self-attention

강사 : 백병인

pi.paek@modulabs.co.kr

모두의연구소 Research Scientist

Transformer

Were we so different?
They have much to learn.
But I've seen goodness in them.. - Optimus Prime

http://all4desktop.com/4181266-transformers-3-optimus-prime.html

Transformer : Attention-only model

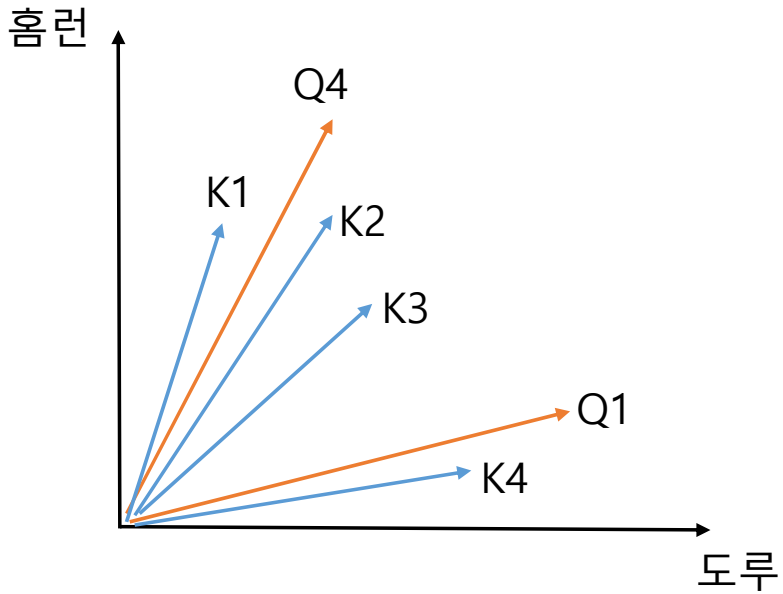
- Vaswani et al. [2017] Attention is All you need

- Can we **remove recurrences, convolutions** from sequence models?
- Sequential computations hinder parallelization.

- With Transformer, Yes!!
- The first simple sequence transduction model architecture, **based solely on attention** mechanism.
- Transformer established state-of-the-art scores, after much shorter training time.

포지션 부여 게임 (1)

- 당신은 프로야구팀 감독이다.
  - 9명의 타자들을 데리고 1번~9번까지의 타순 Qi을 짜야 한다.
  - Q1=1번타자 -> 누구로 하지?
  - Q4=4번타자 -> 누구로 하지?
- 9명의 타자 X = {K1, K2, …, K9}
- 포지션별 적합도 : Qi와 K_j의 cos similarity
- 포지션별 적합도를 어텐션으로 삼아 Xj를 재해석하자. Ci = Attn(Kj, Qi)

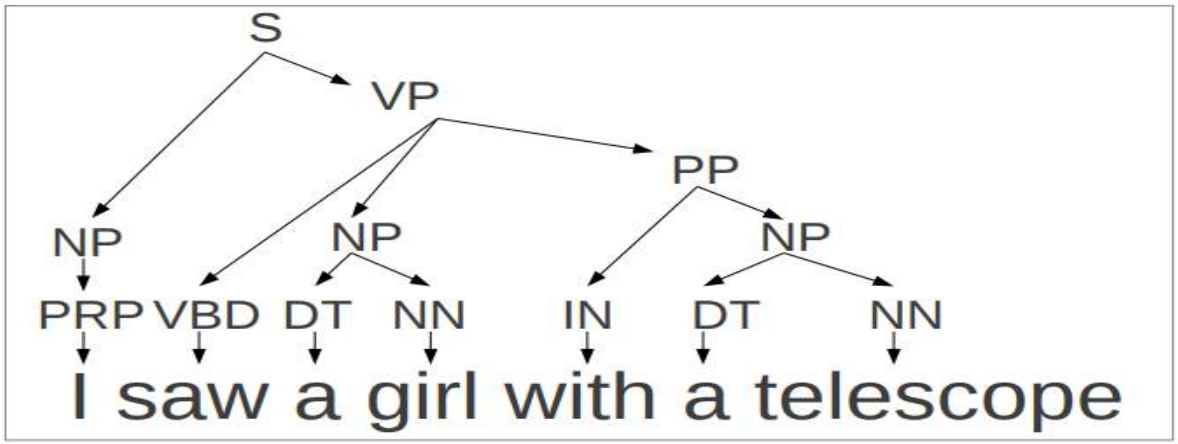


포지션 부여 게임 (2)

