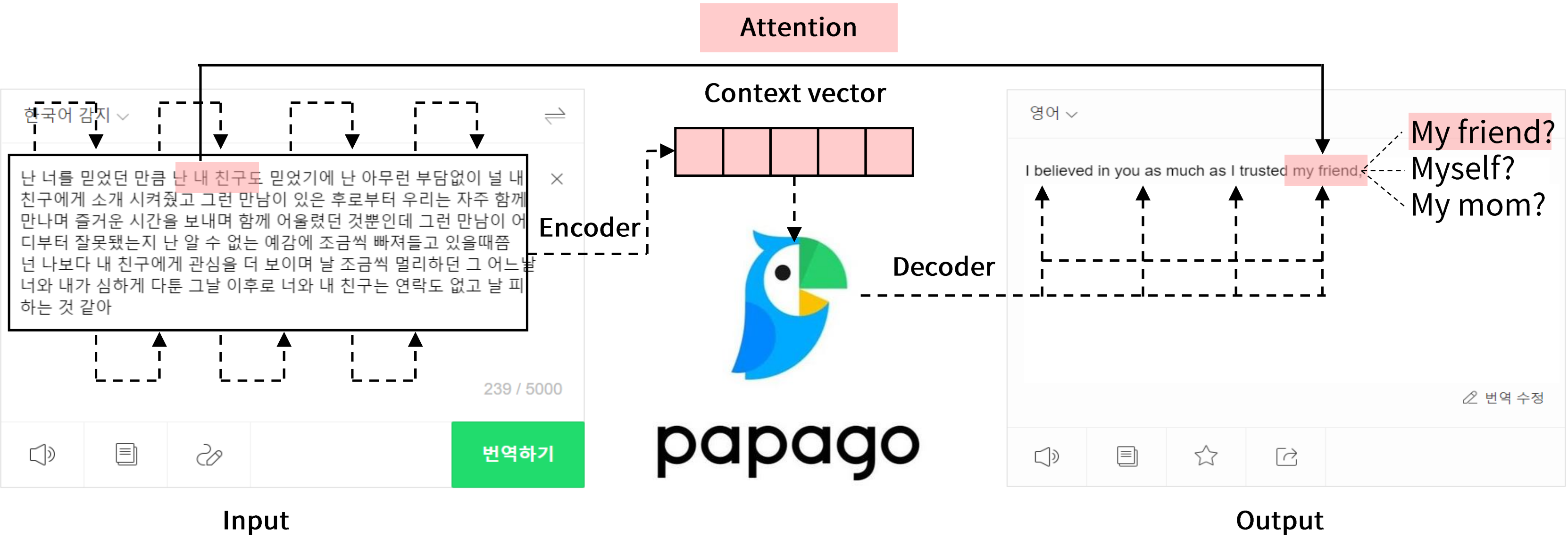


Input

Output

- **Long input & output sequence**
  - Long term dependency
  - Vanishing gradient
  - Limitation of fixed context vector

❖ Machine Translation: Seq2seq

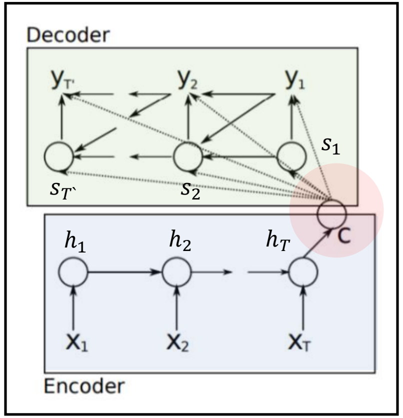


▪ Attention

- Decoder가 특정 시점의 단어를 출력할 때 encoder의 정보 중 연관성이 있는 부분에 주목하여 adaptive context vector를 만든다.
  = Relieve the encoder from the burden of having to encode all information into a fixed length vector.

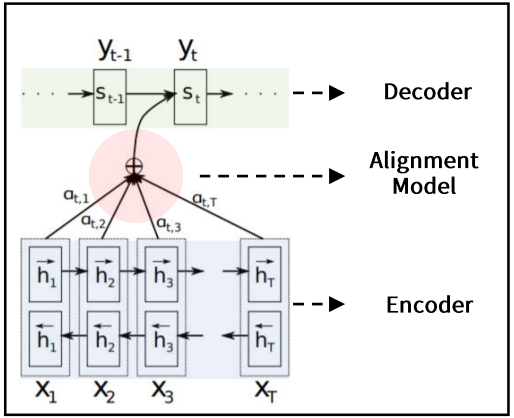
❖ Machine Translation: Seq2seq with Attention

Seq2seq



Attention

Seq2seq with Attention



$$p(y_t|y_{t-1},\ldots,y_1,c) = g(s_t,y_{t-1},c)$$

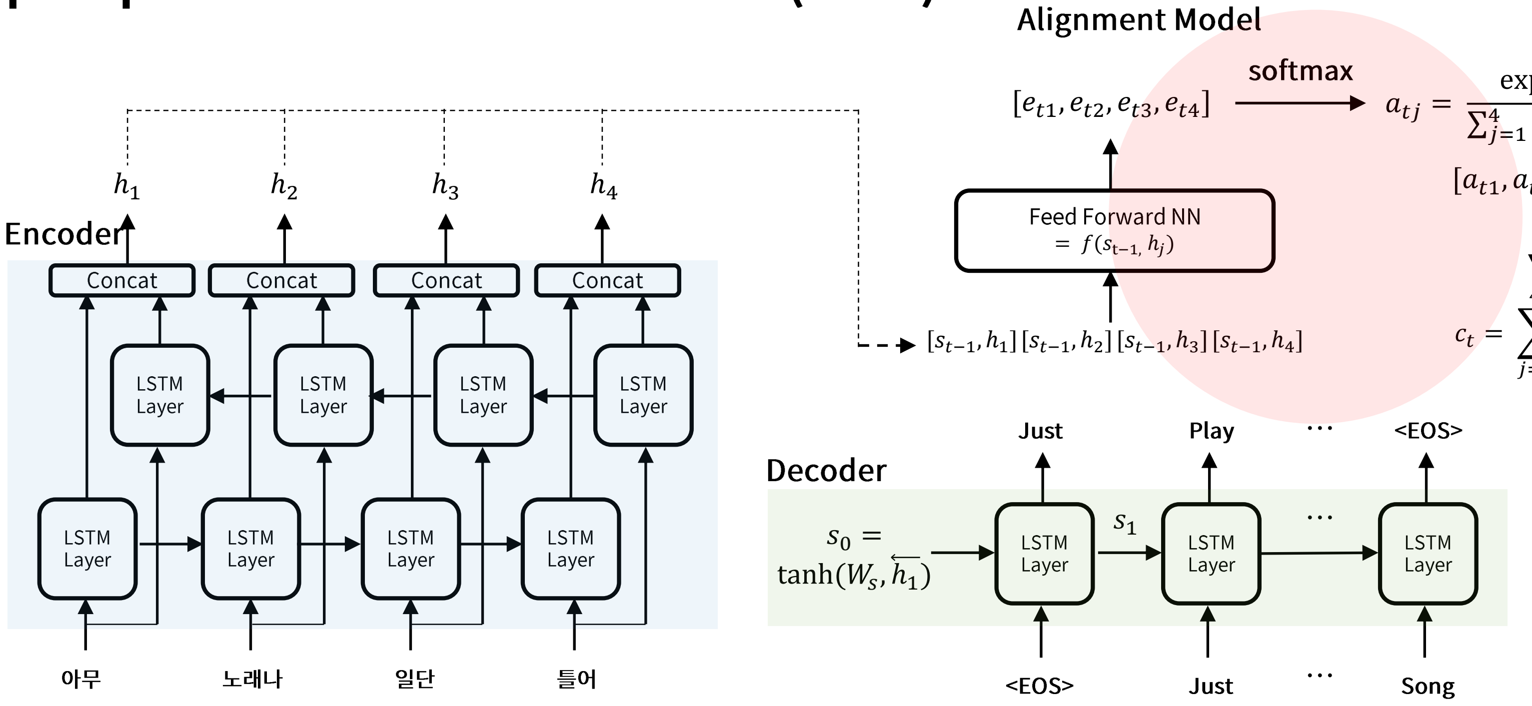
$$p(y_t|y_{t-1},\ldots,y_1,c) = g(s_t,y_{t-1},c_t),\ c_t = \sum_{i=1}^{T} a_i h_i$$

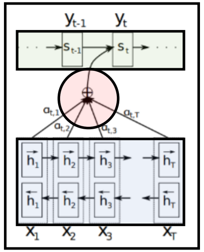
▪ Attention

- Decoder가 특정 시점의 단어를 출력할 때 encoder의 정보 중 연관성이 있는 부분에 주목하여 adaptive context vector를 만든다.
  = Relieve the encoder from the burden of having to encode all information into a fixed length vector.

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

❖ Seq2seq with Attention – Bahdanau(2014)

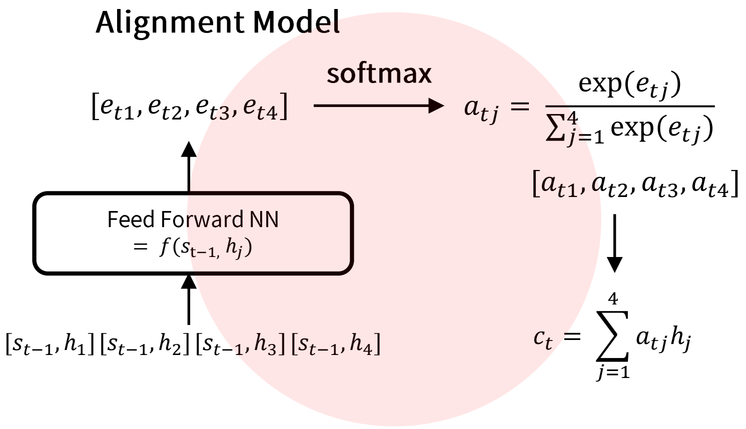


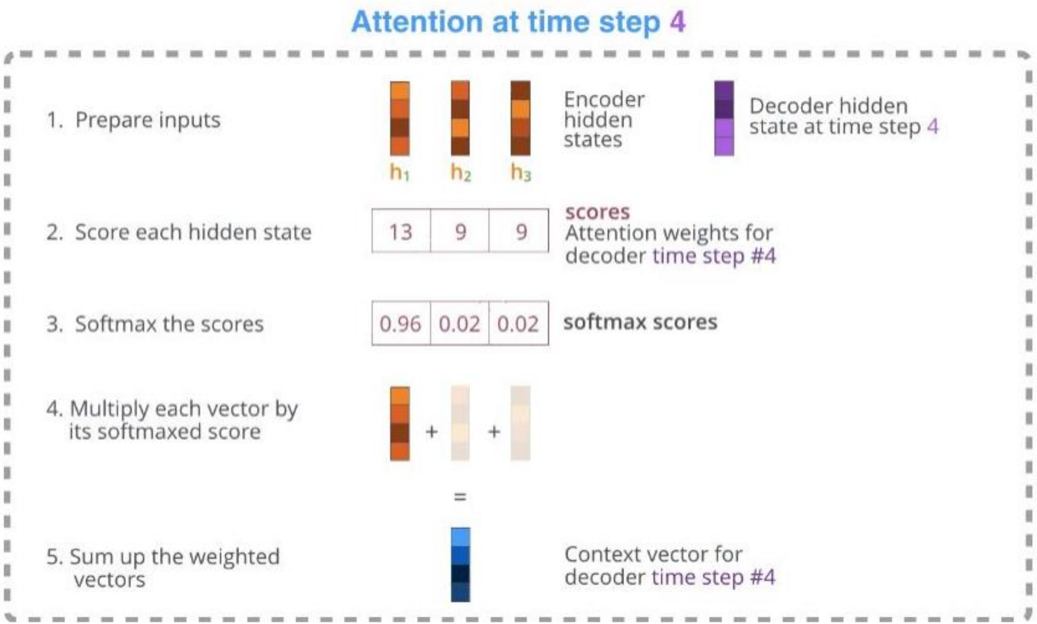


Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." *arXiv preprint arXiv:1409.0473* (2014).

