Transformer,BERT/GPT series

ADVISOR: Jia-Ling Koh
PRESENTER: Xiao-Yuan Hung
DATE: 2022/08/23

1

Review

Paper
https://arxiv.org/pdf/2106.01702.pdf
Slides
http://140.122.184.128/presentation/22-07-12/PsyQA-%20A%20Chinese%20Dataset%20for%20Generating%20Long%20Counseling%20Text%20for%20Mental%20Health%20Support_slide.pdf

PsyQA:Question

- **Chinese dataset of Psychological health support in the form of Question-Answer pair**

Question (Post Title)

为什么有些事情越想心越闷?
The more I think about some things, the more upset I feel. Why?

Description (Post Content)

比如家里人说我和别人相亲的事，我是不太主动比较被动的人……又怪自己，又恨别人到处说，搞得心里很难受很郁闷，这该怎么办?
For example, my family asked me to go on a blind date with others. I am not an active but passive person. ... I blame myself and blame others for speaking ill of me everywhere, making me very uncomfortable and depressed. What should I do?

Keywords

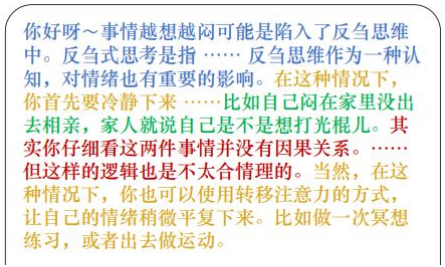
情绪 表达情绪 情绪调节 情绪智力
Emotion, Emotion Expression, Emotion Regulation, EQ

你好呀～事情越想越闷可能是陷入了反刍思维中。反刍式思考是指 …… 反刍思维作为一种认知，对情绪也有重要的影响。在这种情况下，你首先要冷静下来 ……比如自己闷在家里没出去相亲，家人就说自己是不是想打光棍儿。其实你仔细看这两件事情并没有因果关系。……但这样的逻辑也是不太合情理的。当然，在这种情况下，你也可以使用转移注意力的方式，让自己的情绪稍微平复下来。比如做一次冥想练习，或者出去做运动。

Hi ~ The more you think about it, the more depressed you feel. This is maybe because you are trapped in ruminant thinking. Ruminant thinking means that ..... Ruminant thinking, as a form of cognition, also has an important effect on emotion. In this case, you need to calm down first… For example, you stayed at home and didn't go out for a blind date, and your family said that you just wanted to be single. When you look at it carefully, there is no causal relationship between the two events. ..... But this logic doesn't work. Of course, in this case, you can also distract your attention to calm yourself down a bit. Take a meditation practice, or go outside to exercise.

Strategy Identification

- **Task**
  - Assign a strategy label to each sentence in a long answer
- **Implement method**
  - RoBERTa
- **Data Preparation**
  - Dataset
    - Annotated part of PsyQA
  - Randomly split them into train (80%), dev (10%) and test (10%) sets.



Task

- **Input**
  - question S_Q (question sentences)
  - description S_D (description sentences)
  - keyword set K (composed by at most 4 keywords)
- **Output**
  - a long counseling text consisting of multiple sentences that could give helpful comforts and advice mimicking a mental health counselor

Model Pretraining

- **Data**
  - 50K articles related to psychology and mental health support from Yixinli
- **Model**
  - GPT-2 (Generative Pre-Training)

GPT-2 generate

Question:

遇到困难，如何让自己的心志变得更加坚强？(How to be stronger when encountering difficulties?)

Description:

当面临一些困难的时候心志不够坚定的人会感到慌乱，那么如何在还没有遇见困难之前让自己变得更加坚强？When faced with some difficulties, people who are not determined enough will feel panic. How to make ourselves stronger before we meet some difficulties?

Keywords:

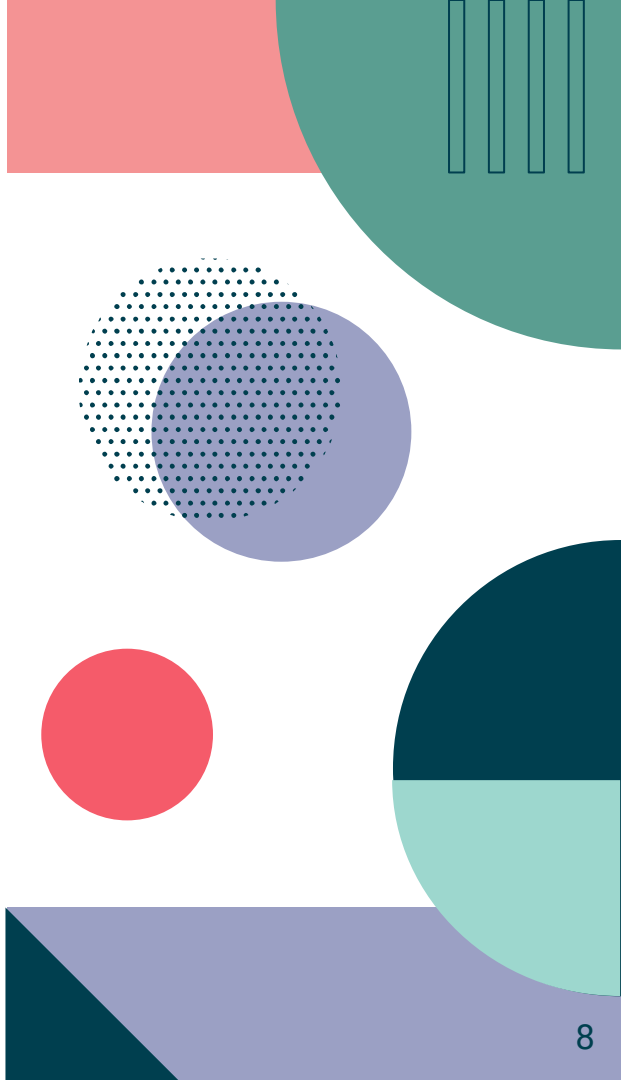
成长,性格完善,自我成长 (self-growth, personality improvement, self-development)

