CS231n Quiz Review

24.11.07
유하영

Quiz 1. Attention & Self Attention

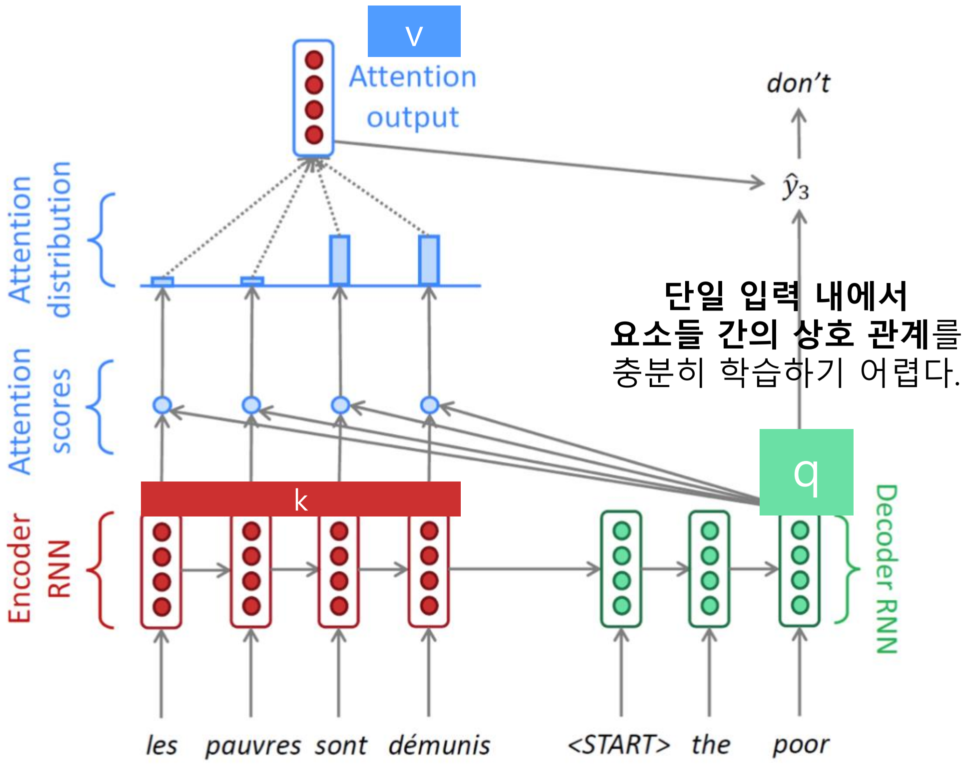
11. Attention operation is permutation invariant.