Data Mining Quality Analytics

❖ Seq2seq with Attention – Bahdanau(2014)





https://jalammar.github.io/visualizing-neural-machine-translation-mechanics-of-seq2seq-models-with-attention/

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

Data Mining Quality Analytics

❖ Seq2seq with Attention – Luong(2015)

- Alignment model의 input으로 $s_{t\text{-}1}$가 아닌 s_t을 사용
- 다양한 similarity function $f(s_{t,}\ h_j)$ 제시

Alignment Model

$$[s_t, h_1] \quad [s_t, h_2] \quad [s_t, h_3] \quad [s_t, h_4]$$

$$\boxed{f(s_{t,}\ h_j)}$$

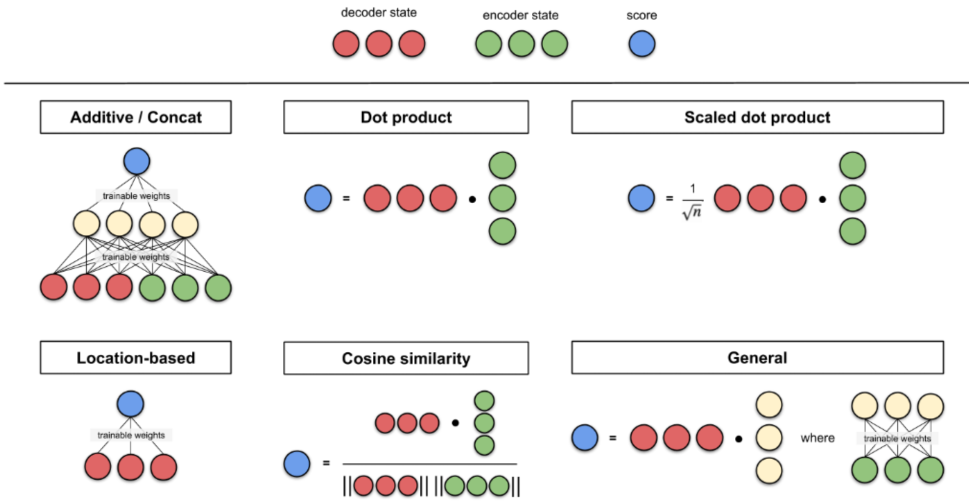
$$[e_{t1}, e_{t2}, e_{t3}, e_{t4}] \xrightarrow{\text{softmax}} a_{tj} = \frac{\exp(e_{tj})}{\sum_{j=1}^{4} \exp(e_{tj})}$$

$$[a_{t1}, a_{t2}, a_{t3}, a_{t4}]$$

$$\downarrow$$

$$c_t = \sum_{j=1}^{4} a_{tj} h_j$$

$$f(s_{t,}\ h_j) = e_j$$



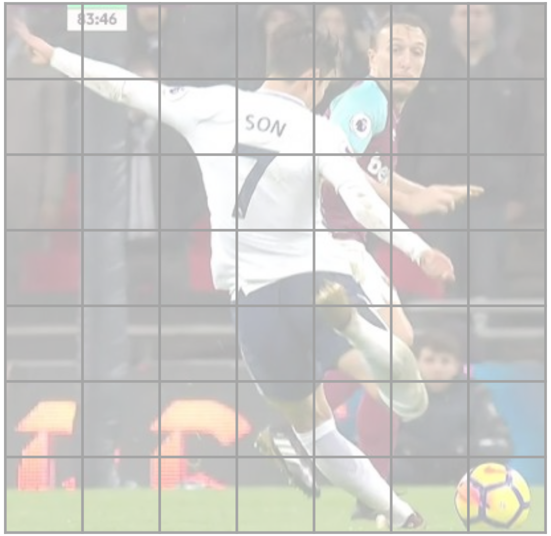
https://towardsdatascience.com/attn-illustrated-attention-5ec4ad276ee3

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

❖ Image Captioning

- 이미지를 입력 데이터로 사용하여 이미지를 설명하는 문장을 생성하는 문제

Input: Image

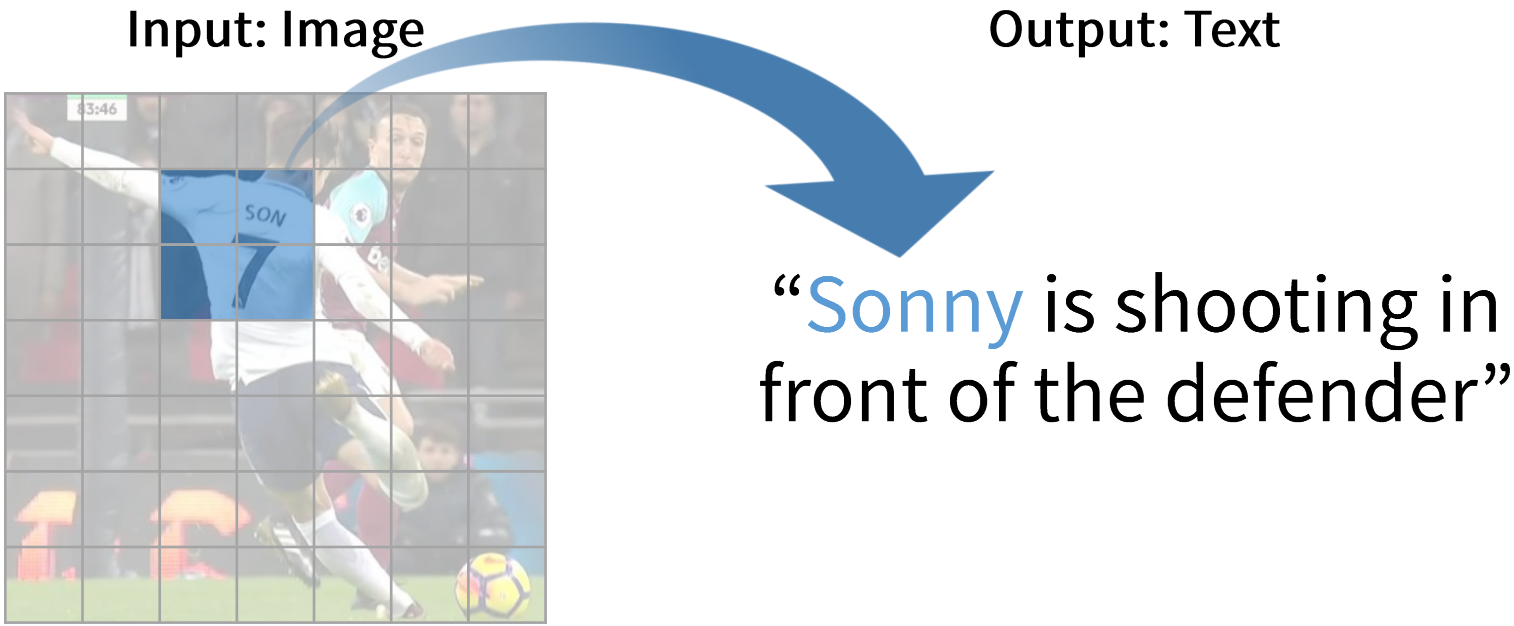


Output: Text

"Sonny is shooting in front of the defender"

❖ Image Captioning

- 이미지를 입력 데이터로 사용하여 이미지를 설명하는 문장을 생성하는 문제

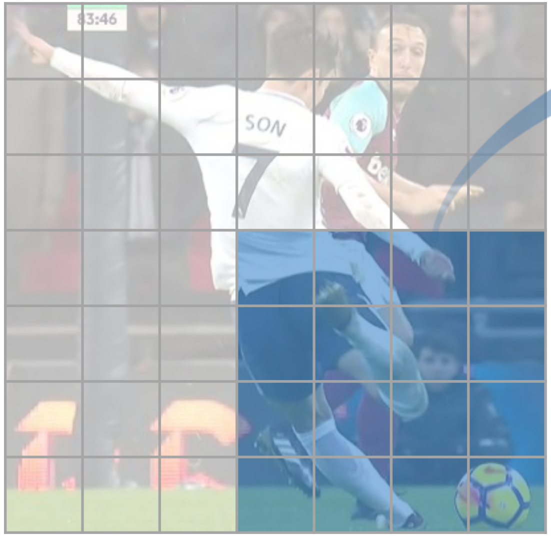


Data Mining Quality Analytics

❖ Image Captioning

- 이미지를 입력 데이터로 사용하여 이미지를 설명하는 문장을 생성하는 문제

Input: Image



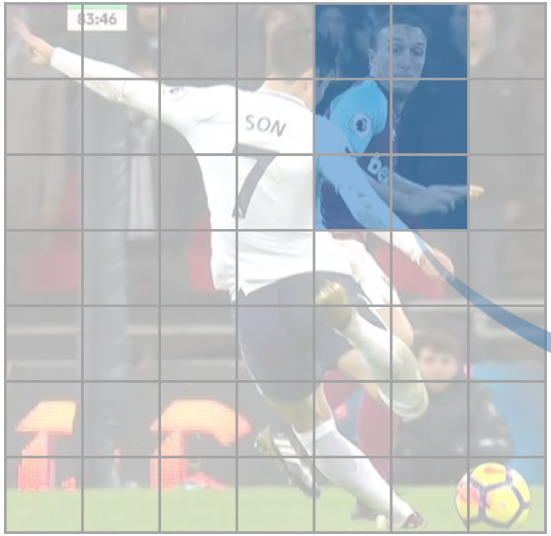
Output: Text

"Sonny is shooting in front of the defender"

❖ Image Captioning

- 이미지를 입력 데이터로 사용하여 이미지를 설명하는 문장을 생성하는 문제

Input: Image



Output: Text

"Sonny is shooting in front of the defender"

Data Mining
Quality Analytics

❖ Paper Review

- Show, Attend and Tell
- Conference on machine learning
- Published in 2015
- Xu, Kelvin, et al.
- Kyunghyun Cho, Yoshua Bengio…

[PDF] **Show**, **attend** and **tell**: Neural image caption generation with visual attention
K Xu, J Ba, R Kiros, K Cho, A Courville… - … conference on machine …, 2015 - jmlr.org
Inspired by recent work in machine translation and object detection, we introduce an attention based model that automatically learns to describe the content of images. We describe how we can train this model in a deterministic manner using standard …

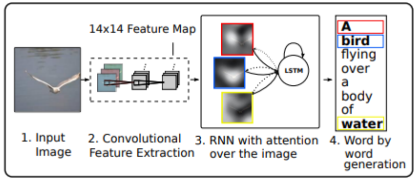
Show, Attend and Tell: Neural Image Caption Generation with Visual Attention

Kelvin Xu	KELVIN.XU@UMONTREAL.CA
Jimmy Lei Ba	JIMMY@PSI.UTORONTO.CA
Ryan Kiros	RKIROS@CS.TORONTO.EDU
Kyunghyun Cho	KYUNGHYUN.CHO@UMONTREAL.CA
Aaron Courville	AARON.COURVILLE@UMONTREAL.CA
Ruslan Salakhutdinov	RSALAKHU@CS.TORONTO.EDU
Richard S. Zemel	ZEMEL@CS.TORONTO.EDU
Yoshua Bengio	FIND-ME@THE.WEB

Abstract

Inspired by recent work in machine translation and object detection, we introduce an attention based model that automatically learns to describe the content of images. We describe how we can train this model in a deterministic manner using standard backpropagation techniques and stochastically by maximizing a variational lower bound. We also show through visualization how the model is able to automatically learn to fix its gaze on salient objects while generating the corresponding words in the output sequence. We validate the use of attention with state-of-the-art performance on three benchmark datasets: Flickr8k, Flickr30k and MS COCO.

Figure 1. Our model learns a words/image alignment. The visualized attentional maps (3) are explained in section 3.1 & 5.4

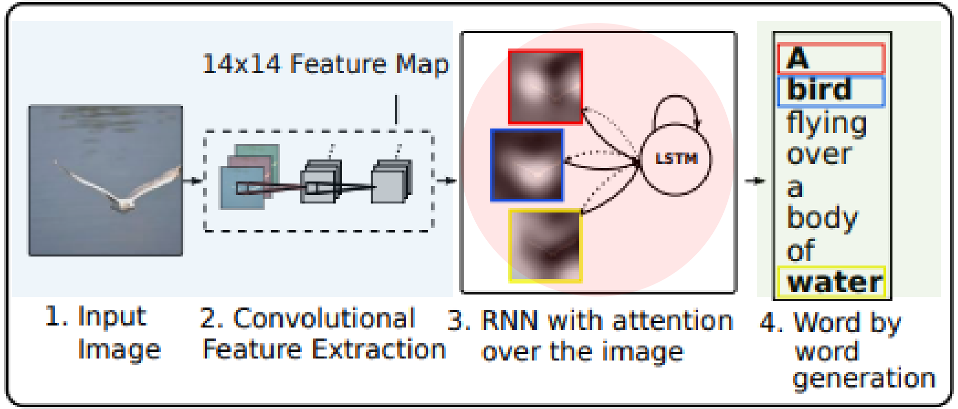


has significantly improved the quality of caption generation using a combination of convolutional neural networks (convnets) to obtain vectorial representation of images and

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

Data Mining Quality Analytics

❖ Model Architecture



Encoder – Feature Extraction

Convolutional Neural Network 모델으로 이미지의 특징을 잘 요약하는 feature map 생성

Attention

Decoder에서 단어를 출력할 때 단어와 가장 유사한 feature map과의 attention score을 계산

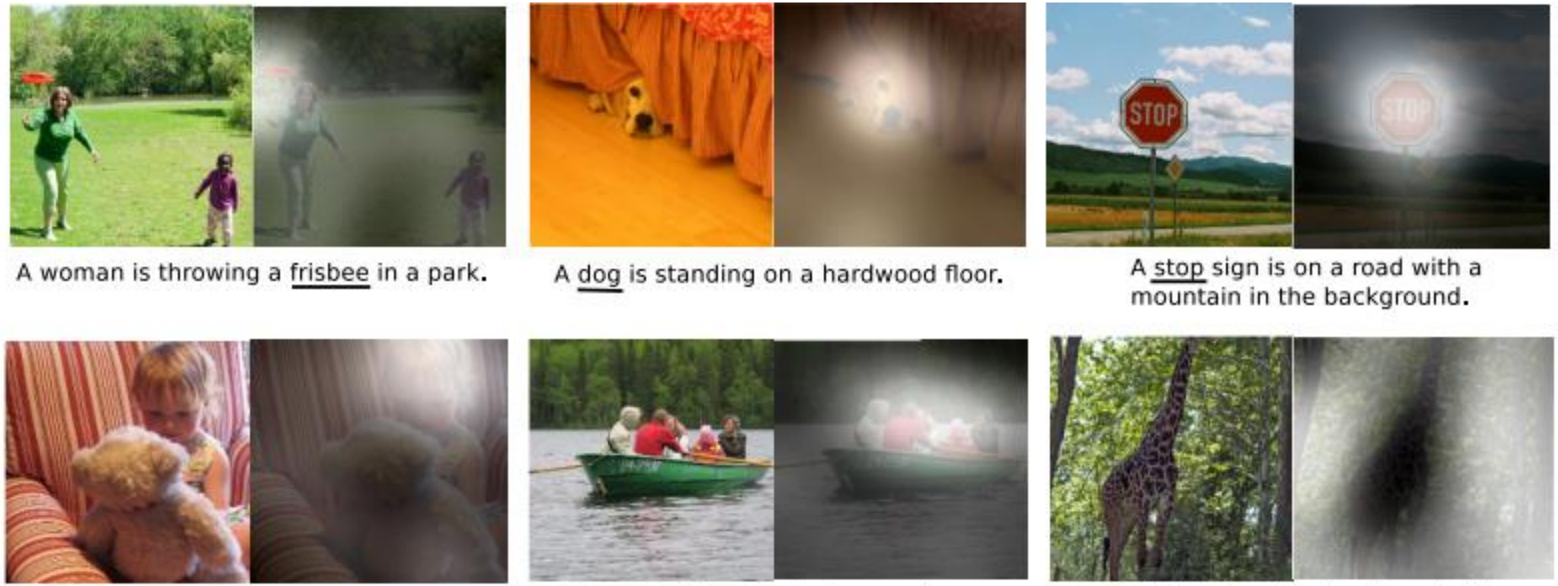
Decoder – Word Generation

이전 시점에 출력된 단어의 정보, feature map, attention score를 decoder의 input으로 사용하여 순차적으로 단어를 출력

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

Data Mining
Quality Analytics

❖ Model Architecture



A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.