- 누구를 4번타자로 하느냐의 문제는 단순히 누가 4번타자감(Q4)에 가장 근접하냐만으로 결정되지 않는다. 특정 선수를 4번타자로 지목하는 것은 4번타자 포지션과 전체 선수들이 얼마나 잘 어울리는지 따져서 결정되는 상대적인 결정이다.
- 최적의 Qi 벡터값은 정해져 있는 것이 아니다. 어떤 상대팀을 만나느냐, 그 상대팀을 이기기 위한 우리 팀의 전략은 어떠해야 하느냐에 따라 Qi는 수정될 수 있다.
- 즉, Q-K의 alignment는 Qi, Kj들끼리의 자체적인 관계 뿐 아니라 Q-K간 관계 및 Q-K의 alignment가 이루고 있는 더 상위 hierarchy 구조에까지 영향을 받아 성립된다.

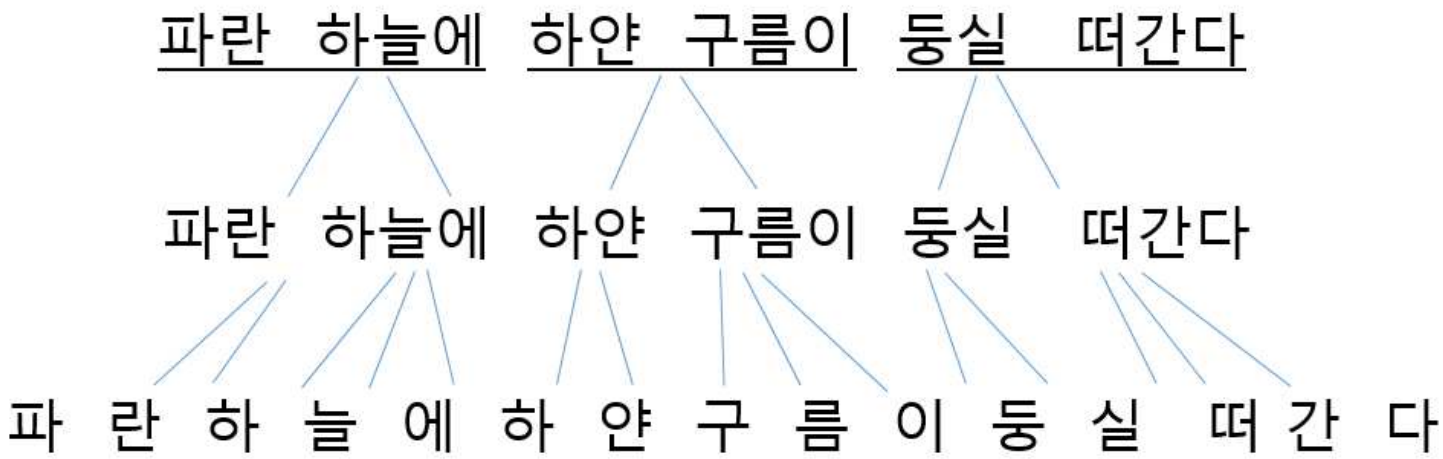
Structured Attention Network?

- 포지셔닝 게임에서 모든 포지션이 동일한 위상을 가지는 것은 아니다.
- 텍스트를 다루는 문제를 생각해 보라. 문장 내 각 포지션(Q) 간에 상호 의존적 내부 구조(parent-children)를 가지는 경우는 어떻게 할 것인가?
- 이 구조를 보다 직접적으로 모델링할 수는 없을까?