0/1 POINT

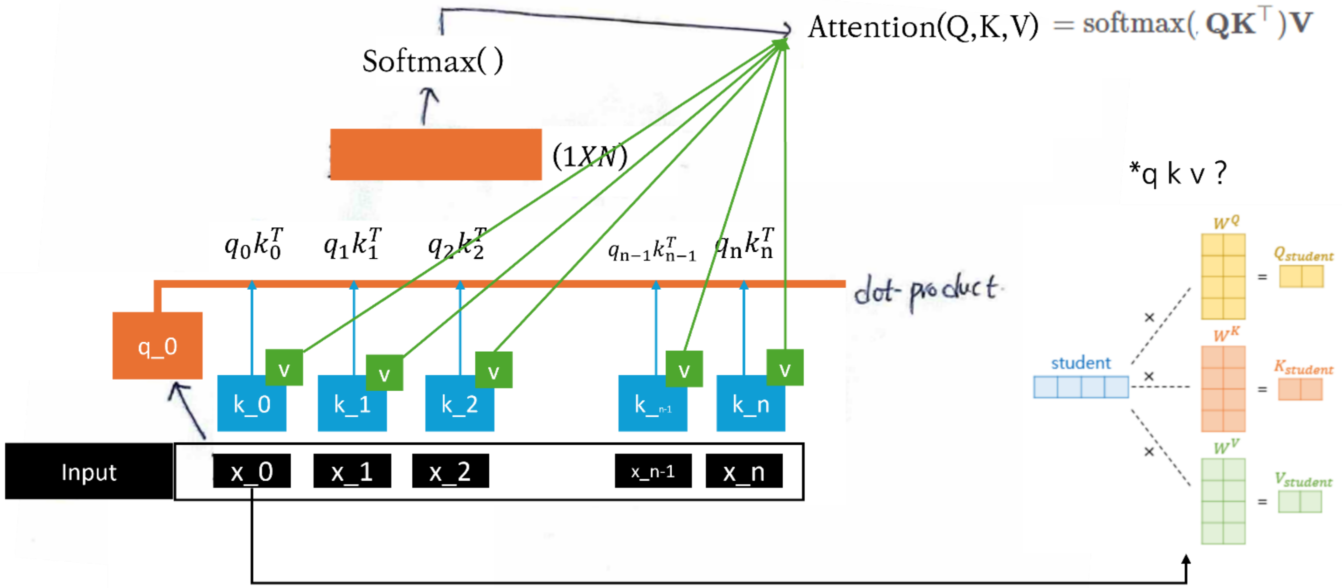
T True

F False

Attention
: 하나의 입력과 다른 입력과의 관계를 바탕으로 중요한 정보에 attention하는 메커니즘



Self Attention
: 단일 입력 내에서 모든 요소들 간의 관계를 학습



Quiz 2. ResNet

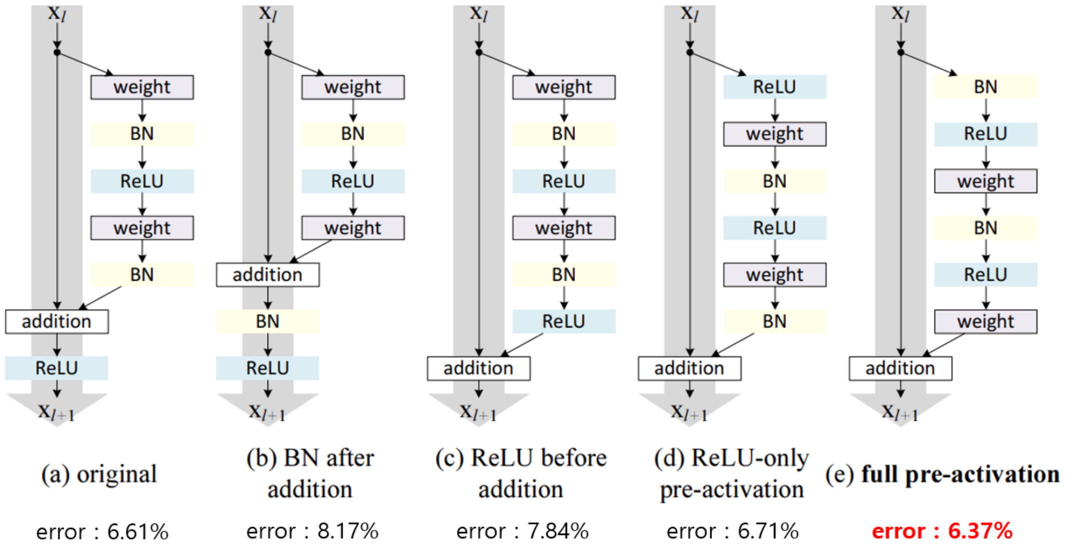
❌ **58.** Residual networks (ResNets) always help mitigate the degradation problem in deep neural networks by allowing training of substantially deeper models.