A little girl sitting on a bed with a teddy bear.

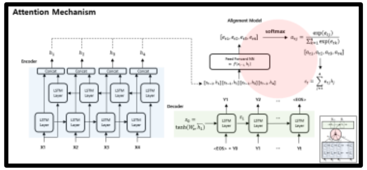
A group of people sitting on a boat in the water.

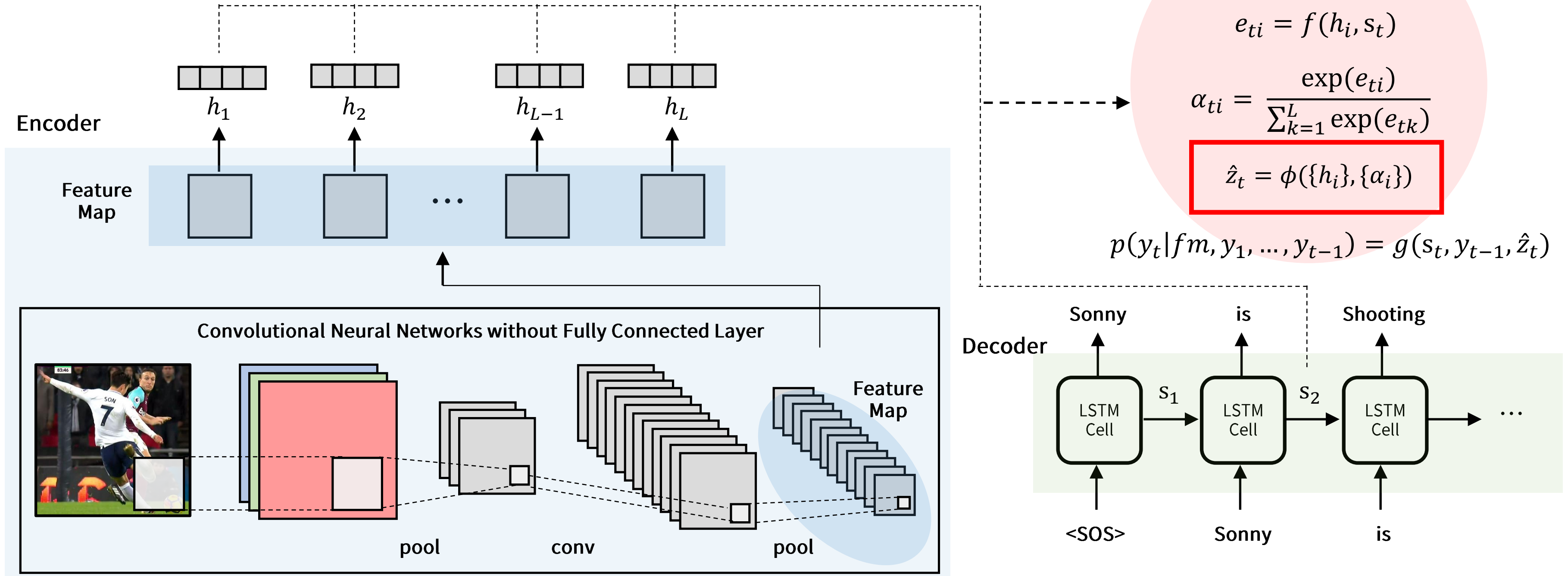
A giraffe standing in a forest with trees in the background.

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

Data Mining Quality Analytics

❖ Model Architecture





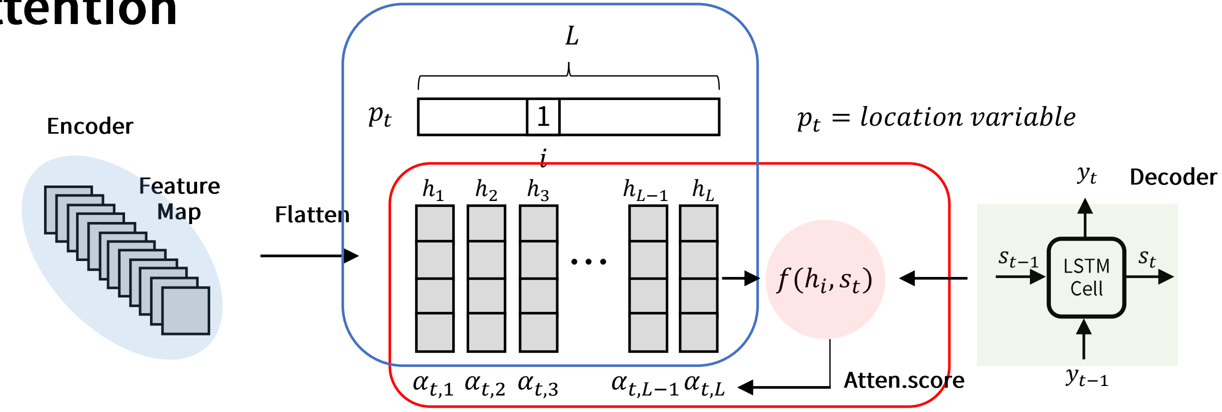
Xu, Kelvin, et al. "Show, attend and tell: Neural image caption generation with visual attention." *International conference on machine learning*. 2015.

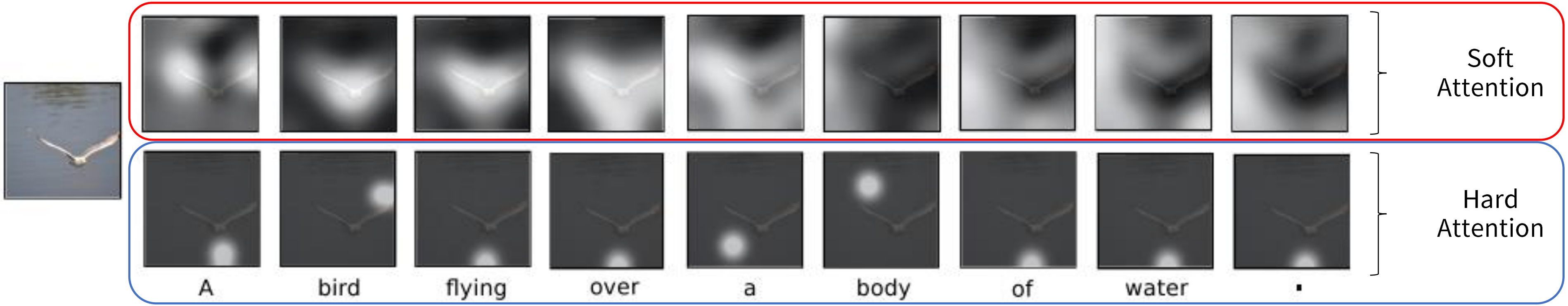
❖ Soft Attention vs Hard Attention

Soft Attention $\hat{z}_t = \phi(\{h_i\},\{\alpha_i\}) = \sum_{k=1}^{L} \alpha_i h_i$

Hard Attention $\hat{z}_t = \phi(\{h_i\},\{\alpha_i\}) = \sum_{i} p_t h_i$

$p(p_{t,i} = 1 | p_{j<t}, h) = \alpha_{t,i}$





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

❖ Soft Attention vs Hard Attention

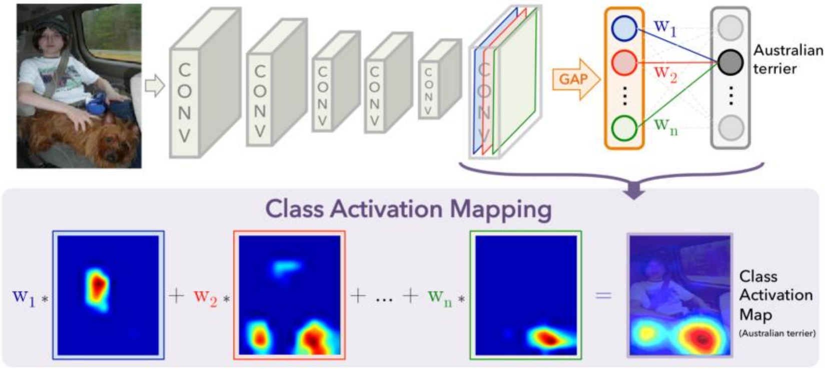
	Soft Attention	Hard Attention
Pros	• Interpretability • End to end learning	• Better performance than soft attention
Cons	• Worse performance than hard attention	• Hard to optimize - Monte Carlo based sampling - REINFORCE

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

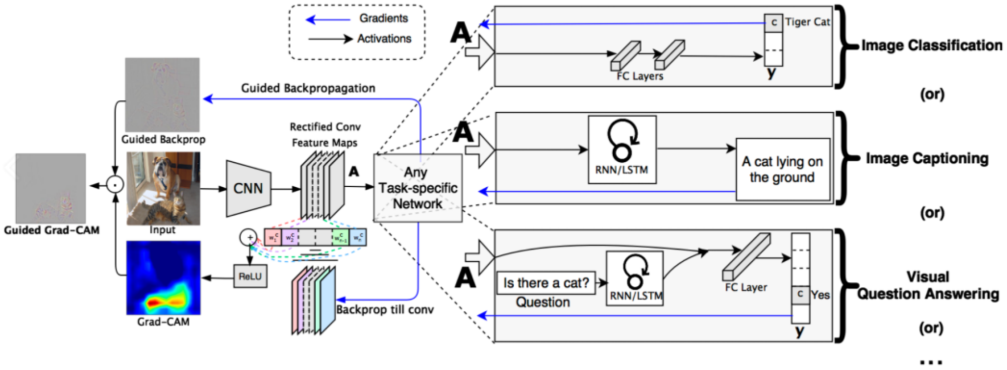
Data Mining
Quality Analytics

❖ Localization vs Attention

- Localization: Interpretability
- Attention: Interpretability + Model performance



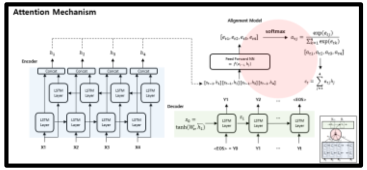
Class Activation Map(CAM)



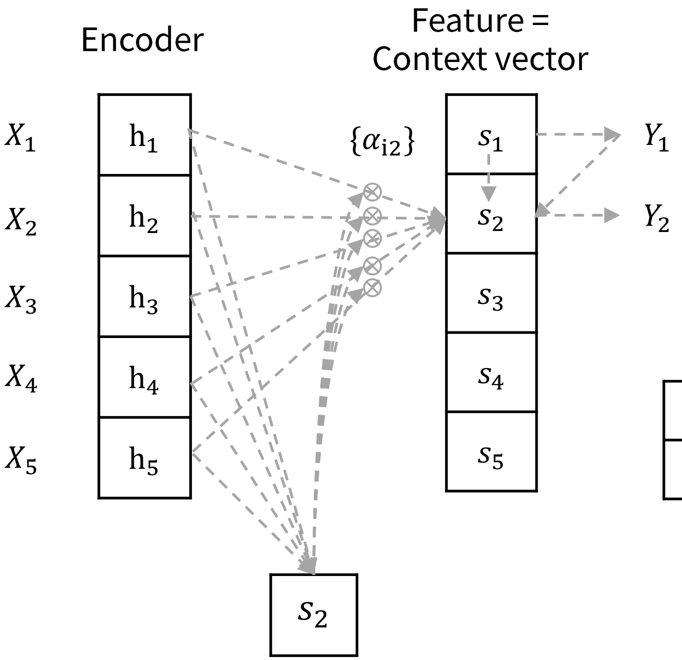
Grad_CAM

Zhou, Bolei, et al. "Learning deep features for discriminative localization." *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2016.
Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." *Proceedings of the IEEE international conference on computer vision.* 2017.

Data Mining
Quality Analytics



❖ Feature Representation



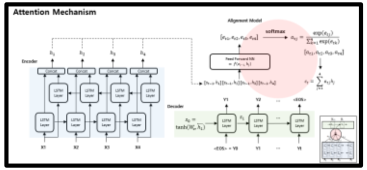
Machine Translation

$$\{\alpha_{i2}\} = softmax(f(\mathrm{h}_i, s_2))$$

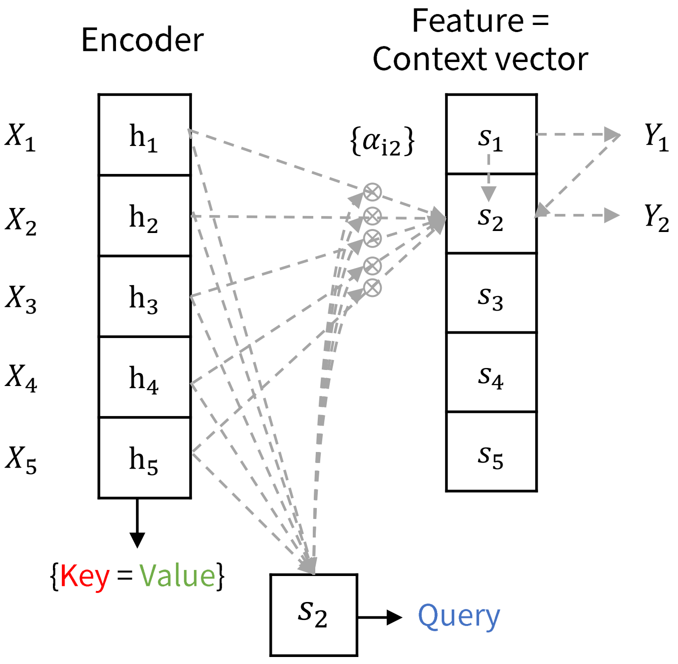
$$feature = \sum_{i=1}^{5} \alpha_{i2} * \mathrm{h}_i$$

Query	Key	Value	Output	Mentee
강현규	백인성	차상원 대리	$0 * int(V_1)$	김태우 과장
	강현규	김태우 과장	$1 * int(V_2)$	
	이영재	조민구 대리	$0 * int(V_3)$	
	조용원	안선희 과장	$0 * int(V_4)$	
	목충협	정지영 사원	$0 * int(V_5)$	