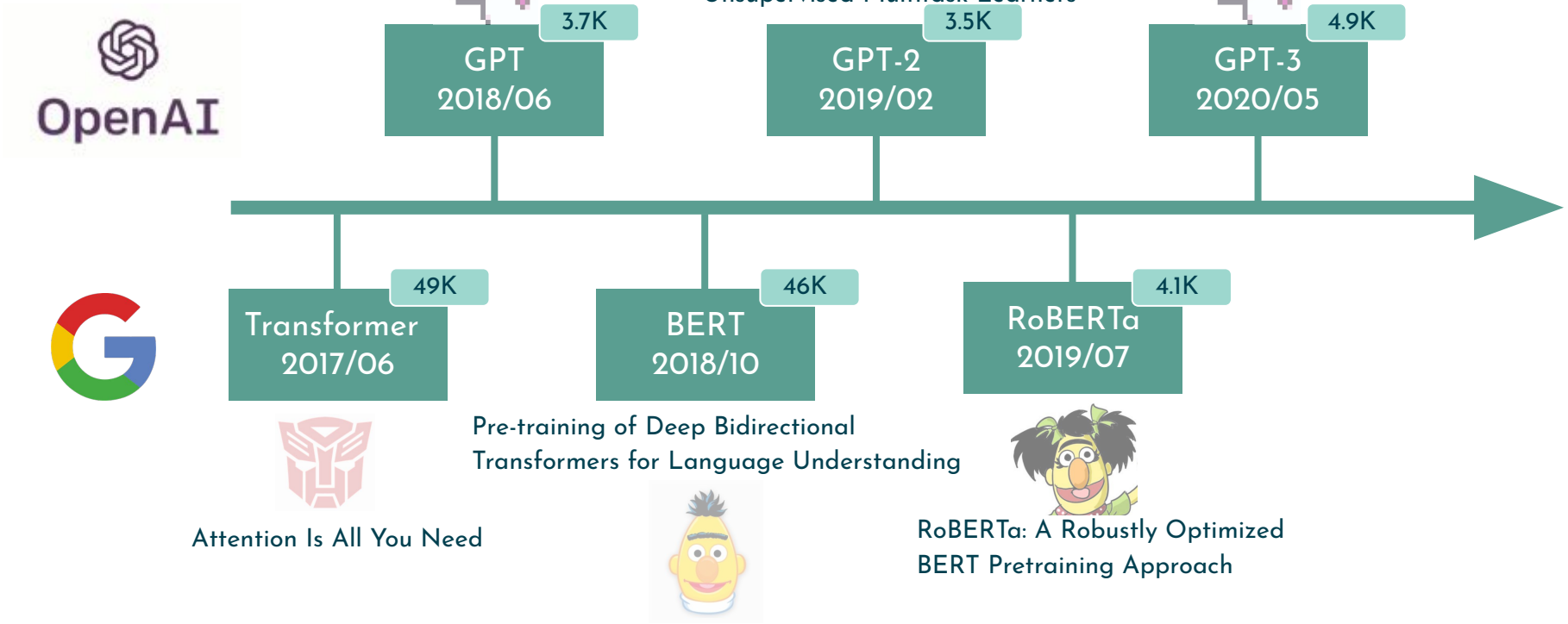
GPT_{ft}+Strategy

要坚持下去，只有这样才可能真正成为你坚持下去的动力。因为我们的目标很多是希望自己能够变得越来越优秀，能够在这一点上给自己带来很大的动力。比如我们在遇到困难的时候，也会努力想要去克服困难，但是很多时候我们会觉得要克服困难并不容易，因为我们并不完美，我们没有办法去付出相应的努力。如果你总是很害怕面对那些困难，那么就需要去克服困难。这一次是我的经验，希望对你有所帮助。当面临困难的时候，可以先先告诉自己这件事情是我能够克服的，我可以变得更好。但是我要告诉自己我可以去克服的。在面对困难的时候我们更需要去克服那些困难。但是我们要相信我们不会是不可控的。当自己在面对困难的时候，有能力去应对困难的时候，那么就能够更加的坚持下去。加油!