Self-attention 개념의 출현

- Kim et al.[2017] Strunctured Attention Networks
- Self-attention : structured attention over the source only to obtain **soft-parents** for each symbol



Structured Attention

- 기존 Attention에 대한 해석

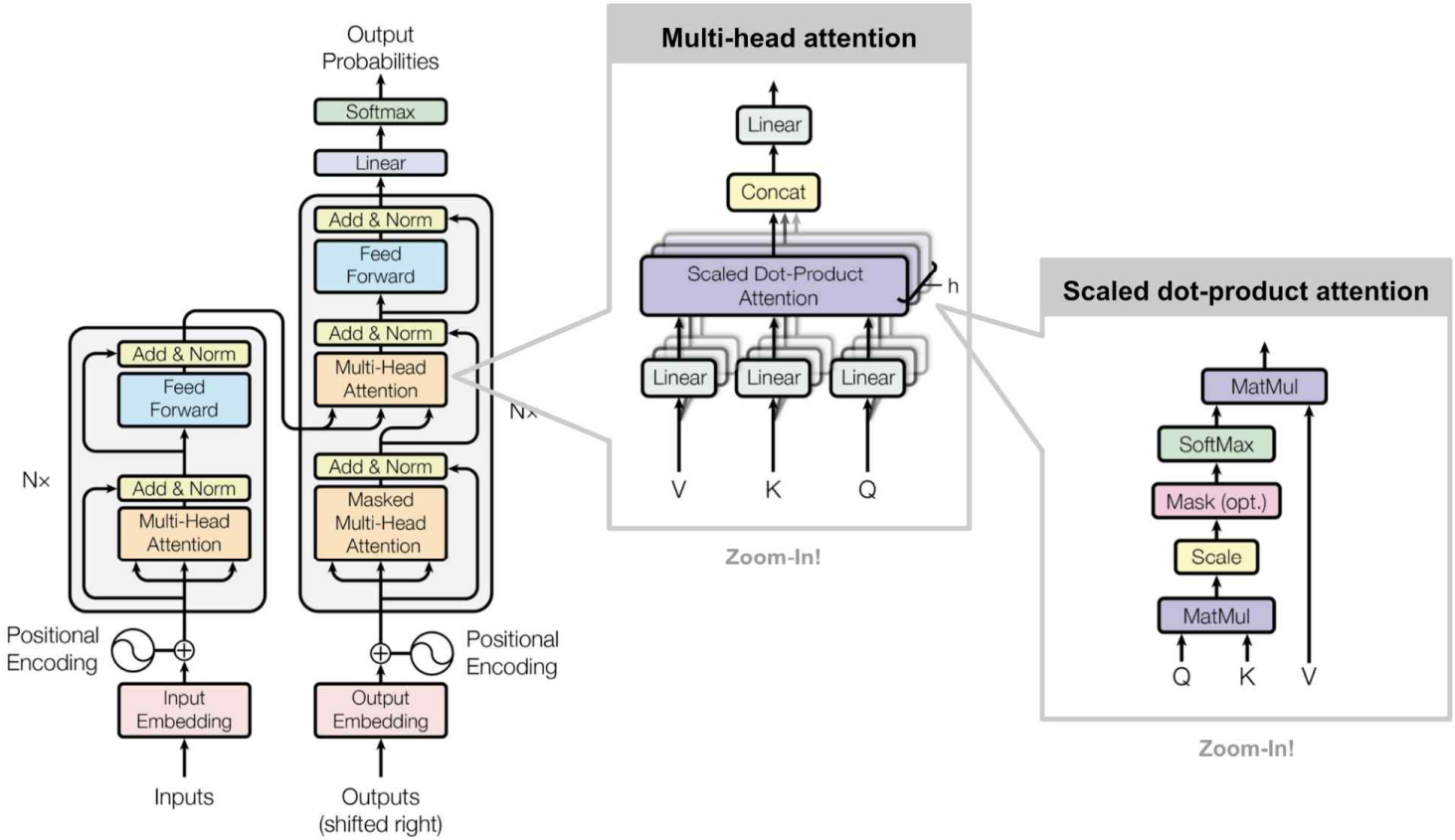
$$\mathbf{c} = \mathbb{E}_{z \sim p(z \mid x,q)}[f(x,z)] = \sum_{i=1}^{n} p(z=i \mid x,q)\mathbf{x}_i$$