Data Mining Quality Analytics



❖ Feature Representation



Machine Translation

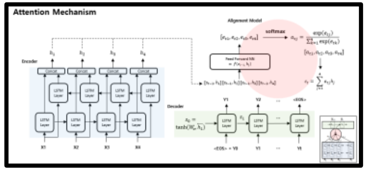
$$\{\alpha_{i2}\} = softmax(f(\mathrm{h}_i, s_2))$$

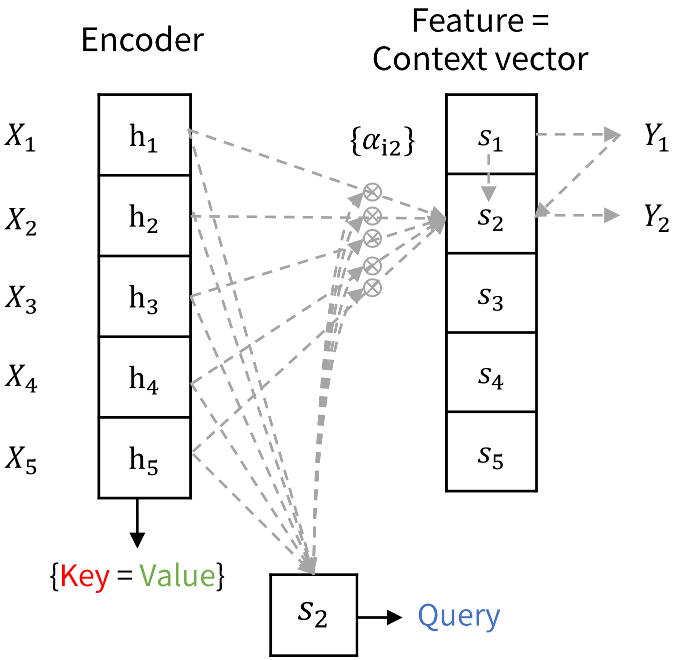
$$feature = \sum_{i=1}^{5} \alpha_{i2} * \mathrm{h}_i$$

Query	Key	Value	Output	Feature
s_1	h_1	h_1	$\alpha_{12} * \mathrm{h}_1$	$\sum_{i=1}^{5} \alpha_{i2} * \mathrm{h}_i$
s_2	h_2	h_2	$\alpha_{22} * \mathrm{h}_2$	
s_3	h_3	h_3	$\alpha_{32} * \mathrm{h}_3$	
s_4	h_4	h_4	$\alpha_{42} * \mathrm{h}_4$	
s_5	h_5	h_5	$\alpha_{52} * \mathrm{h}_5$	

Data Mining Quality Analytics

❖ **Feature Representation**





Machine Translation

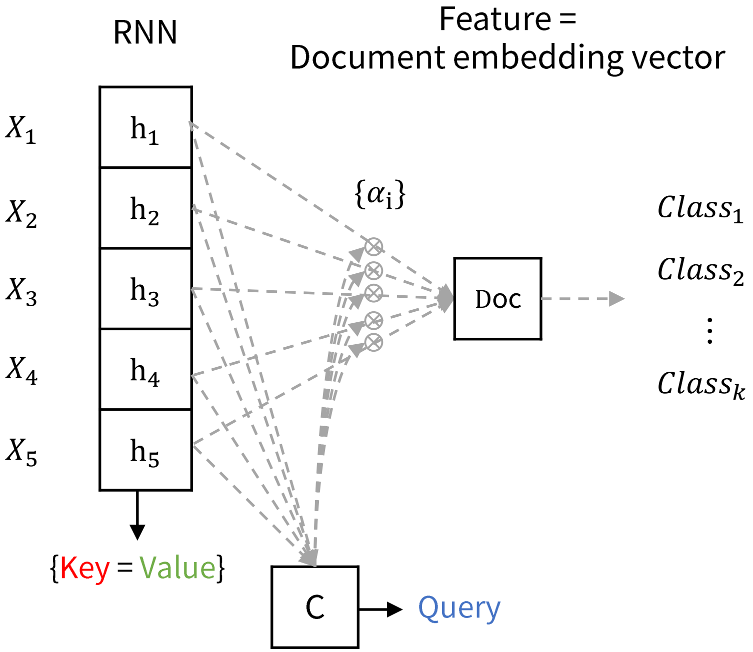
$$\{\alpha_{i2}\} = softmax(f(\mathrm{h}_i, s_2))$$

$$feature = \sum_{i=1}^{5} \alpha_{i2} * \mathrm{h}_i$$

⬇

$$A(q, K, V) = \sum_{\mathrm{i}} softmax(f(K, q))\, V$$

❖ Feature Representation



Document Classification

$$\{\alpha_i\} = softmax(f(h_i, c))$$

$$feature = \sum_{i=1}^{5} \alpha_i * h_i$$

⬇

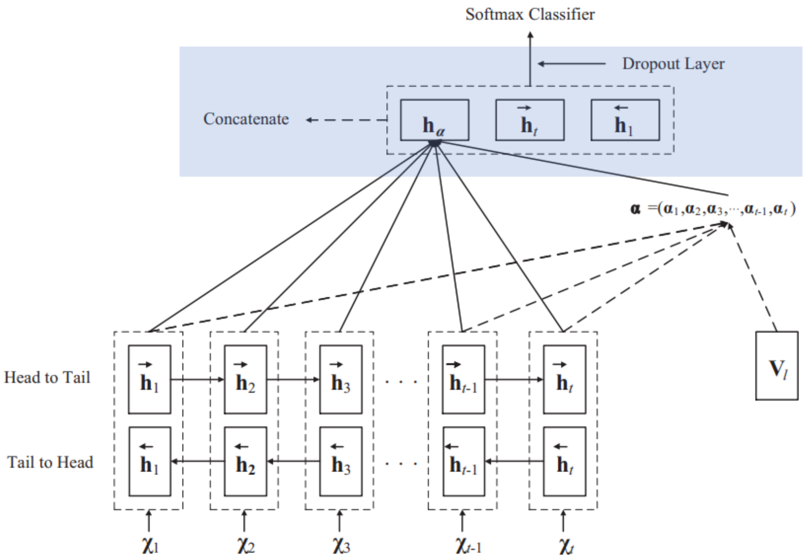
$$A(q, K, V) = \sum_{i} softmax(f(K, q))\, V$$

Generalized Attention Form

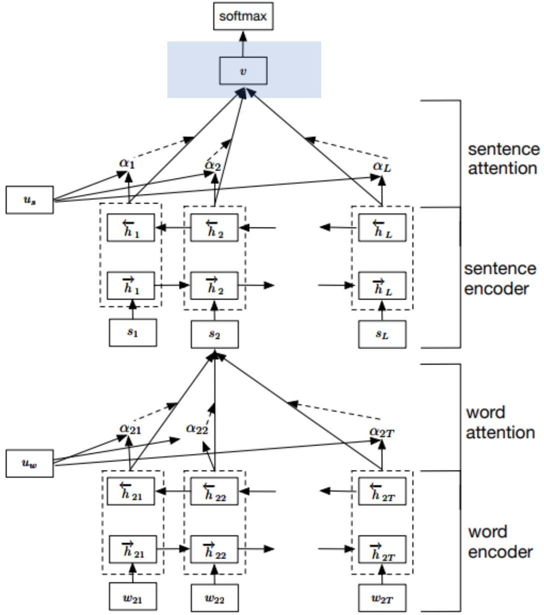
Data Mining
Quality Analytics

❖ Document Classification: RNN-based

- 문서 또는 문장을 보다 더 잘 표현하는 feature를 생성



Bidirectional RNN with Attention



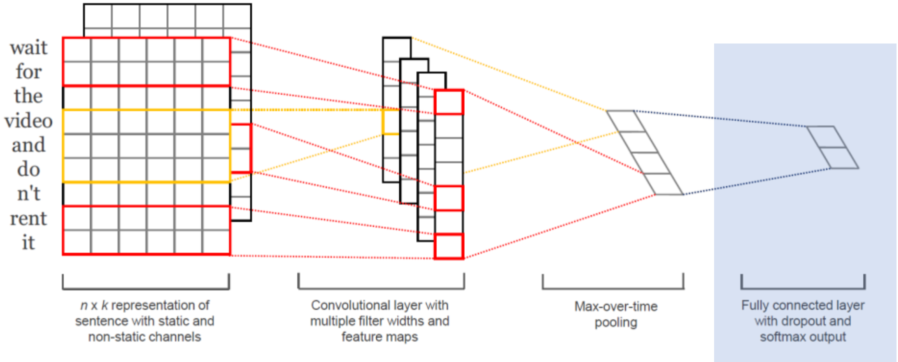
Hierarchical Attention Network

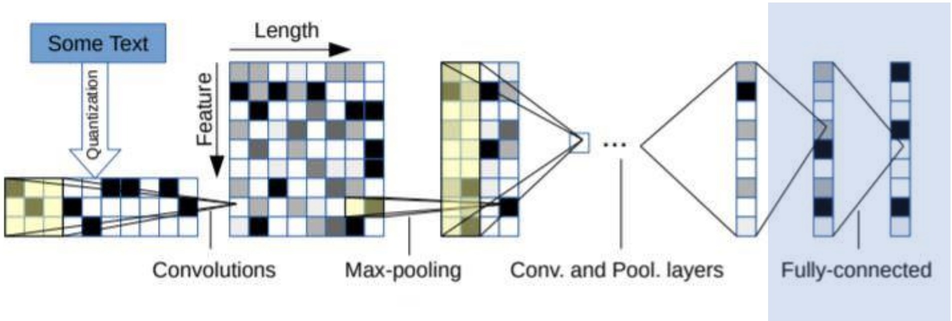
Liu, Tengfei, et al. "Recurrent networks with attention and convolutional networks for sentence representation and classification." *Applied Intelligence* 48.10 (2018): 3797-3806.
Yang, Zichao, et al. "Hierarchical attention networks for document classification." *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics* 2016.

Data Mining Quality Analytics

❖ Document Classification: CNN-based

- 문서 또는 문장을 보다 더 잘 표현하는 feature를 생성





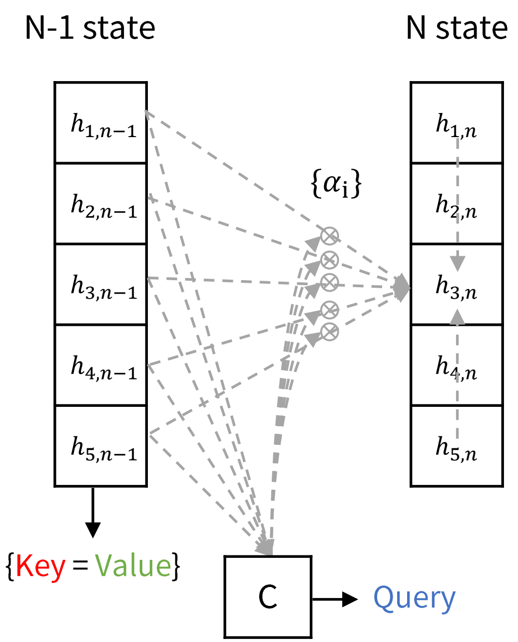
TextCNN

Character-level CNN

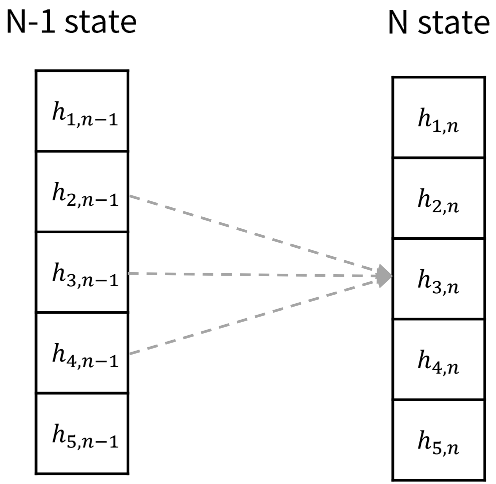
Kim, Yoon. "Convolutional neural networks for sentence classification." *arXiv preprint arXiv:1408.5882* (2014).
Zhang, Xiang, Junbo Zhao, and Yann LeCun. "Character-level convolutional networks for text classification." *Advances in neural information processing systems.* 2015.