0/1 POINT

- T True
- F False

2. Pre-Activation

Batch Normalization과
ReLU 위치에 따른 성능 평가 지표



1. Residual Block

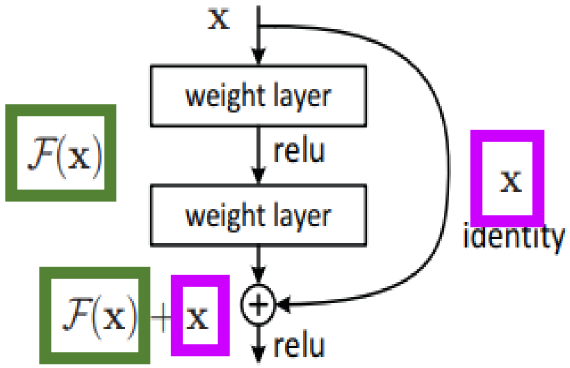
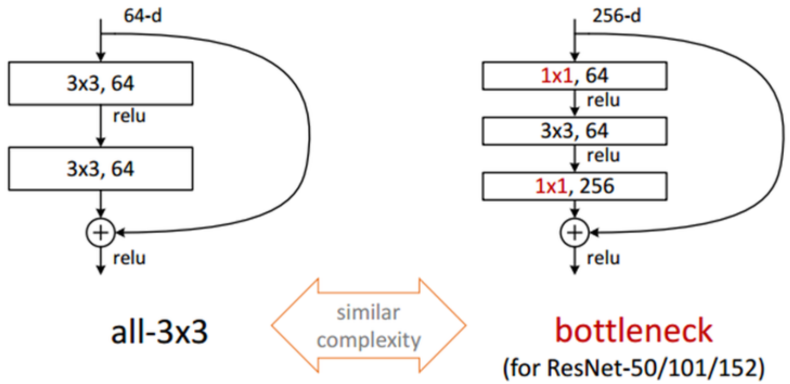


Figure 2. Residual learning: a building block.

$$F(x) + x$$

입력을 출력에 더해줌으로써 기울기가 직접 전달될 수 있게 하여
Gradient Vanishing 문제 완화

3. Bottleneck



dimension을 줄이기 위한 목적.
-> 파라미터 감소

ResNet-50 이상에서 사용

Quiz 3. Batch Normalization

7. Batch normalization can have an implicit regularizing effect, especially with smaller minibatches.

0/1 POINT

- T True
- F False

간단 논세

Layer Normalization

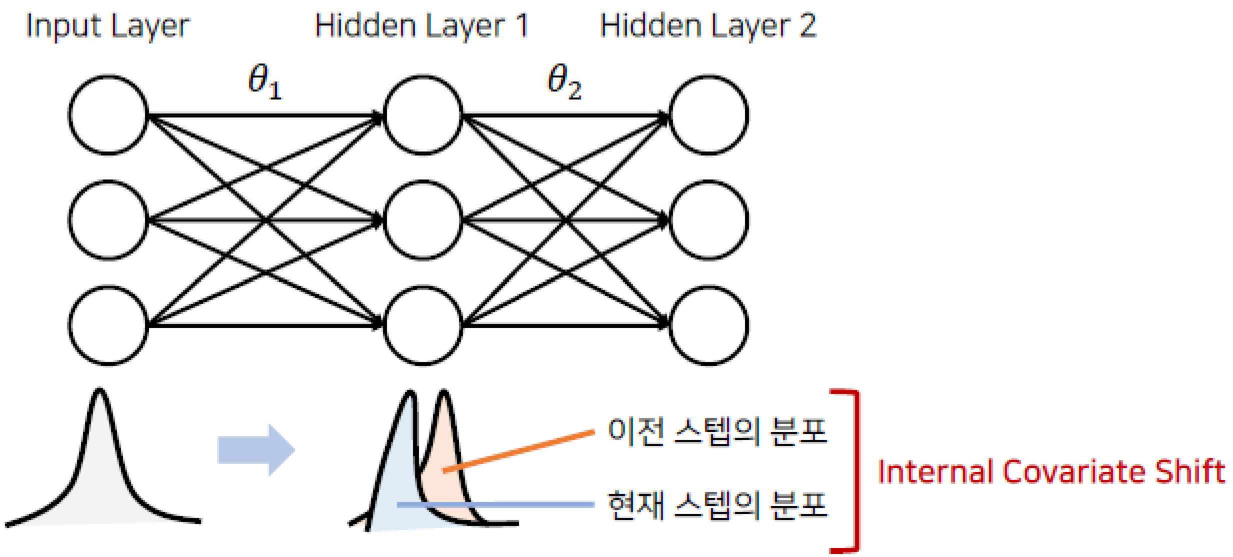
Lei Ba, Jimmy, Jamie Ryan Kiros, and Geoffrey E. Hinton. "Layer normalization." *ArXiv e-prints* (2016): arXiv-1607.