2

Models





https://www.youtube.com/watch?v=nzqIFIcCSWQ

OpenAI

Improving Language Understanding by Generative Pre-Training

GPT 2018/06

3.7K

Language Models are Unsupervised Multitask Learners

GPT-2 2019/02

3.5K

Language Models are Few-Shot Learners

GPT-3 2020/05

4.9K

Transformer 2017/06

49K

Attention Is All You Need

BERT 2018/10

46K

Pre-training of Deep Bidirectional Transformers for Language Understanding

RoBERTa 2019/07

4.1K

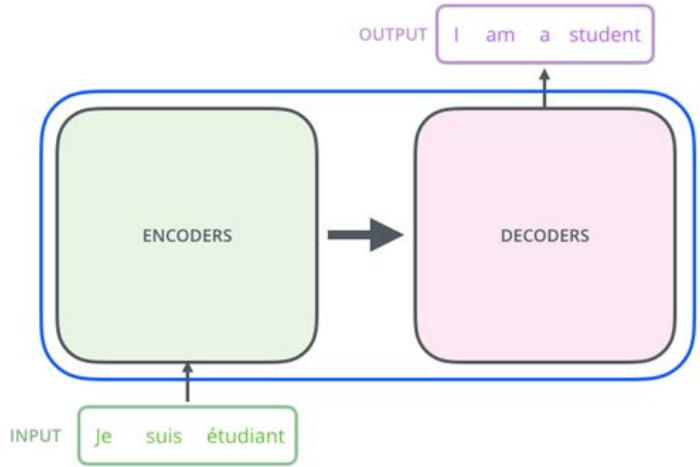
RoBERTa: A Robustly Optimized BERT Pretraining Approach

Source

- **Attention Is All You Need** (Times Cited 49340)
  - https://arxiv.org/pdf/1706.03762.pdf

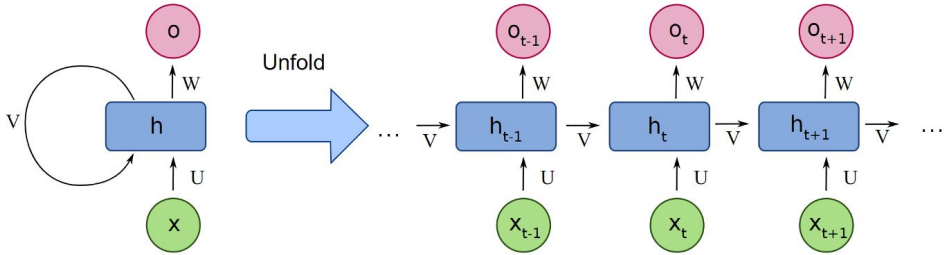
Transformer

- Propose a new seq2seq model with multi-headed self-attention
- Apply to machine translation and other tasks



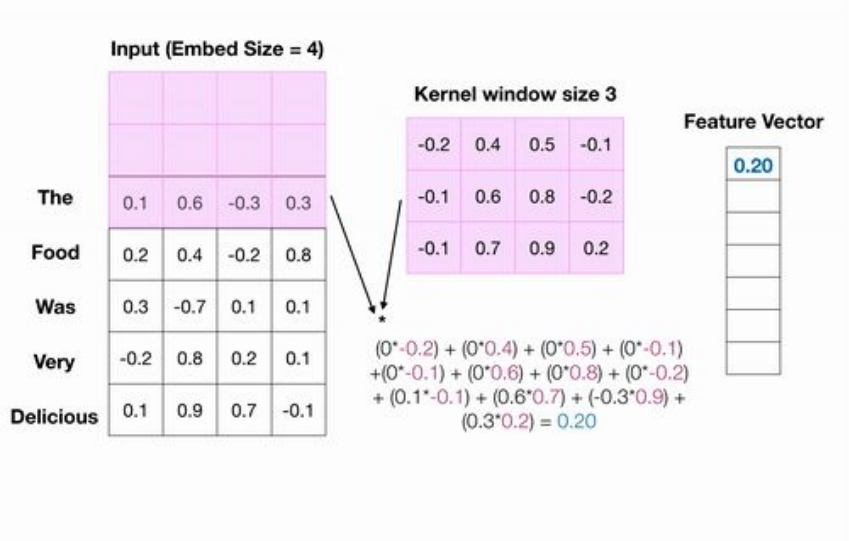
https://jalammar.github.io/illustrated-transformer/

RNN-based



- **Disadvantages**
  - hard to parallel
  - historical information loss

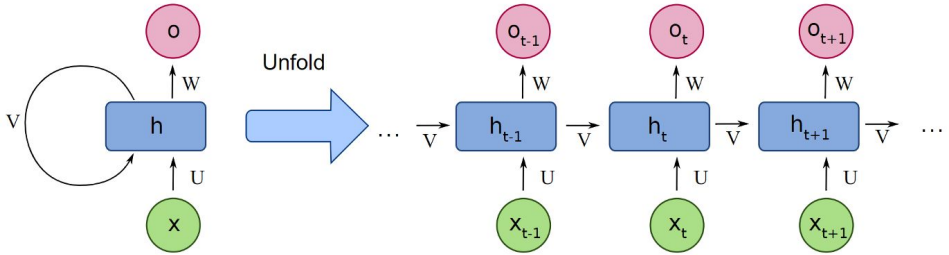
CNN-based



- **Disadvantages**
  - Requires many layers to read longer sentences

RNN https://medium.com/deeplearningbrasilia/deep-learning-recurrent-neural-networks-f9482a24d010
CNN https://chiranthancv95.medium.com/convolutional-sequence-to-sequence-learning-in-nlp-8ec25072de19