❖ Attention to Self-Attention

$$A(q, K, V) = \sum_{\mathrm{i}} softmax(f(K, q))\, V$$



RNN-Based

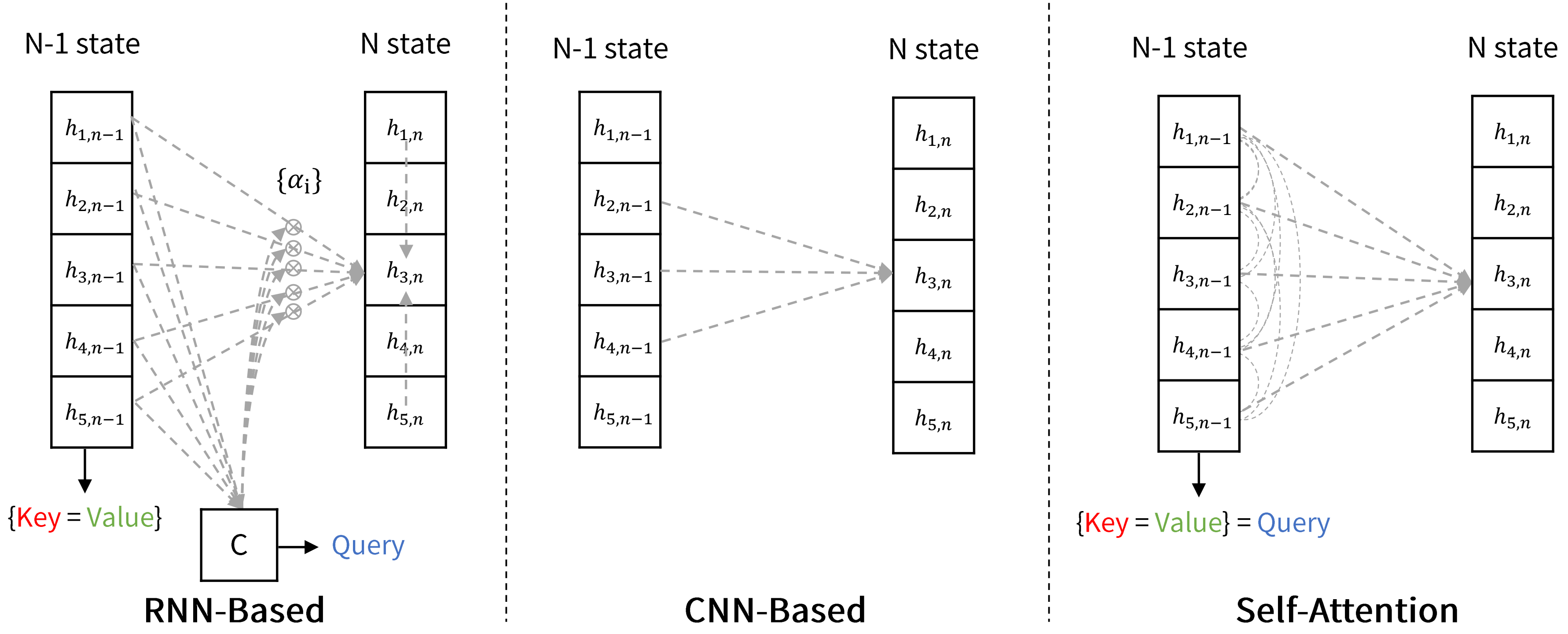


CNN-Based

- RNN-based
  - 입력 시퀀스를 순차적으로 처리하여 병렬처리 어려움
  - 연산 시간, 계산 복잡도 多
- CNN-based
  - 병렬처리는 쉽지만 거리가 먼 단어와의 관계를 학습하기에는 여러 합성곱 연산 필요
- 더 나은 feature 계산 방법?
  - RNN, CNN 구조 사용 X
  - Key = Value = Query
  - f = dot product 사용

❖ Attention to Self-Attention

$$A(q, K, V) = \sum_{i} softmax(f(K, q))\, V$$



Data Mining
Quality Analytics

❖ Paper Review

- Attention is All You Need
- NIPS
- Published in 2017
- A Vaswani, et al, Google Brain, Research
- A Vaswani, N Parmar…

Attention is all you need
A Vaswani, N Shazeer, N Parmar… - Advances in neural …, 2017 - papers.nips.cc
The dominant sequence transduction models are based on complex recurrent orconvolutional neural networks in an encoder and decoder configuration. The best performing such models also connect the encoder and decoder through an attentionm …

Attention Is All You Need

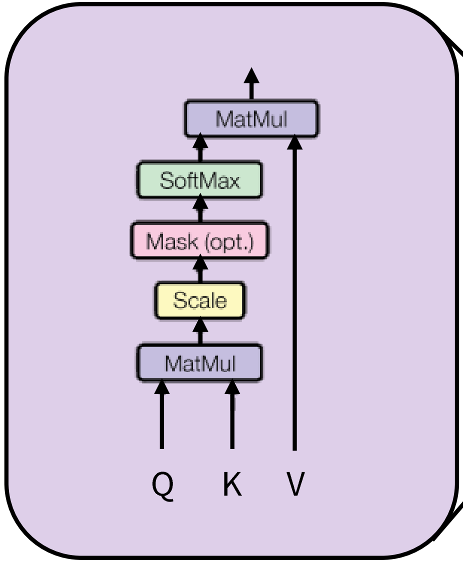
Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

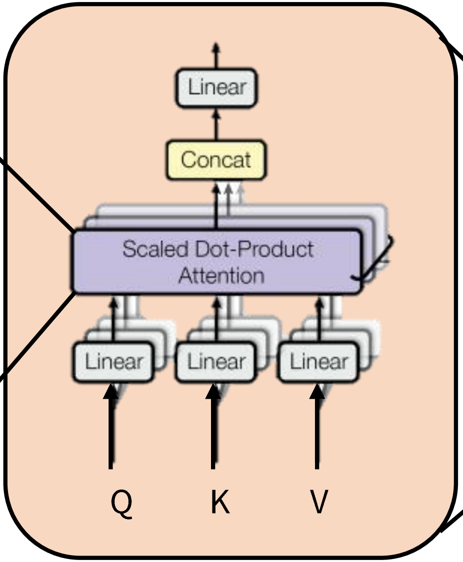
The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.0 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature.

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

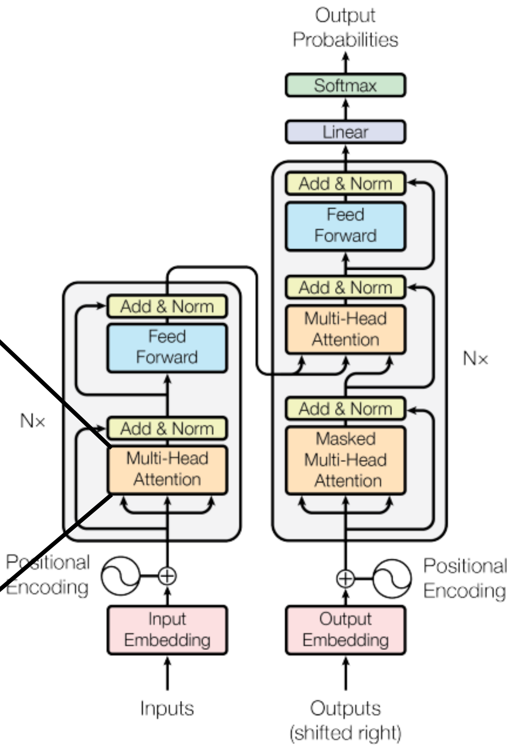
❖ Transformer



Scaled Dot-Product Attention



Multi-Head Attention

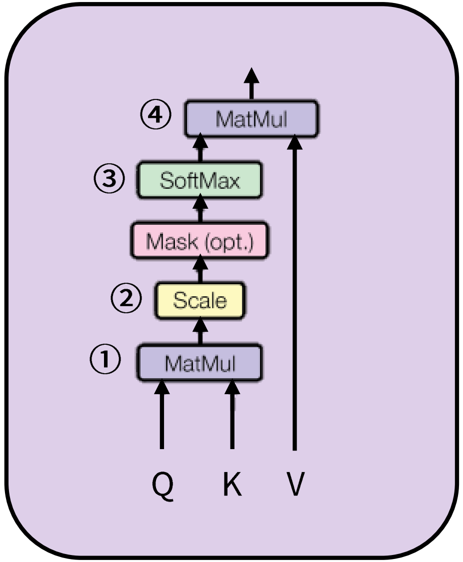


Transformer

http://jalammar.github.io/

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

❖ Transformer



Scaled Dot-Product
Attention

Generalized Attention Form

$$A(q, K, V) = \sum_{i} softmax(f(K, q))\, V$$

① MatMul

$$f(K, Q) = QK^T$$

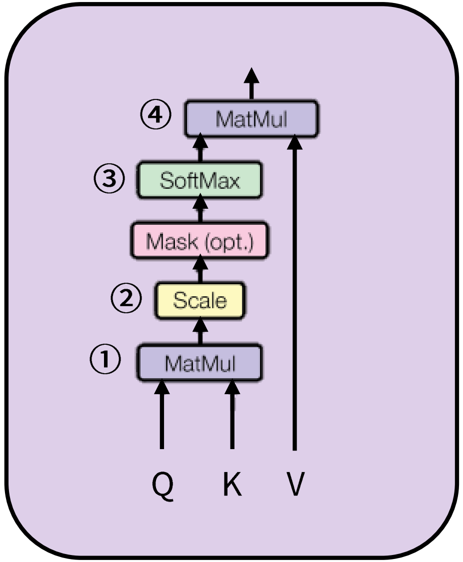
While the two are similar in theoretical complexity, dot-product attention is much faster and more space-efficient in practice, since it can be implemented using highly optimized matrix multiplication code. (⋯)

http://jalammar.github.io/

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

Data Mining Quality Analytics

❖ Transformer



Scaled Dot-Product Attention

Generalized Attention Form

$$A(q, K, V) = \sum_{i} softmax(f(K, q))\, V$$

① MatMul $\quad f(K, Q) = QK^T$

② Scale $\quad \frac{QK^T}{\sqrt{d_k}}$

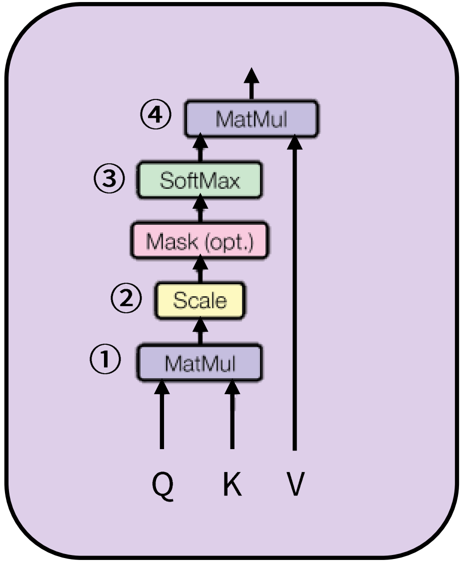
We suspect that for large values of dk, the dot products grow large in magnitude, pushing the softmax function into regions where it has extremely small gradients . To counteract this effect, we scale the dot products (···)

http://jalammar.github.io/

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

Data Mining Quality Analytics

❖ Transformer



Scaled Dot-Product Attention

Generalized Attention Form

$$A(q,K,V) = \sum_{i} softmax(f(K,q))\,V$$

① MatMul: $f(K,Q) = QK^T$

② Scale: $\frac{QK^T}{\sqrt{d_k}}$

③ Softmax: $softmax(\frac{QK^T}{\sqrt{d_k}})$

④ MatMul: $softmax(\frac{QK^T}{\sqrt{d_k}})V$

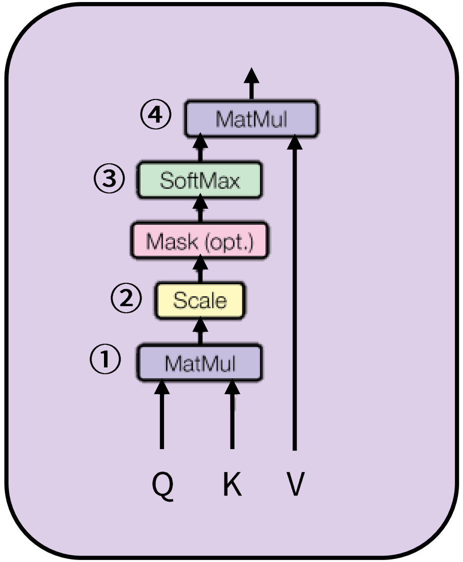
Self-Attention

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

http://jalammar.github.io/

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

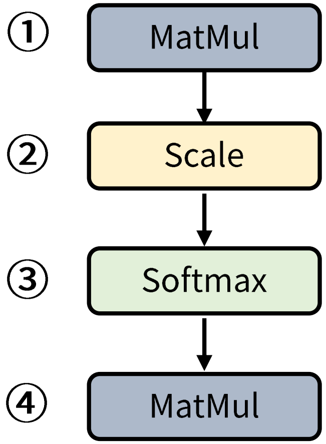
❖ Transformer

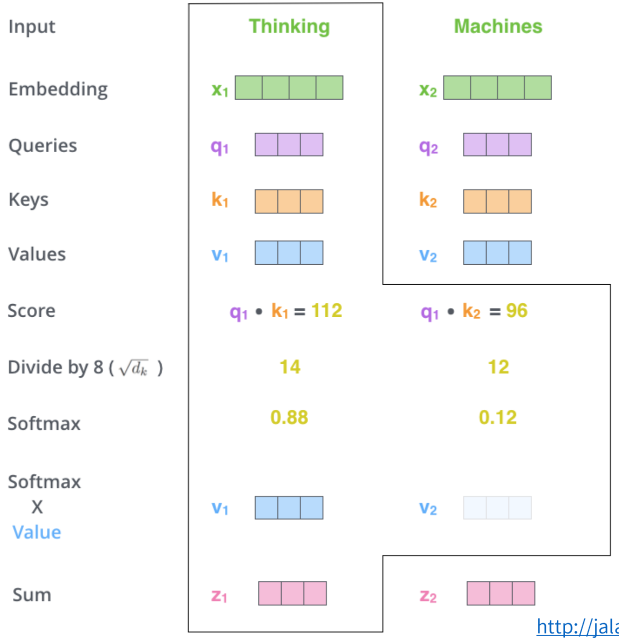


Scaled Dot-Product Attention

Self-Attention

$$A(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$





http://jalammar.github.io/

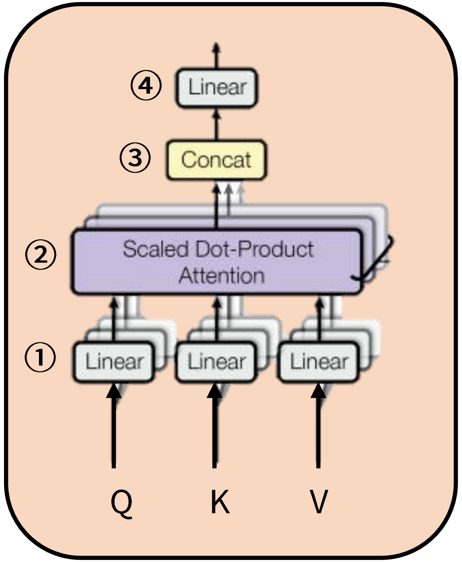
Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems*. 2017..