24.11.07
유하영

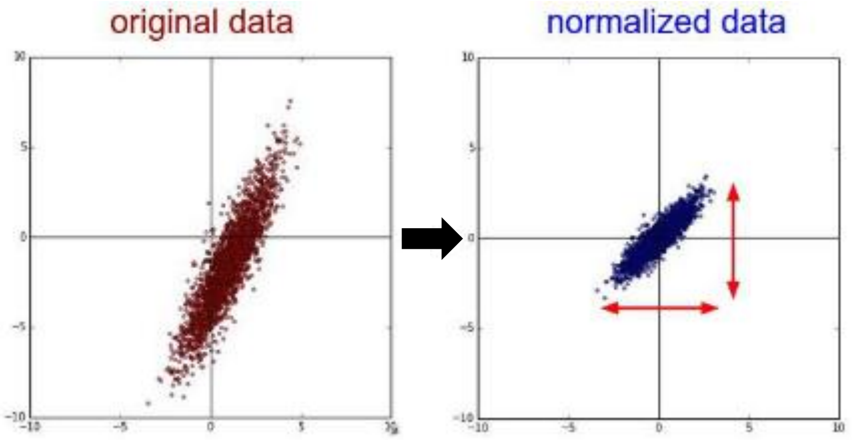
Introduction

긴 학습 시간 소요

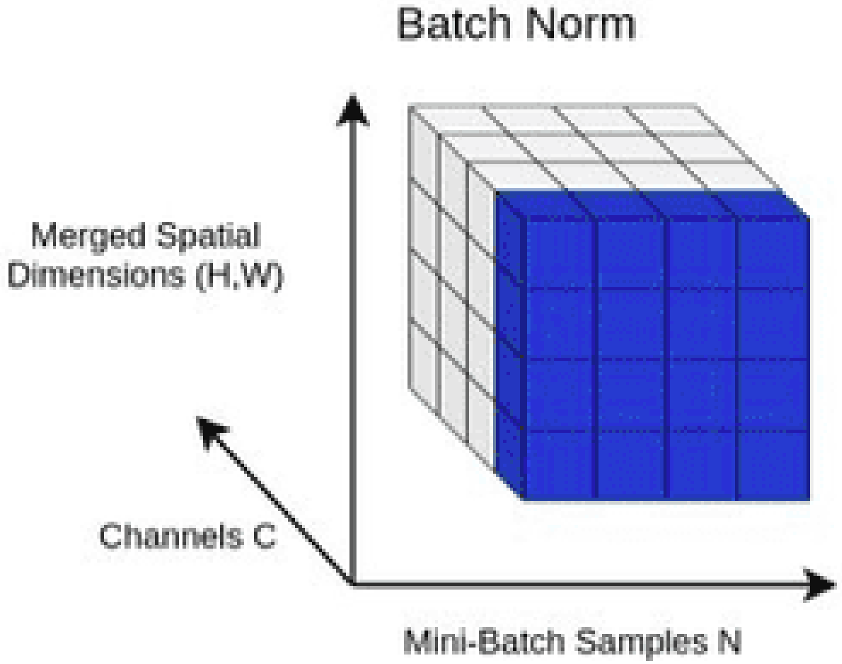
Covariate Shift



Normalization



Batch Normalization



Input: Values of x over a mini-batch: $\mathcal{B} = \{x_{1...m}\}$;
Parameters to be learned: γ, β