- Q-K의 relation인 Z를 discrete한 latent variable로 생각한다면, atttention p(z|x, q)는 CRF(conditional random field)로 볼 수 있다.

$$c = \mathbb{E}_{z \sim p(z \mid x,q)}[f(x,z)] = \sum_{C} \mathbb{E}_{z \sim p(z_C \mid x,q)}[f_C(x, z_C)]$$

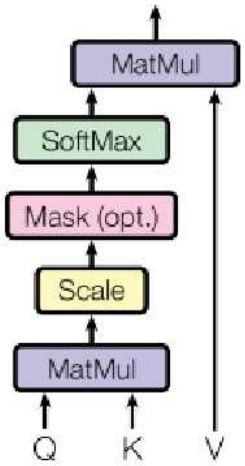
- 이제 layer 구조로 attention을 쌓아 보자.

Transformer Full Model Architecture

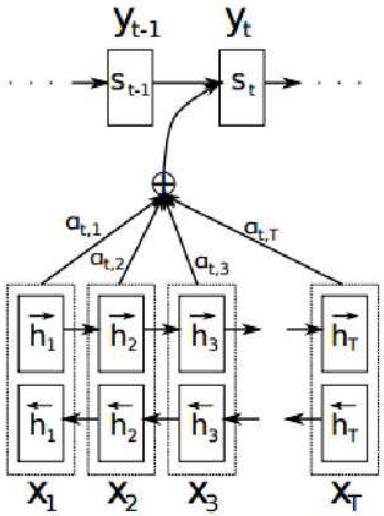


https://lilianweng.github.io/lil-log/2018/06/24/attention-attention.html

Scaled Dot-product Attention



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



$$p(y_i|y_1, \ldots, y_{i-1}, \mathbf{x}) = g(y_{i-1}, s_i, c_i)$$

$$s_i = f(s_{i-1}, y_{i-1}, c_i)$$

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

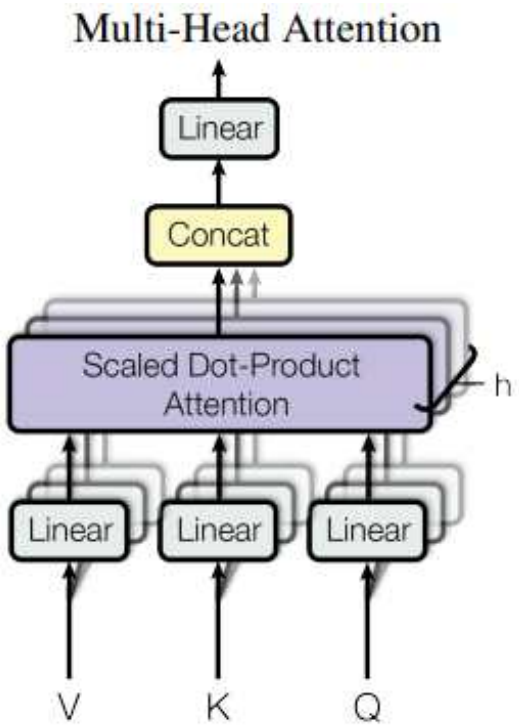
$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})}$$

$$e_{ij} = a(s_{i-1}, h_j)$$

Q K

attention은 V를 해석하는 encoder
Q=K 가 된다는 것은, 포지션 정보를 해석해야 할 X=V 자기 자신의 하위 구조에서 찾아야 한다는 뜻이다.
e.g) 하늘 = 0.5*하 + 0.5*늘 + 0*(나머지 하위 다)