Data Mining Quality Analytics

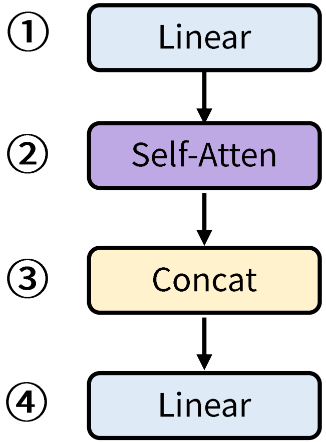
❖ Transformer

Self-Attention

$$A(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$



Multi-Head Attention



Multi-Head Attention

$$Q` = QW_i^Q \quad K` = KW_i^K \quad V` = VW_i^V$$

$$(i = 1 \dots h)$$

$$\text{head}_i = A(Q`, K`, V`)$$

$$[\text{head}_1, \text{head}_2, \dots, \text{head}_h]$$

$$[\text{head}_1, \text{head}_2, \dots, \text{head}_h]W^O$$

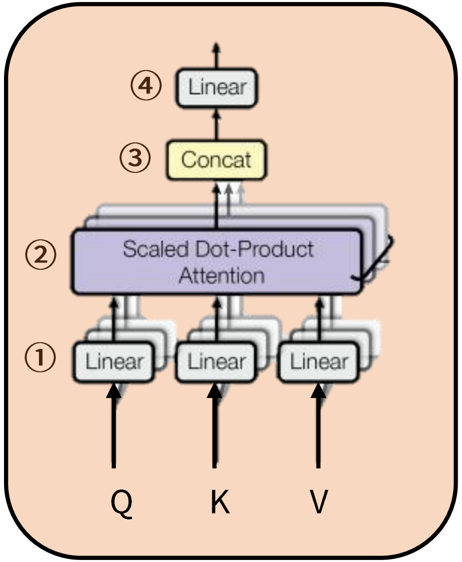
$$MultiHead(Q,K,V) = [\text{head}_1, \text{head}_2, \dots, \text{head}_h]W^O$$

http://jalammar.github.io/

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

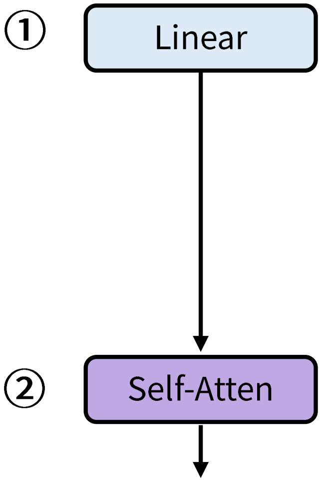
Data Mining
Quality Analytics

❖ Transformer

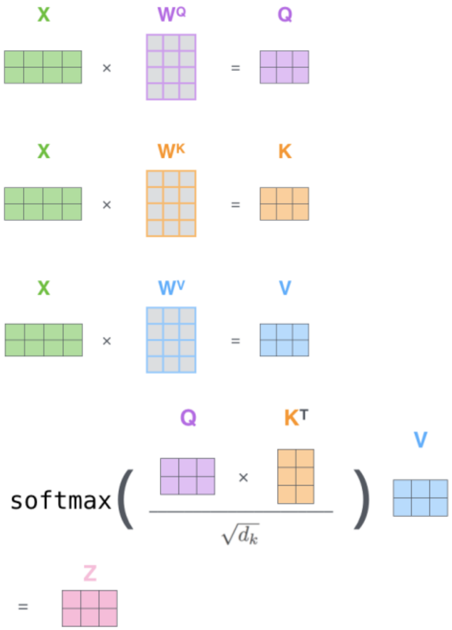


Multi-Head Attention

Self-Attention



$$A(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$



http://jalammar.github.io/

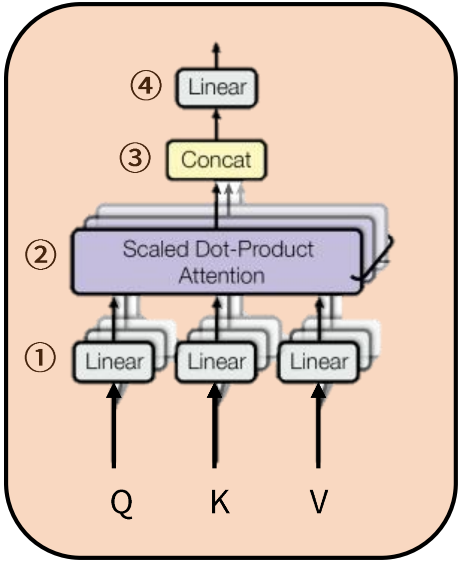
Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems.* 2017..

Data Mining Quality Analytics

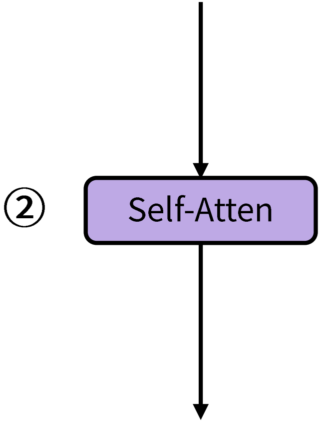
❖ **Transformer**

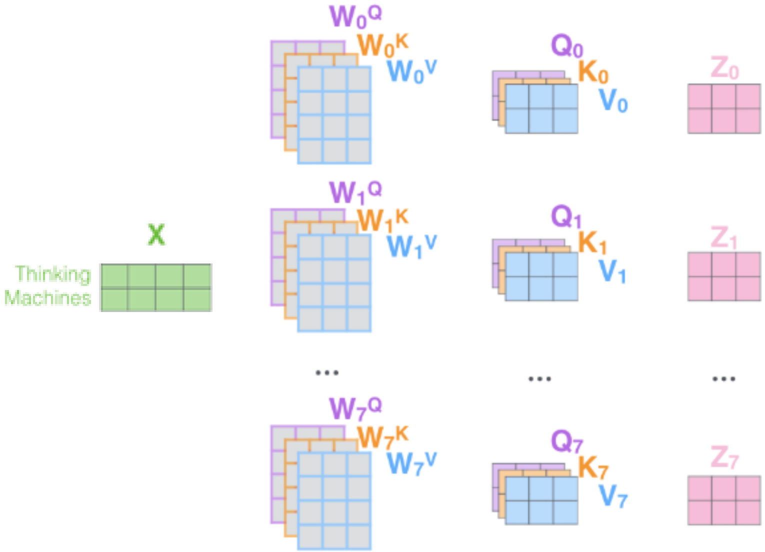
Self-Attention

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



Multi-Head Attention





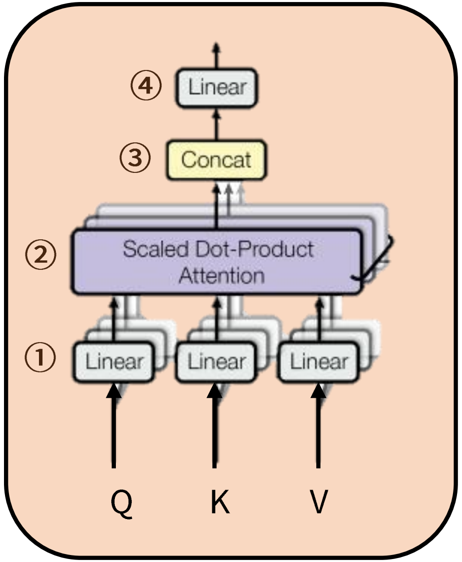
http://jalammar.github.io/

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

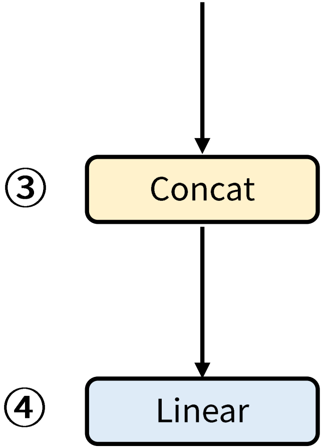
❖ Transformer

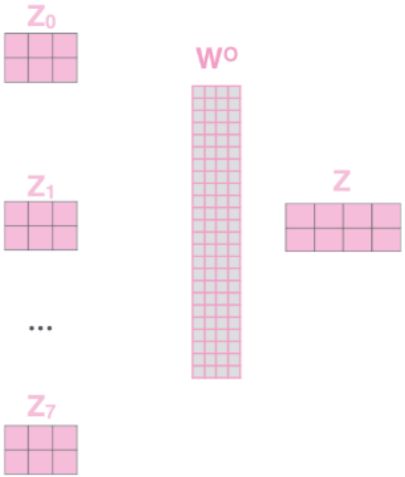
Self-
Attention

$$A(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$



Multi-Head
Attention





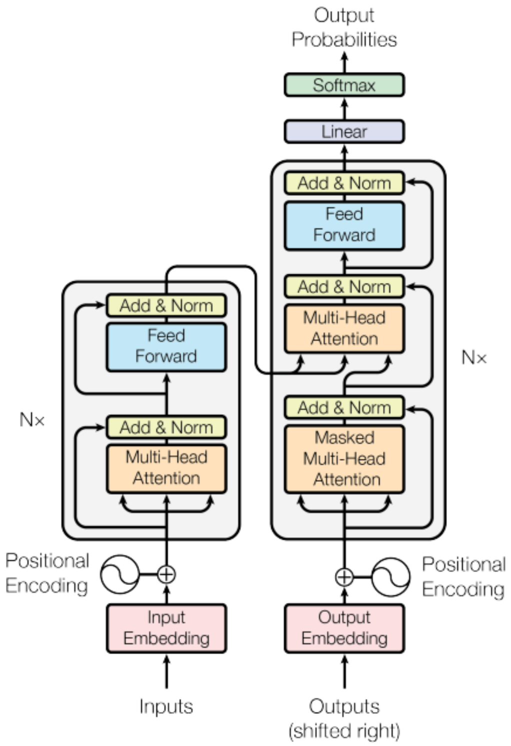
http://jalammar.github.io/

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

Data Mining
Quality Analytics

❖ Contributions

- One is the total computational complexity per layer. (···)
- Another is the amount of computation that can be parallelized, (···)
- The third is the path length between long-range dependencies in the network. Learning long-range dependencies is a key challenge in many sequence transduction tasks. (···)
- As side benefit, self-attention could yield more interpretable models.



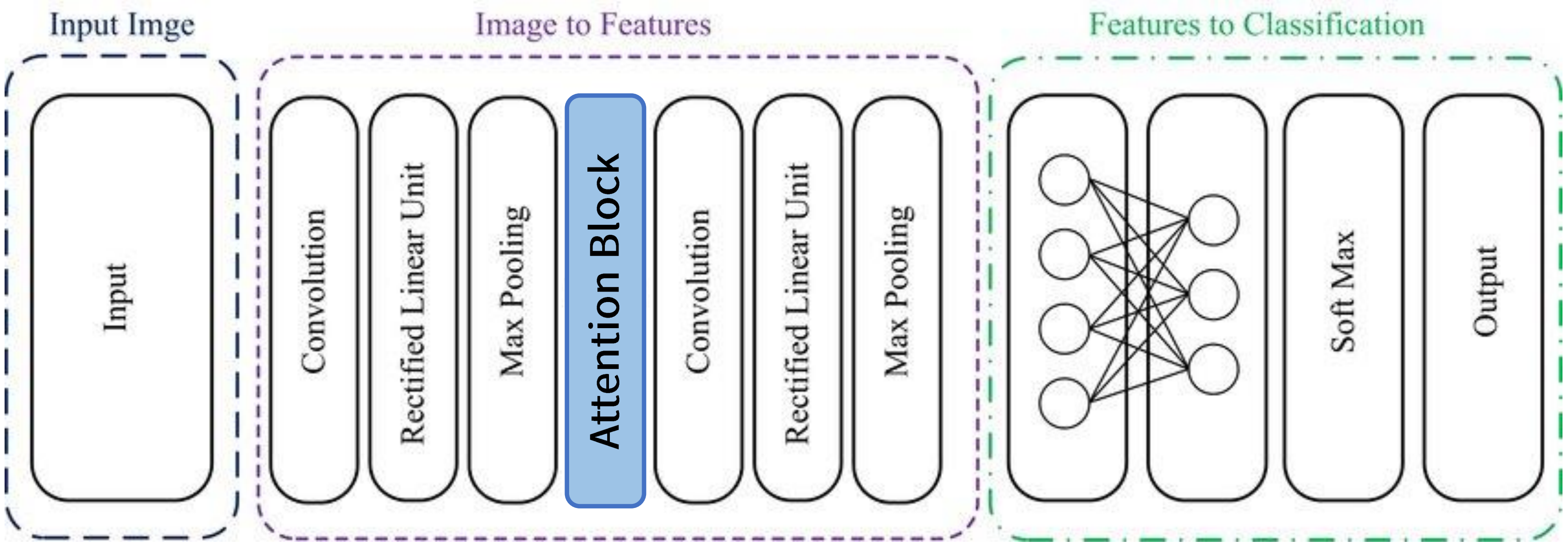
Transformer http://jalammar.github.io/

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

Data Mining
Quality Analytics

❖ Attention: CNN Architecture

- 이미지의 feature을 보다 더 잘 생성하기 위해 convolution layer 사이에 attention block 삽입



https://www.researchgate.net/publication/333112125_An_islanding_detection_method_based_on_image_classification_with_convolution_neural_network

❖ Various Attention Algorithms: CNN

- 이미지의 feature을 보다 더 잘 생성하기 위해 convolution layer 사이에 attention block 삽입

Squeeze-and-excitation networks
J Hu, L Shen, G Sun - … of the IEEE conference on computer …, 2018 - openaccess.thecvf.com
Convolutional neural networks are built upon the convolution operation, which extracts informative features by fusing spatial and channel-wise information together within local receptive fields. In order to boost the representational power of a network, several recent …
☆ 99 2267회 인용 관련 학술자료 전체 16개의 버전 »

- Jie Hu et al., University of Oxford
- Computer Vision and Pattern Recognition(CVPR) in 2018
- Channel attention
- Winner of ILSVRC 2017

Cbam: Convolutional block attention module
S Woo, J Park, JY Lee… - Proceedings of the …, 2018 - openaccess.thecvf.com
Abstract We propose **Convolutional Block Attention Module** (CBAM), a simple and effective **attention module** that can be integrated with any feed-forward **convolutional** neural networks. Given an intermediate feature map, our **module** sequentially infers **attention** maps …
☆ 99 378회 인용 관련 학술자료 전체 9개의 버전 »

- S Woo, J Park et al., KAIST, LUNIT
- European Conference on Computer Vision(ECCV) in 2018
- Channel attention + Spatial attention