Output: $\{y_i = \mathrm{BN}_{\gamma,\beta}(x_i)\}$

$$\mu_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \text{// mini-batch mean}$$

$$\sigma_{\mathcal{B}}^2 \leftarrow \frac{1}{m}\sum_{i=1}^{m} (x_i - \mu_{\mathcal{B}})^2 \qquad \text{// mini-batch variance}$$

$$\widehat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}} \qquad \text{// normalize}$$

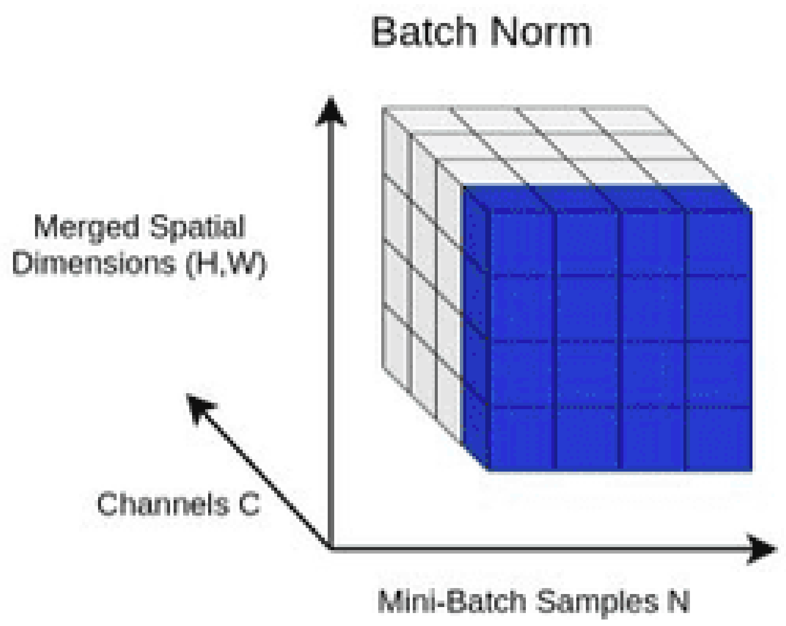
$$y_i \leftarrow \gamma \widehat{x}_i + \beta \equiv \mathrm{BN}_{\gamma,\beta}(x_i) \qquad \text{// scale and shift}$$

Algorithm 1: Batch Normalizing Transform, applied to activation x over a mini-batch.

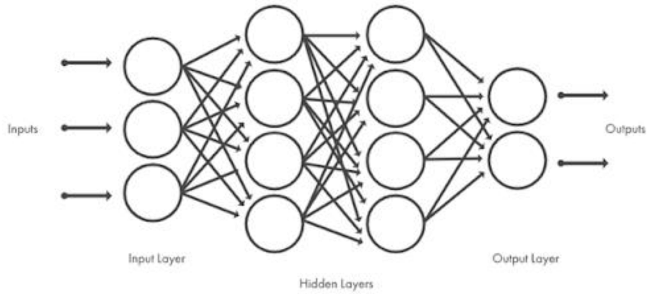
1. Sequence Length Variability in Recurrent Neural Networks (RNNs)
2. Limitations in Online Learning and Large Distributed Models

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. ICML, 2015.