Multi-Head Attention (1)



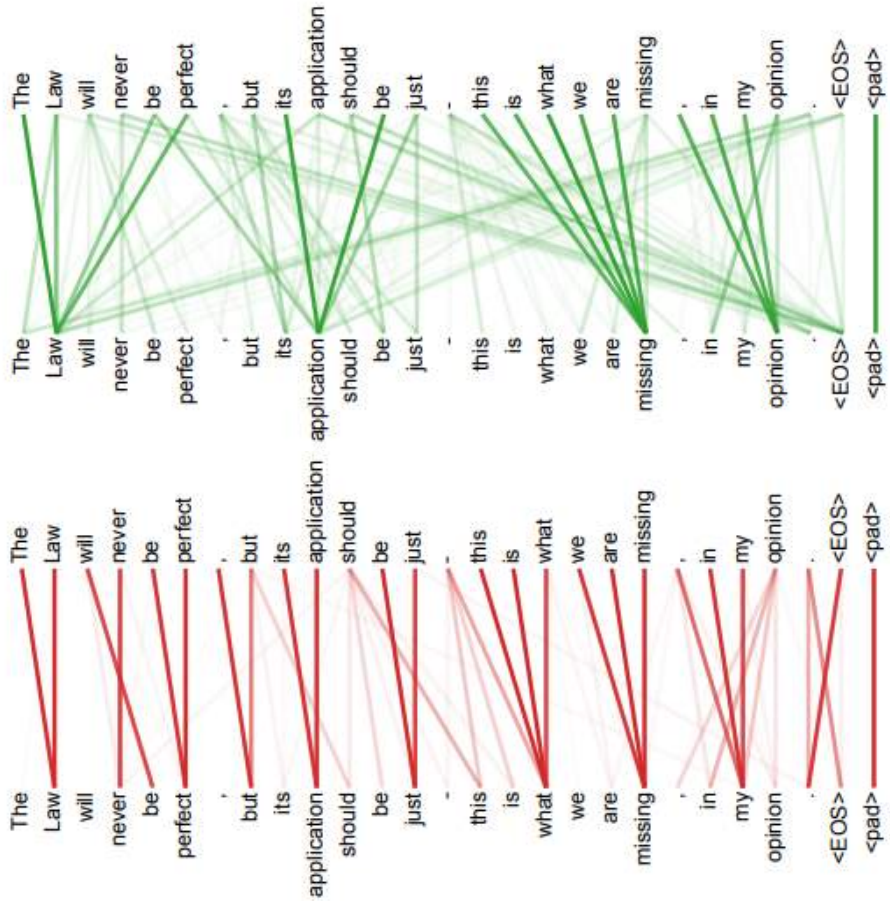
- Why Multi-Head??
- Q-K의 relation을 한가지 관점으로만 규정할 수 있을까?

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_\text{h})W^O$$
$$\text{where head}_\text{i} = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

Multi-Head Attention (2)