Non-local neural networks
X Wang, R Girshick, A Gupta… - Proceedings of the IEEE …, 2018 - openaccess.thecvf.com
Both convolutional and recurrent operations are building blocks that process one local neighborhood at a time. In this paper, we present non-local operations as a generic family of building blocks for capturing long-range dependencies. Inspired by the classical non-local …
☆ 99 871회 인용 관련 학술자료 전체 10개의 버전 »

- X Wang, Abhinav Gupta, Carnegie Mellon University
- Ross Girshick, Kaiming He, Facebook AI Research
- Computer Vision and Pattern Recognition(CVPR) in 2018
- Non-local self-attention

Stand-alone self-attention in vision models
P Ramachandran, N Parmar, A Vaswani, I Bello… - arXiv preprint arXiv …, 2019 - arxiv.org
Convolutions are a fundamental building block of modern computer vision systems. Recent approaches have argued for going beyond convolutions in order to capture long-range dependencies. These efforts focus on augmenting convolutional models with content-based …
☆ 99 10회 인용 관련 학술자료 전체 5개의 버전 »

- X Wang, Abhinav Gupta, Carnegie Mellon University
- P. Ramachandran, A. Vaswani et al,. Google Research, Brain
- arXiv in 2019
- Local self-attention

Data Mining Quality Analytics

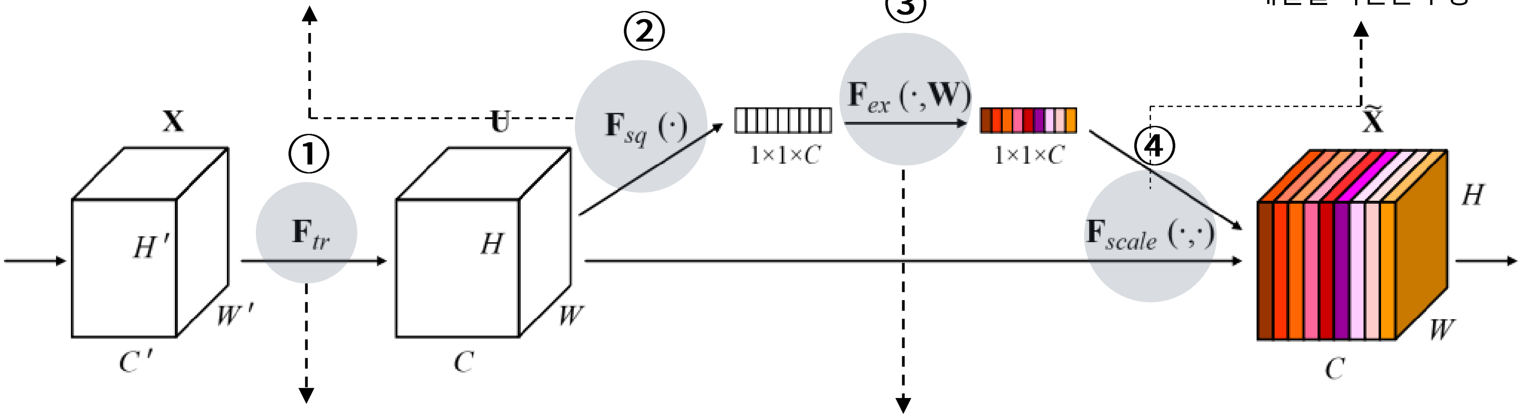
❖ Squeeze-and-Excitation Networks(SENet) - 2018

$$F_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i,j) = z_c$$

② Channel-wise global average pooling을 통해 채널별 특성 추출

$$F_{\text{scale}}(u_c, s_c) = s_c u_c = \widetilde{x_c}$$

④ 이전 layer에 채널 특성 벡터의 dot product를 통한 채널별 어텐션 수행



$$F_{\text{tr}}(\text{X}) = \text{v}_c * \text{X} = u_c$$

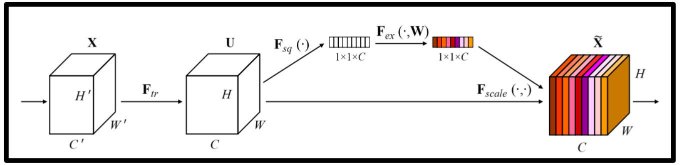
① Conv 연산을 통해 차원 변환

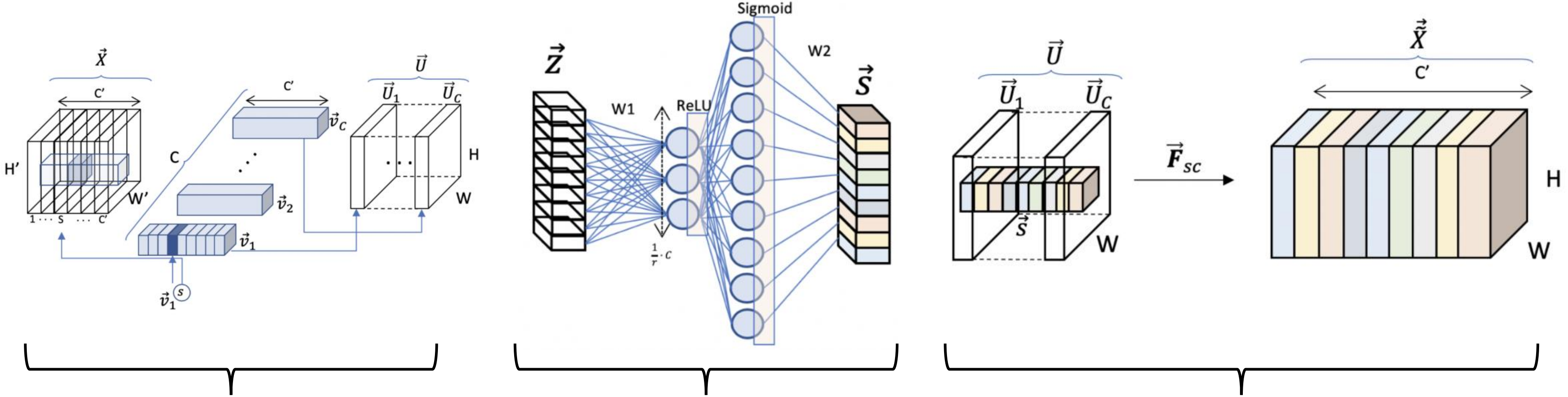
$$F_{ex}(z, W) = sigmoid(W_2 Relu(W_1 z) = s$$

③ 채널별 특성의 관계를 학습하는 과정을 거쳐 채널 특성 벡터를 추출

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

❖ Squeeze-and-Excitation Networks(SENet) - 2018



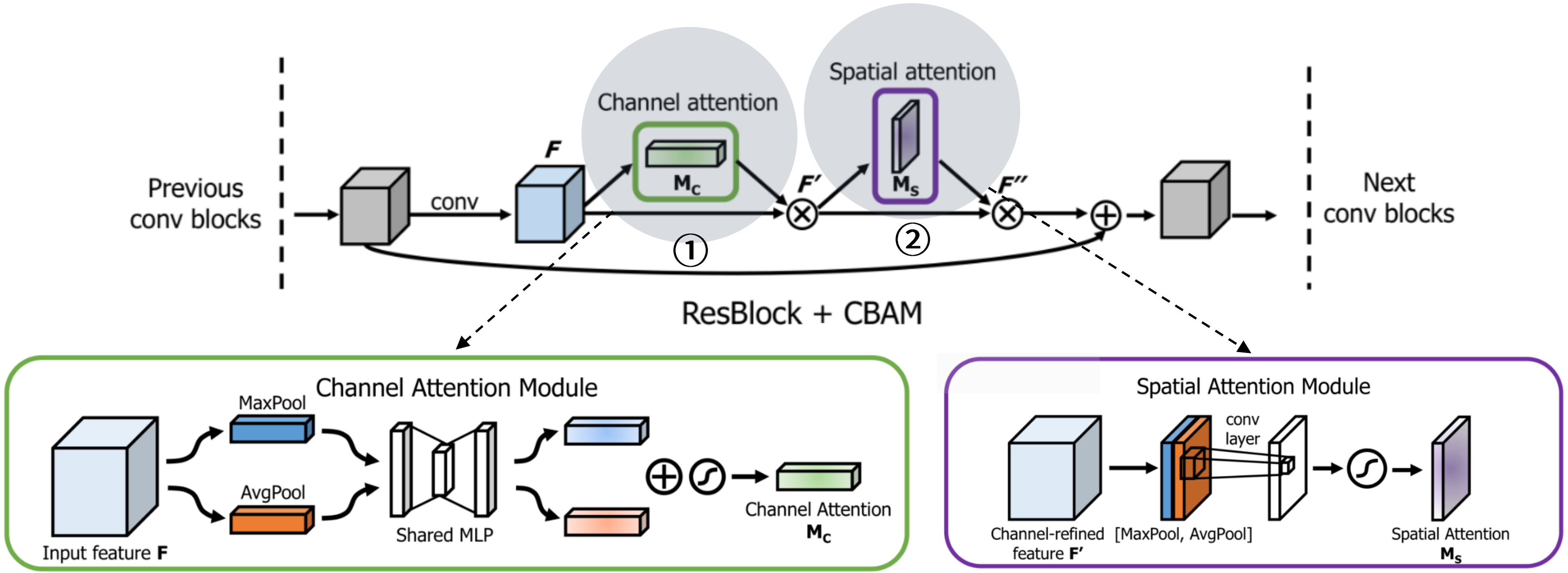


① Transform ② Squeeze ③ Excitation ④ Scaling

https://towardsdatascience.com/understanding-and-visualizing-se-nets-1544aff0fc68

Data Mining Quality Analytics

❖ Convolutional Block Attention Module(CBAM) - 2018

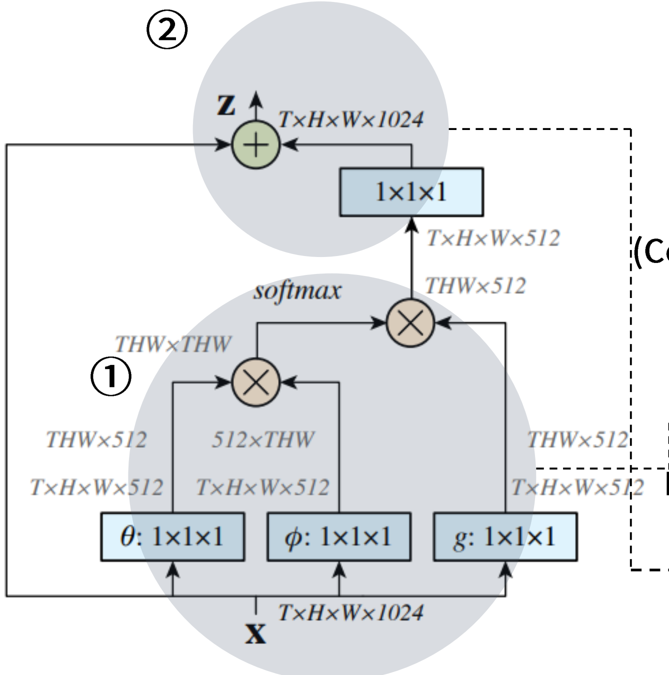


Woo, Sanghyun, et al. "Cbam: Convolutional block attention module." *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018.

Data Mining Quality Analytics

❖ Non-Local Neural Networks - 2018

- The third is the path length between long-range dependencies in the network. Learning long-range dependencies is a key challenge in many sequence transduction tasks. (⋯)



Non-local layer

Local Layer (Convolution)

$$y_{ij} = \sum_{a,b \in N} W_{i-a,j-b}, x_{ab}$$

Non-local layer

$$① \; y_i = \frac{1}{C(x)} \sum_{\forall j} softmax(f(\theta(x_i), \phi(x_j)))g(x_j)$$

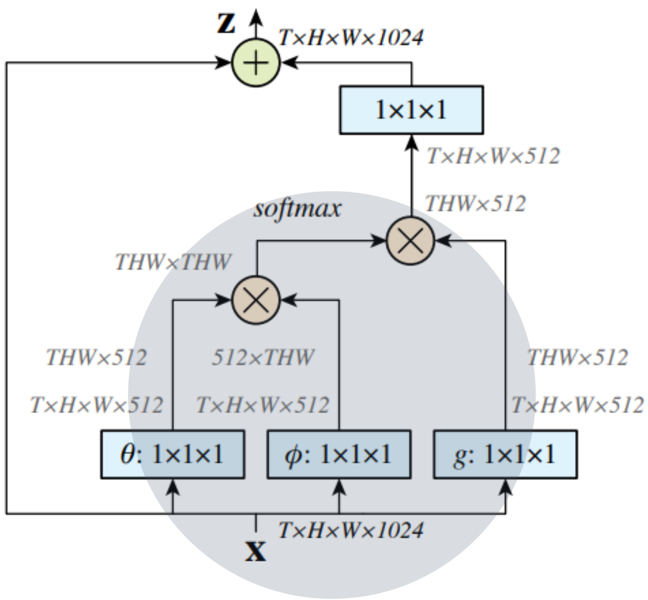
$$② \; z_i = W_z y_i + x_i$$

Wang, Xiaolong, et al. "Non-local neural networks." *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2018.

Data Mining
Quality Analytics

❖ Non-Local Neural Networks - 2018

- The third is the path length between long-range dependencies in the network. Learning long-range dependencies is a key challenge in many sequence transduction tasks. (⋯)



Non-local layer

Non-local layer

$$y_i = \frac{1}{C(x)} \sum_{\forall j} softmax(f(\theta(x_i), \phi(x_j)))g(x_j)$$

$C(x) = 1 \quad \theta, \phi, g = Linear\ function \quad f = dot\ product$

$y_i, x_i, x_j \rightarrow Y, X, X$

$$Y = softmax(X^T W_\theta^T W_\phi X) W_g X$$

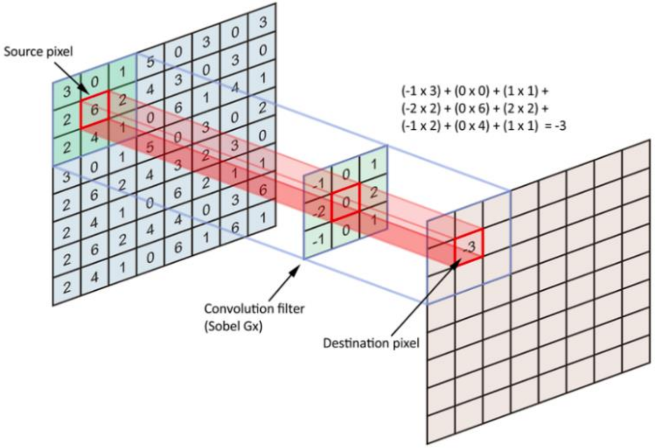
Self Attention

$$A(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Wang, Xiaolong, et al. "Non-local neural networks." *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2018.