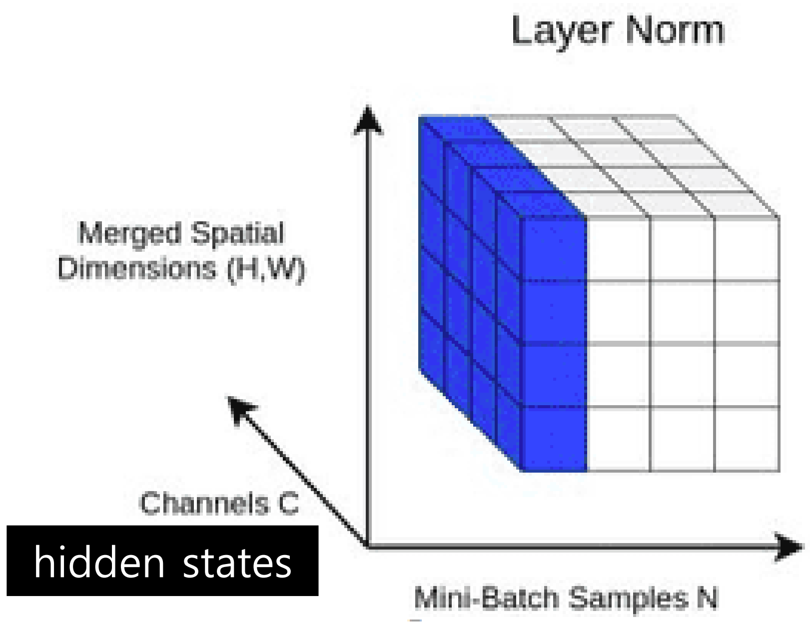
Layer Normalization



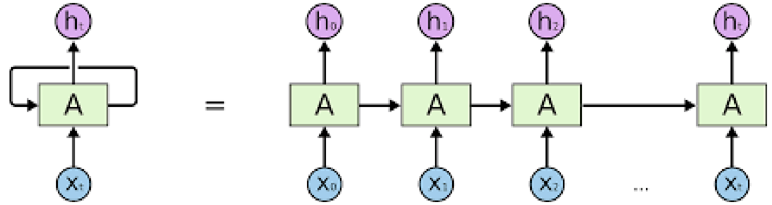
$$\bar{a}_i^l = \frac{g_i^l}{\sigma_i^l}\left(a_i^l - \mu_i^l\right)$$



Feed-forward neural network



$$\mathbf{h}^t = f\left[\frac{\mathbf{g}}{\sigma^t} \odot \left(\mathbf{a}^t - \mu^t\right) + \mathbf{b}\right]$$