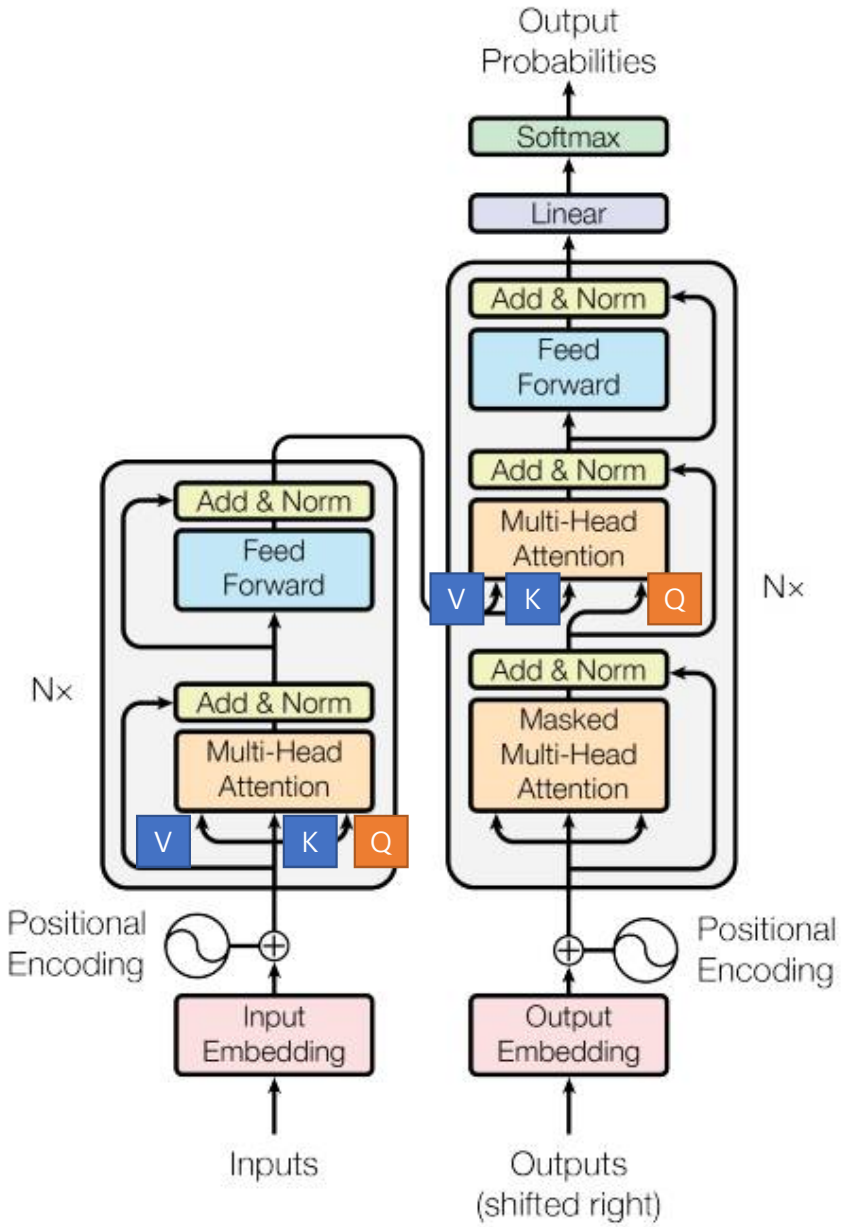
- Multi-Head Attention의 각 Head가 학습한 Q-K relation 은 실제로 다르게 나타난다.
- 이것은 동일한 self-attention 이 다른 task를 수행하도록 학습되었기 때문이다.