RNN-based



- **Disadvantages**
  - hard to parallel
  - historical information loss

RNN https://medium.com/deeplearningbrasilia/deep-learning-recurrent-neural-networks-f9482a24d010
CNN https://chiranthancv95.medium.com/convolutional-sequence-to-sequence-learning-in-nlp-8ec25072de19

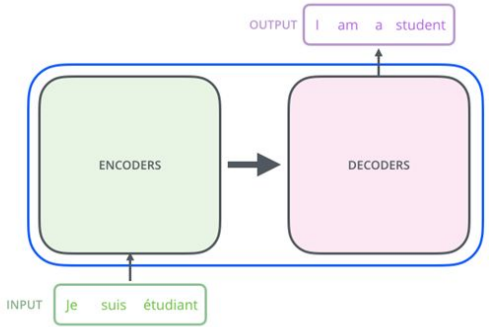
CNN-based

- **Disadvantages**
  - Requires many layers to read longer sentences

RNN https://medium.com/deeplearningbrasilia/deep-learning-recurrent-neural-networks-f9482a24d010
CNN https://chiranthancv95.medium.com/convolutional-sequence-to-sequence-learning-in-nlp-8ec25072de19

Transformer

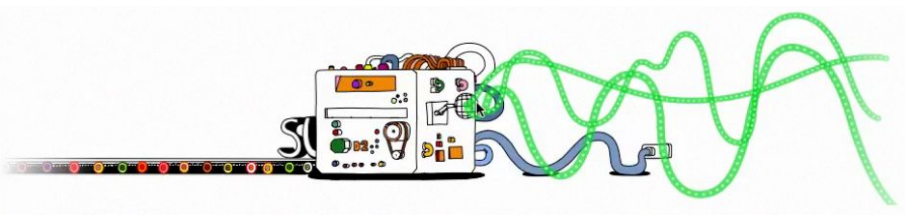
- Sequence transduction model based entirely on attention
- low complexity
- Parallelization
- High quality of translation after being trained for a little time



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

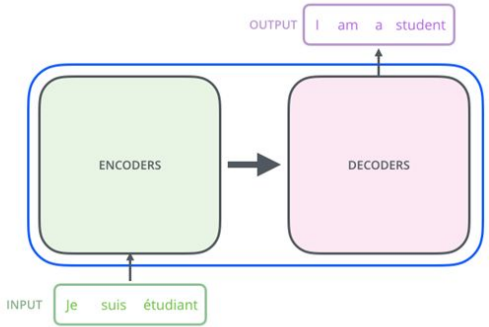
(additional Info.)Sequence Transduction

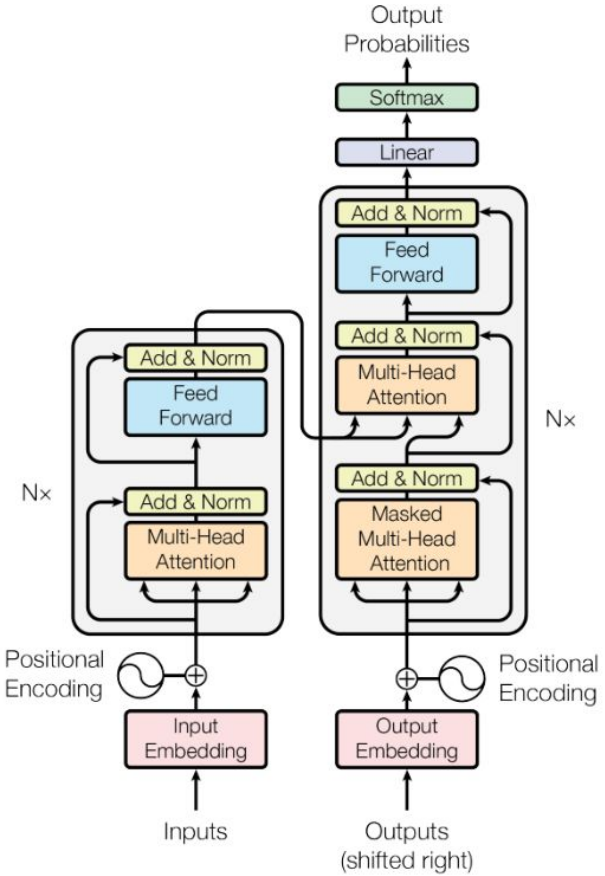
- **Any task where input sequences are transformed into output sequences**
- **Practical example**
  - speech recognition, text-to-speech, machine translation, protein secondary structure prediction…
- **Not so practical**
  - Turing machines, human intelligence…
- **Want a single framework able to handle as many kinds of sequence as possible**