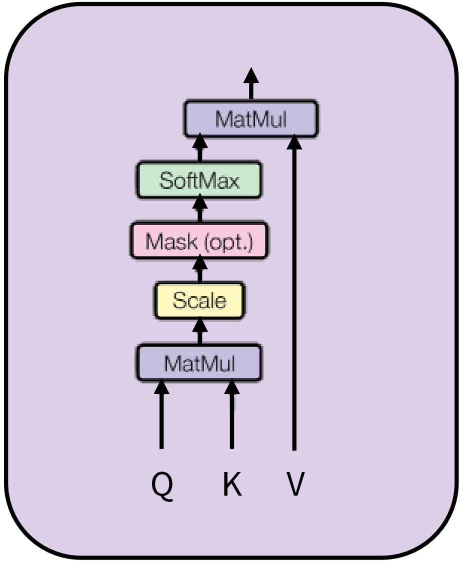
Data Mining
Quality Analytics

❖ Stand-Alone Self-Attention - 2019



$$y_{ij} = \sum_{a,b \in N(i,j)} W_{i-a,j-b} x_{ab}$$

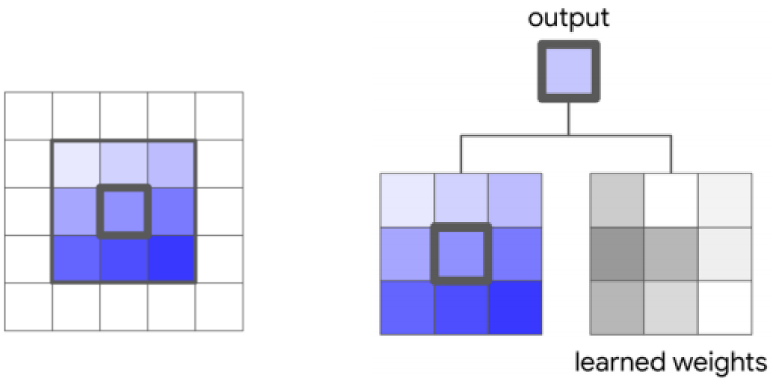
Convolution

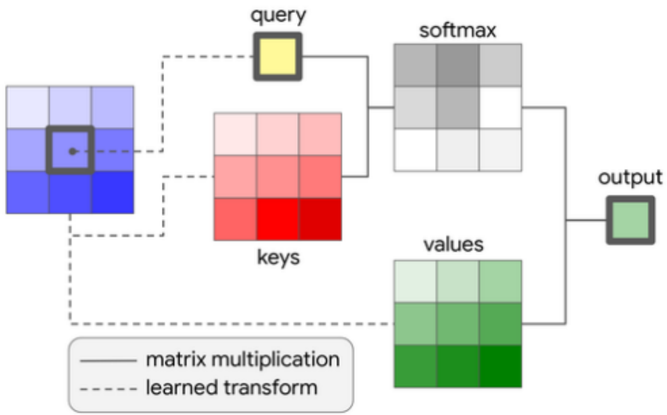


Self-Attention

Ramachandran, Prajit, et al. "Stand-alone self-attention in vision models." *arXiv preprint arXiv:1906.05909* (2019).

❖ Stand-Alone Self-Attention - 2019





$$y_{ij} = \sum_{a,b \in N(i,j)} W_{i-a,j-b} x_{ab}$$

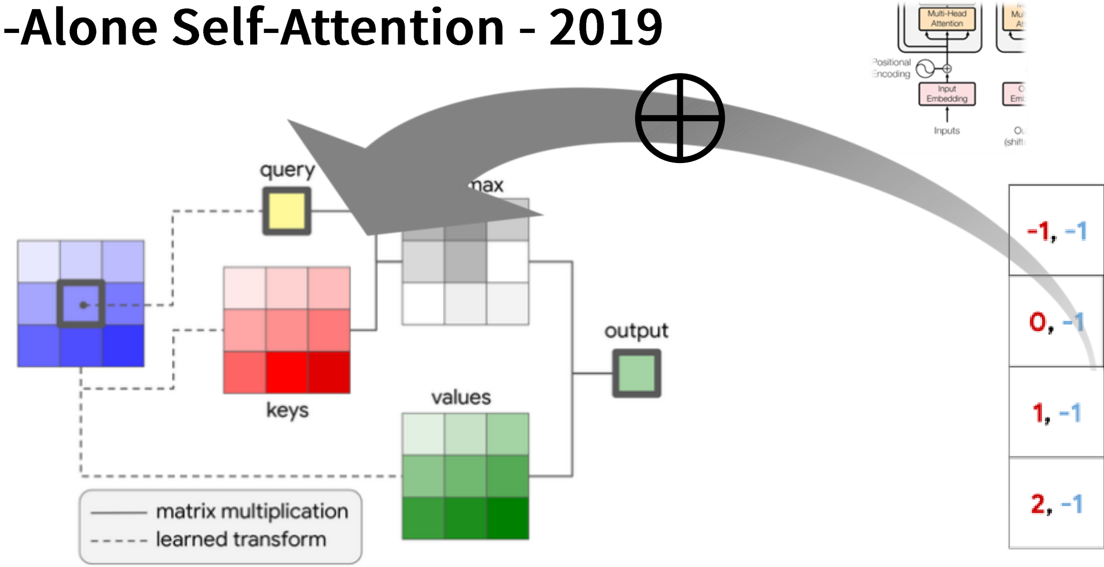
$$y_{ij} = \sum_{a,b \in N(i,j)} softmax_{ab}(q_{ij}^T k_{ab}) v_{ab}$$

$$q_{ij} = W_Q x_{ij} \quad k_{ab} = W_K x_{ab} \quad v_{ab} = W_V x_{ab}$$

Convolution **Self-Attention**

Ramachandran, Prajit, et al. "Stand-alone self-attention in vision models." *arXiv preprint arXiv:1906.05909* (2019).

❖ Stand-Alone Self-Attention - 2019



-1, -1	-1, 0	-1, 1	-1, 2
0, -1	0, 0	0, 1	0, 2
1, -1	1, 0	1, 1	1, 2
2, -1	2, 0	2, 1	2, 2

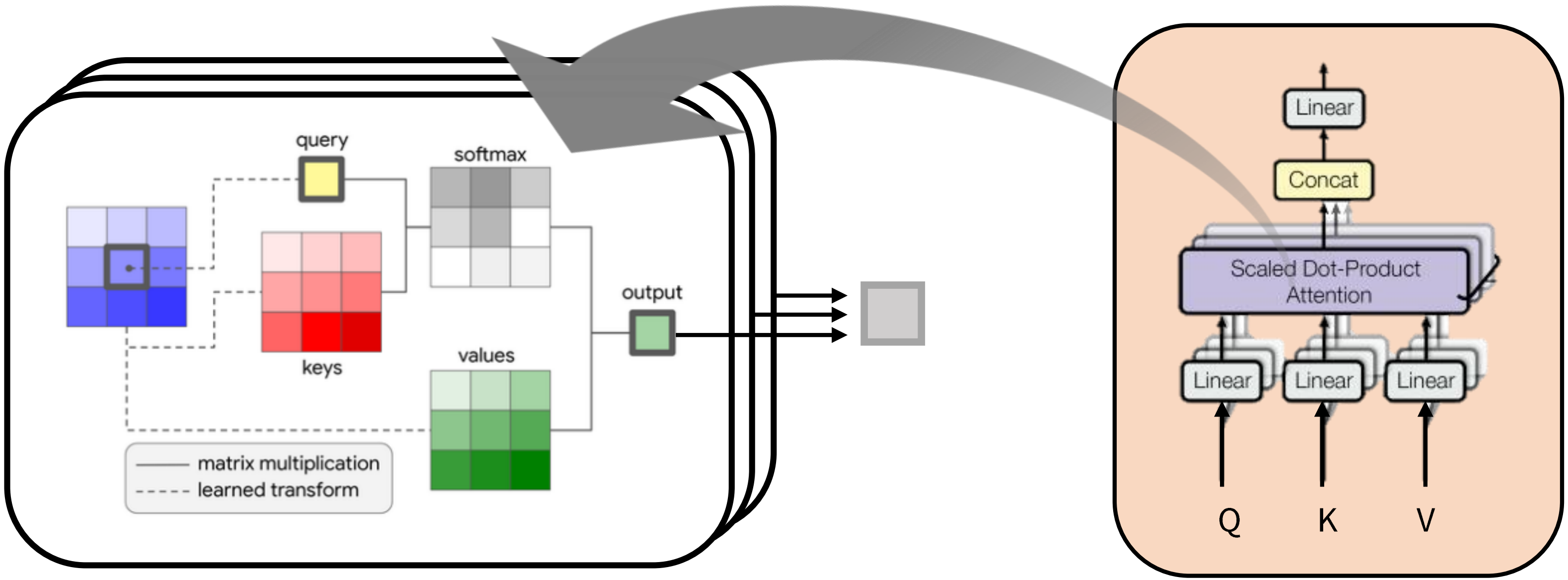
$$y_{ij} = \sum_{a,b \in N(i,j)} softmax_{ab}(q_{ij}^T k_{ab}) v_{ab}$$

$$y_{ij} = \sum_{a,b \in N(i,j)} softmax_{ab}(q_{ij}^T k_{ab} + q_{ij}^T r_{a-i,b-j}) v_{ab}$$

Ramachandran, Prajit, et al. "Stand-alone self-attention in vision models." *arXiv preprint arXiv:1906.05909* (2019).

Data Mining
Quality Analytics

❖ Stand-Alone Self-Attention - 2019



Multi-Head Attention

Ramachandran, Prajit, et al. "Stand-alone self-attention in vision models." *arXiv preprint arXiv:1906.05909* (2019).

Data Mining
Quality Analytics

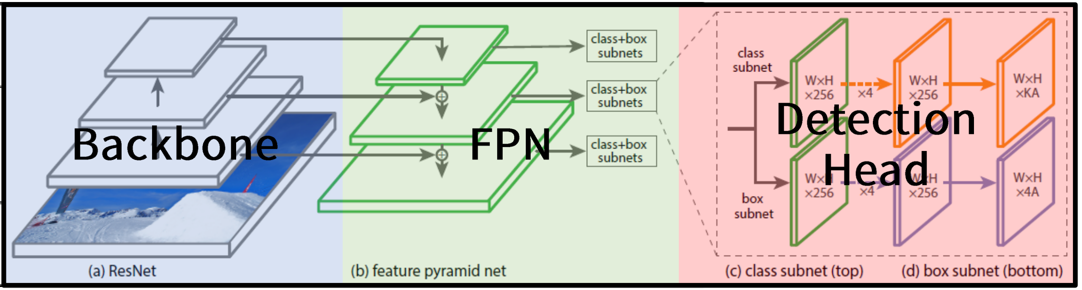
❖ Stand-Alone Self-Attention - 2019

- Image classification – ImageNet dataset

	ResNet-26			ResNet-38			ResNet-50		
	FLOPS (B)	Params (M)	Acc. (%)	FLOPS (B)	Params (M)	Acc. (%)	FLOPS (B)	Params (M)	Acc. (%)
Baseline	4.7	13.7	74.5	6.5	19.6	76.2	8.2	25.6	76.9
Conv-stem + Attention	4.5	10.3	**75.8**	5.7	14.1	**77.1**	7.0	18.0	**77.4**
Full Attention	4.7	10.3	74.8	6.0	14.1	76.9	7.2	18.0	**77.6**

- Object detection – COCO dataset

Detection Heads + FPN	Backbone
Convolution	Baseline Conv-stem + Attention Full Attention
Attention	Conv-stem + Attention Full Attention



Ramachandran, Prajit, et al. "Stand-alone self-attention in vision models." *arXiv preprint arXiv:1906.05909* (2019).

❖ Conclusion

- Attention은 '모델이 더 나은 성능을 위해 집중적으로 학습해야 할 부분까지도 학습하도록 하자.' 라는 관점에서 딥러닝 모델링에 필수적인 방법론
- 해석력 + 모델 성능 향상의 두가지 이점
- Visual attention은 computer vision 분야 중 가장 활발하게 연구되고 있는 분야 중 하나
- 목적은 동일하지만 다양한 형태의 attention 방법론이 존재함
- 최근 CNN 모델의 block 중 하나로 활용되는 것이 아니라 convolution 연산까지 대체할 수 있다는 가능성 확인
- 텍스트, 이미지 뿐만 아니라 시그널 등 다양한 데이터의 모델링에도 효과적일 것으로 생각되어 연구실에서 진행하고 있는 다양한 영역의 연구 또는 프로젝트에서 활용하였으면 좋겠음.

❖ Reference

▪ Attention

- Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." *arXiv preprint arXiv:1409.0473* (2014).
- Luong, Minh-Thang, Hieu Pham, and Christopher D. Manning. "Effective approaches to attention-based neural machine translation." *arXiv preprint arXiv:1508.04025* (2015).
- Xu, Kelvin, et al. "Show, attend and tell: Neural image caption generation with visual attention." *International conference on machine learning.* 2015.

▪ Self-Attention

- Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems. 2017.
- Hu, Jie, Li Shen, and Gang Sun. "Squeeze-and-excitation networks." *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2018.
- Woo, Sanghyun, et al. "Cbam: Convolutional block attention module." *Proceedings of the European Conference on Computer Vision (ECCV).* 2018.
- Wang, Xiaolong, et al. "Non-local neural networks." *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2018.
- Ramachandran, Prajit, et al. "Stand-alone self-attention in vision models." *arXiv preprint arXiv:1906.05909* (2019).

감사합니다