- Vaswani et al. [2017] Attention is All you need, p.15

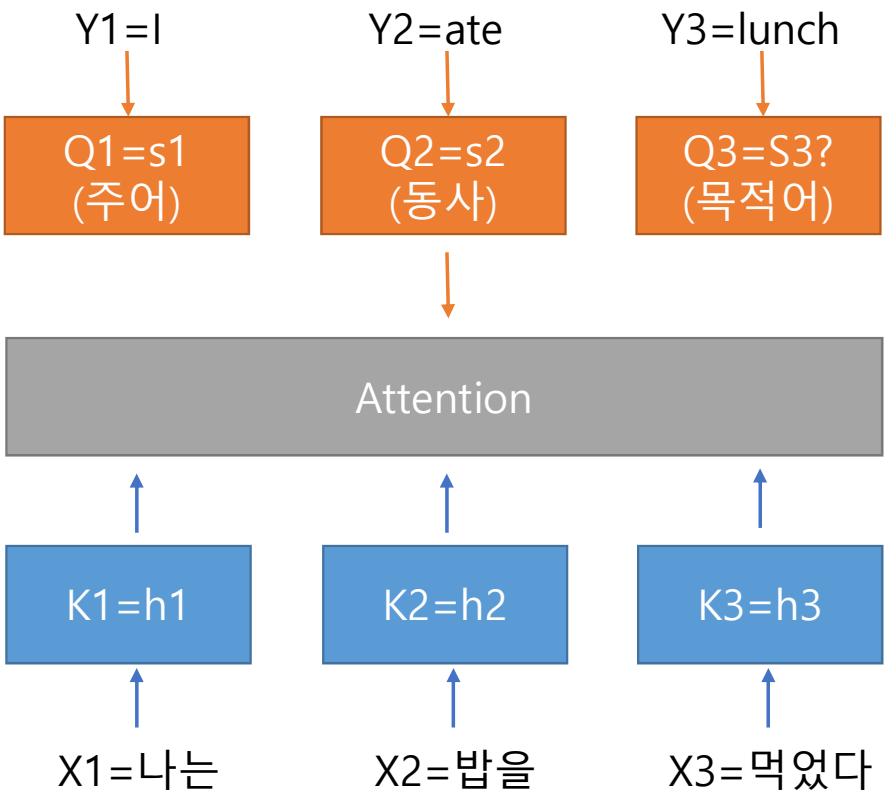


https://arxiv.org/pdf/1706.03762.pdf

Full Image



NMT with attention



C3 = 0.1*h1 + 0.75*h2 + 0.15*h3

=> 목적어 포지션 Q3를 결정할 때는
문장 X={X1,X2,X3}를 C3로 해석하세요.

No RNN -> positional encoding

- Input encoding을 RNN으로 하지 않다 보니 발생하는 문제
  => I ate lunch 와 Lunch ate I 가 구분이 안되는 문제.
  => 1-d CNN encoder에도 동일한 문제는 발생한다.
  => input word embedding에 positional encoding vector를 더해서 입력으로 처리한다.

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

A1. HMM, MEMM, CRF

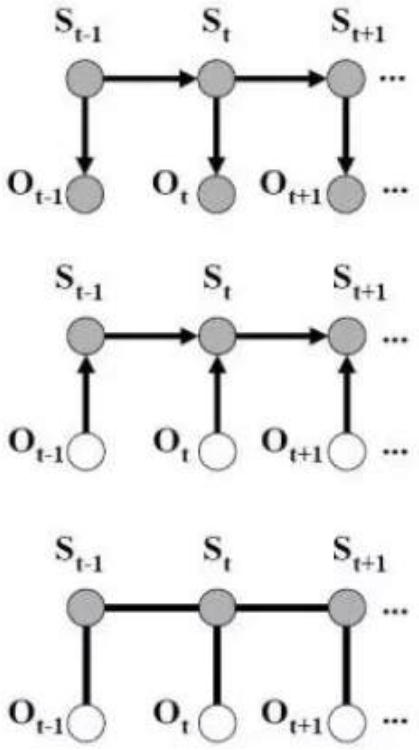
$\vec{s} = s_1, s_2, \ldots s_n \qquad \vec{o} = o_1, o_2, \ldots o_n$

HMM $P(\vec{s}, \vec{o}) \propto \prod_{t=1}^{|\vec{o}|} P(s_t \mid s_{t-1}) P(o_t \mid s_t)$

MEMM $P(\vec{s} \mid \vec{o}) \propto \prod_{t=1}^{|\vec{o}|} P(s_t \mid s_{t-1}, o_t)$

$$\propto \prod_{t=1}^{|\vec{o}|} \frac{1}{Z_{s_{t-1}, o_t}} \exp\left(\sum_j \lambda_j f_j(s_t, s_{t-1}) + \sum_k \mu_k g_k(s_t, x_t) \right)$$

CRF $P(\vec{s} \mid \vec{o}) \propto \frac{1}{Z_{\vec{o}}} \prod_{t=1}^{|\vec{o}|} \exp\left(\sum_j \lambda_j f_j(s_t, s_{t-1}) + \sum_k \mu_k g_k(s_t, x_t) \right)$



The advantage of CRF is that CRF resolve the label bias problem which can be happened in the MEMM model by global normalization.

https://liqiangguo.wordpress.com/2011/04/19/memm-v-s-crf/