Model Architecture

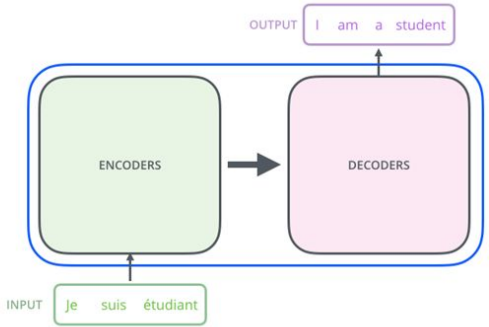
- **Encoder Input**
- **Decoder Output**

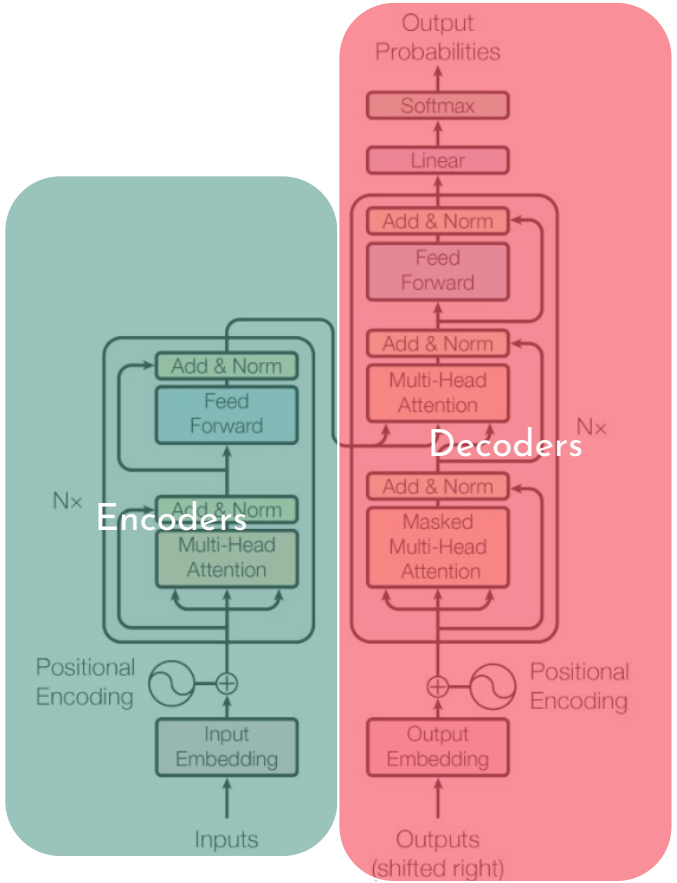




Model Architecture

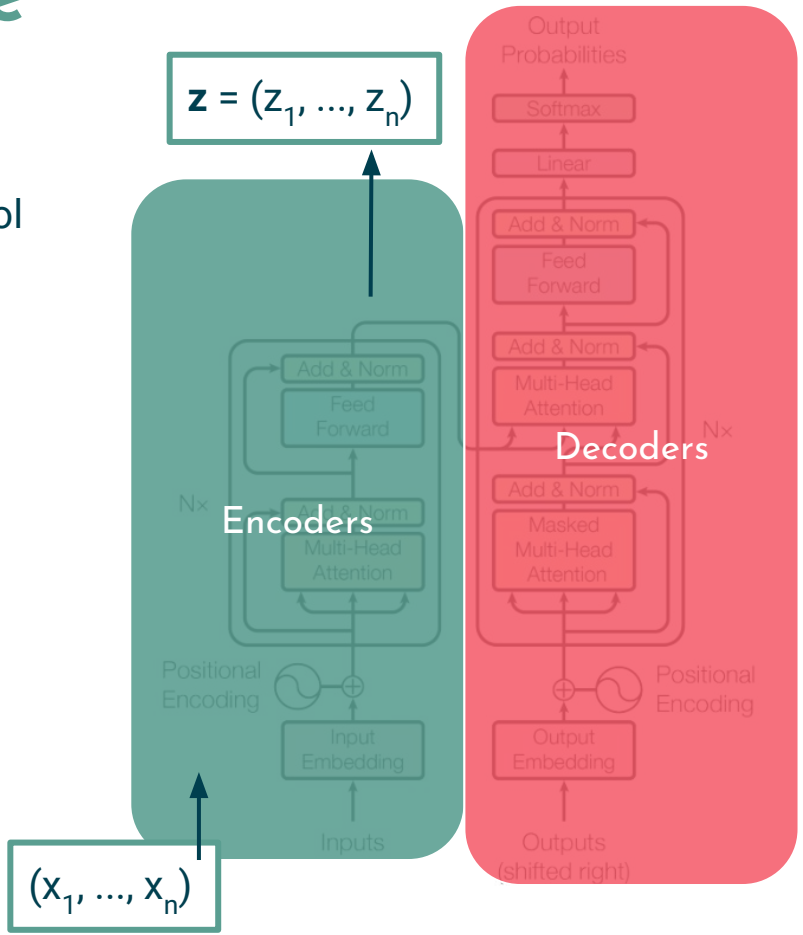
- **Encoder Input**
- **Decoder Output**