RNN

Experimental results

6.2 Teaching machines to read and comprehend

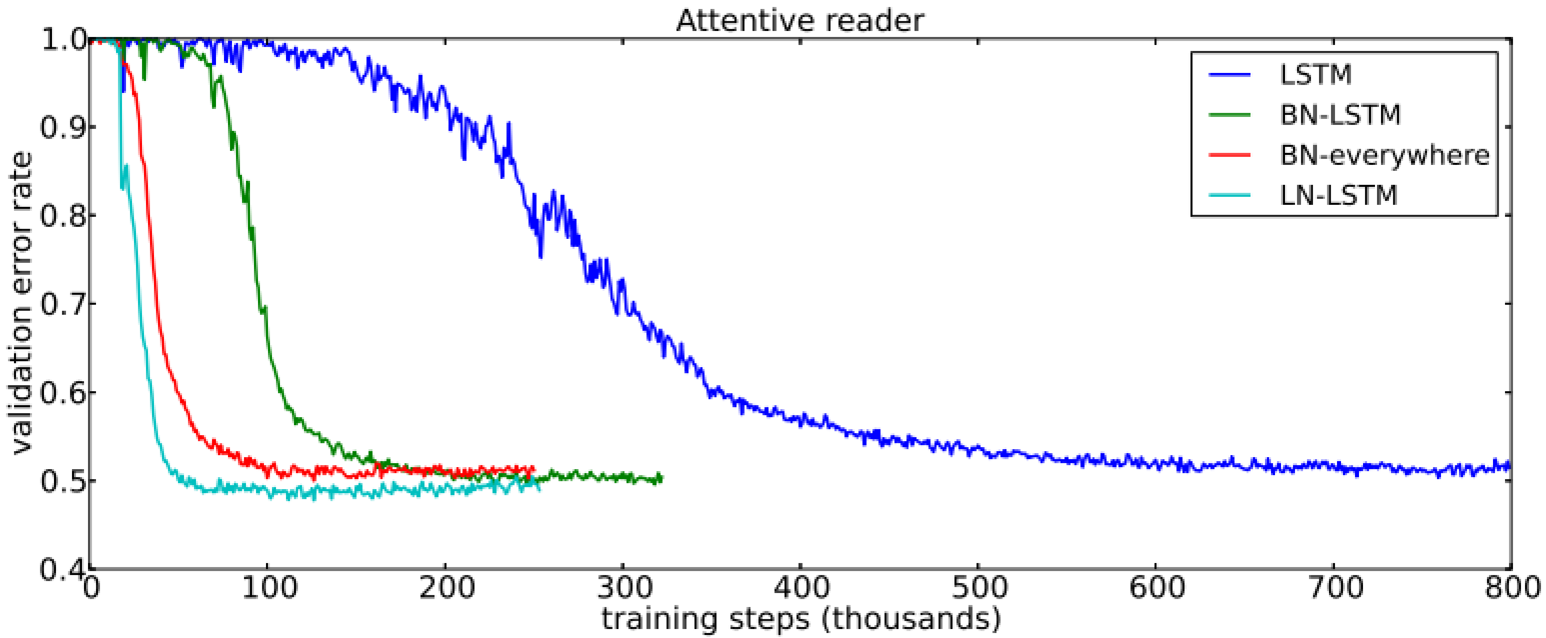


Figure 2: Validation curves for the attentive reader model. BN results are taken from [Cooijmans et al., 2016].

6.7 Convolutional Networks

Quiz 3. Normalization

7. Batch normalization can have an implicit regularizing effect, especially with smaller minibatches.

0/1 POINT

T True

F False

"regularizing effect"

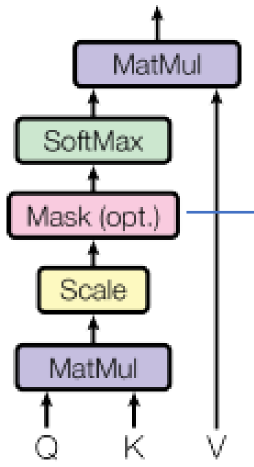
- batch마다 달라지는 통곗값
- shift 파라미터로 인한 정규화 값의 위치 조정

Scaled Dot-Product Attention

* Scaled Dot-Product Attention

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

Scaled Dot-Product Attention



<pad> 토큰의 학습 방지

Query, Key 벡터의 차원을 d_k,

Q, K의 요소들이 평균 0, 분산 1인 독립적인 랜덤변수라고 가정

$$Q \cdot K = \sum_{i=1}^{d_k} q_i k_i$$

* 기대값

$$E[q_i k_i] = E[q_i]\,E[k_i] = 0$$

$$E[Q \cdot K] = \sum_{i=1}^{d_k} E[q_i k_i] = 0$$

* 분산

$$Var(q_i k_i) = Var(q_i) \cdot Var(k_i) = 1$$

$$Var(Q \cdot K) = Var(\sum_{i=1}^{d_k} q_i k_i) = \sum_{i=1}^{d_k} Var(q_i k_i) = d_k$$

분산 $= d_k$ → 차원 d_k가 커질수록 분산이 비례하여 증가함

→ 내적값의 분산을 1로 맞추기 위해 $\sqrt{d_k}$로 scaling

$$\frac{Q \cdot K}{\sqrt{d_k}} = \frac{\sum_{i=1}^{d_k} q_i k_i}{\sqrt{d_k}}$$

분산의 성질에 의하여

$$Var(\frac{Q \cdot K}{\sqrt{d_k}}) = \frac{Var(Q \cdot K)}{d_k} = \frac{d_k}{d_k} = 1$$

Scaling을 수행하면 내적 결과의 분산은 1이 됨

차원이 증가하더라도 내적값의 크기가 적절한 범위로 유지되므로,
Attention weight가 특정 위치로 치우치는 현상을 방지할 수 있다.