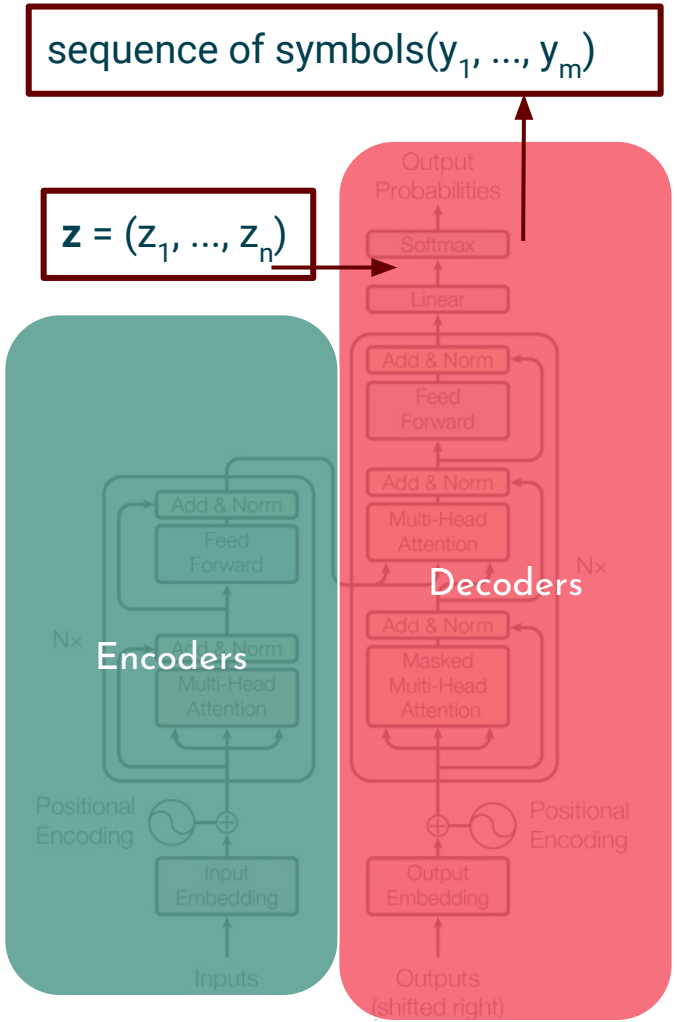
Model Architecture

- **Encoder Input**
  - maps an input sequence of symbol representations$(x_1, ..., x_n)$ to a sequence of continuous representations $\mathbf{z} = (z_1, ..., z_n)$
- **Decoder Output**



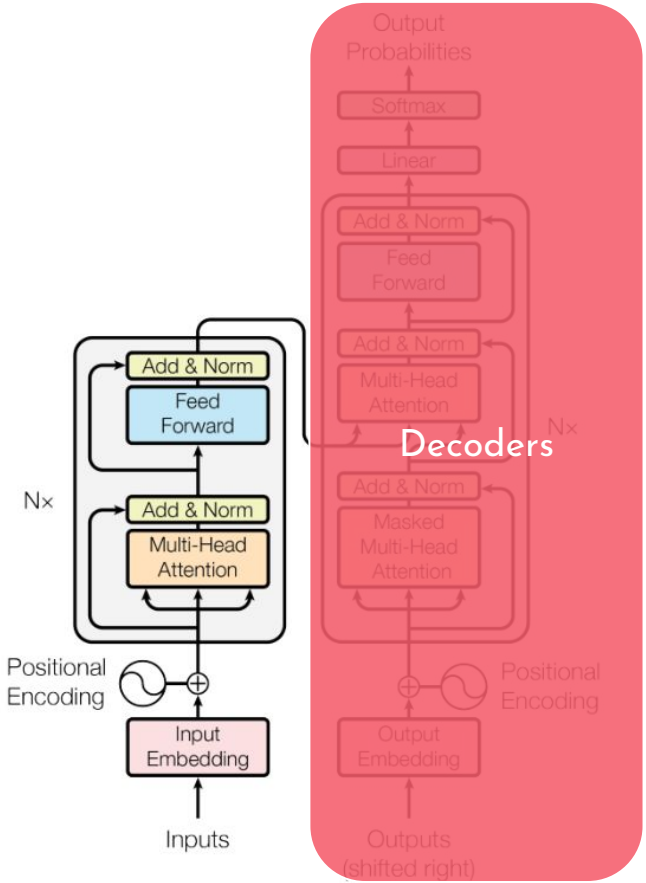
Model Architecture

- Encoder Input
- **Decoder Output**
  - generates an output sequence of symbols (y_1, ..., y_m) one element at a time



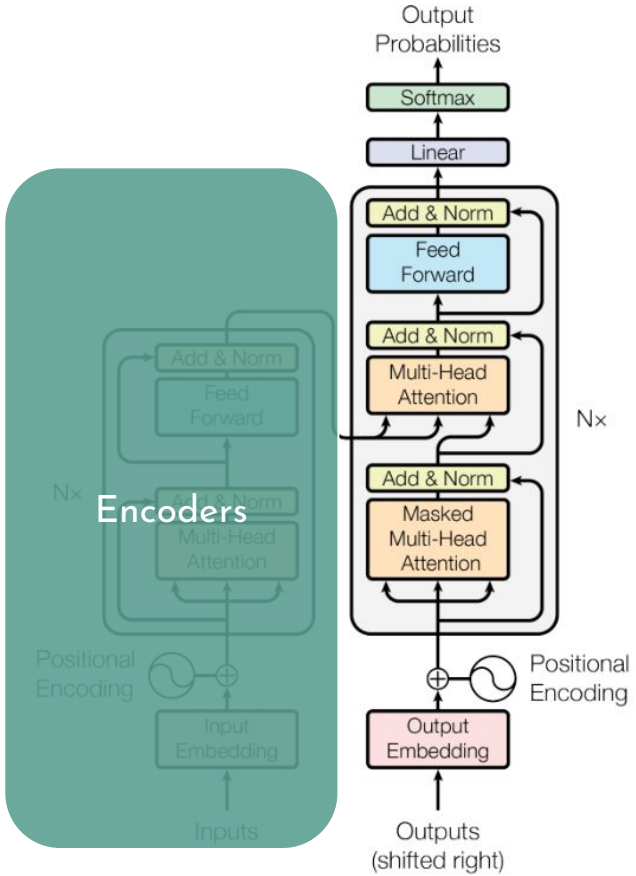
Encoder

- **layer N**
- **Sub-layers**
  - Multi-Head Attention
  - Feed Forward
- **output of each sub-layer**
  - LayerNorm(x + Sublayer(x))
- **Positional Encoding**

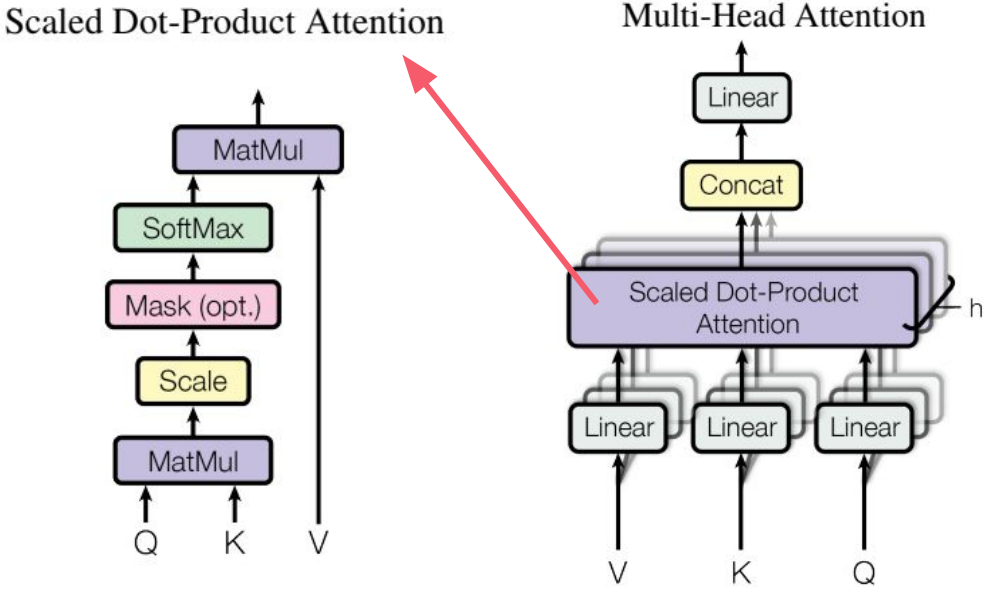


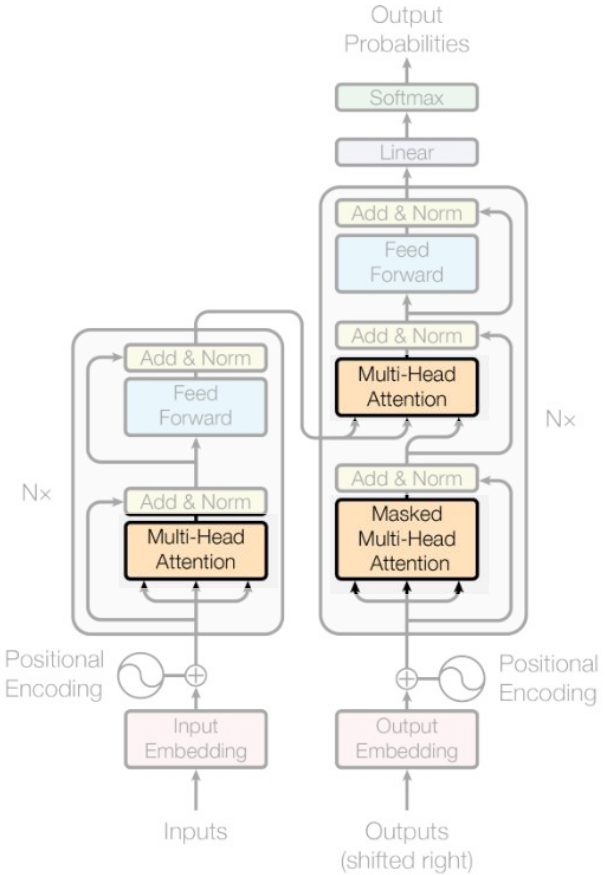
Decoder

- **layer N**
- **Sub-layers**
  - Multi-Head Attention
  - Feed Forward
  - Masked Multi-Head Attention
- **output of each sub-layer**
  - LayerNorm(x + Sublayer(x))
- **Positional Encoding**
- **Linear**
- **Softmax**



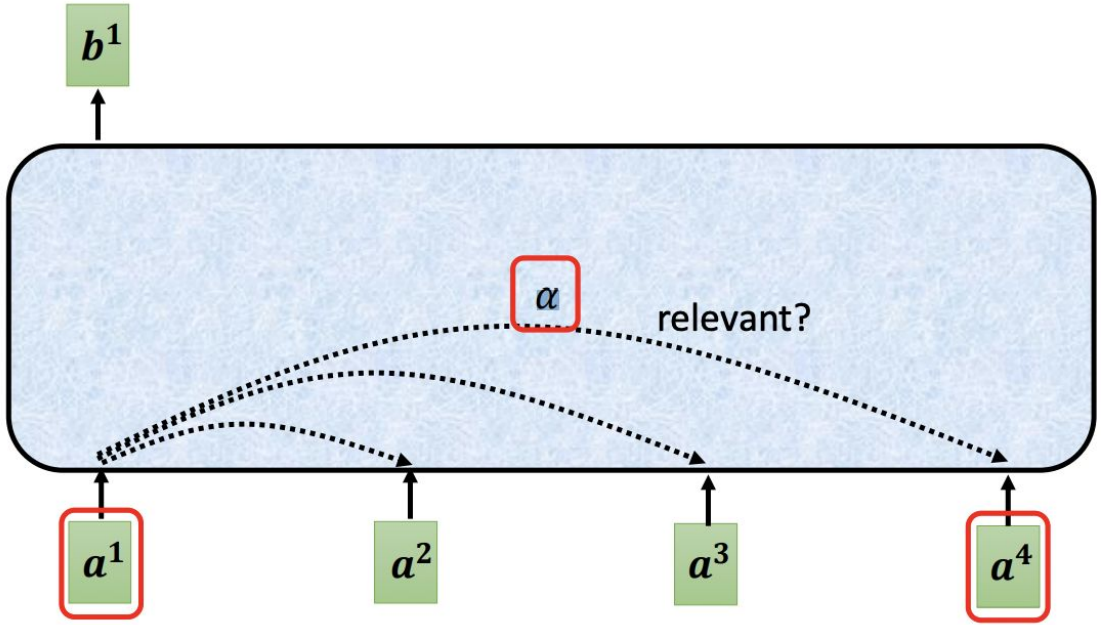
Multi-Head Attention



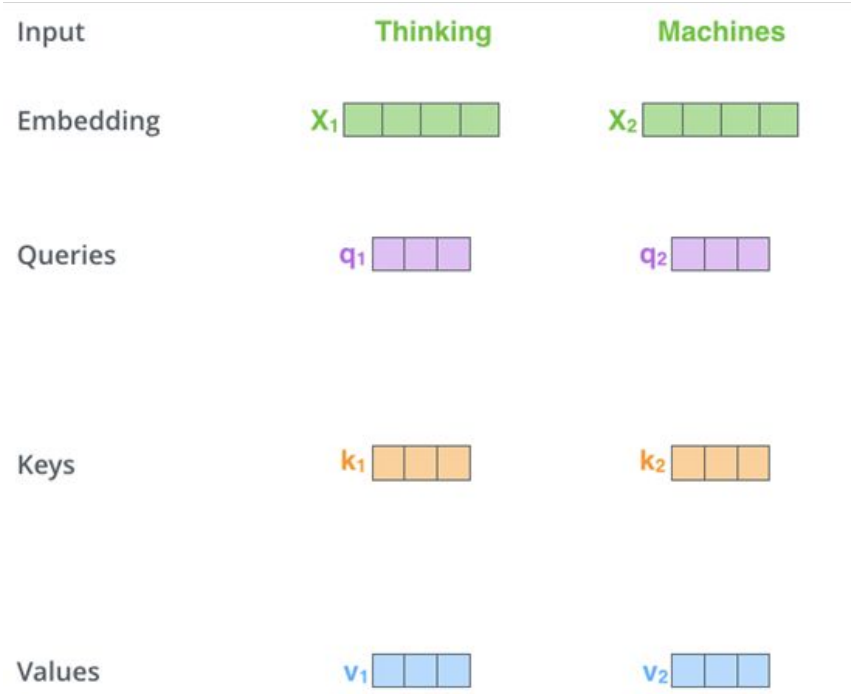


Self-attention

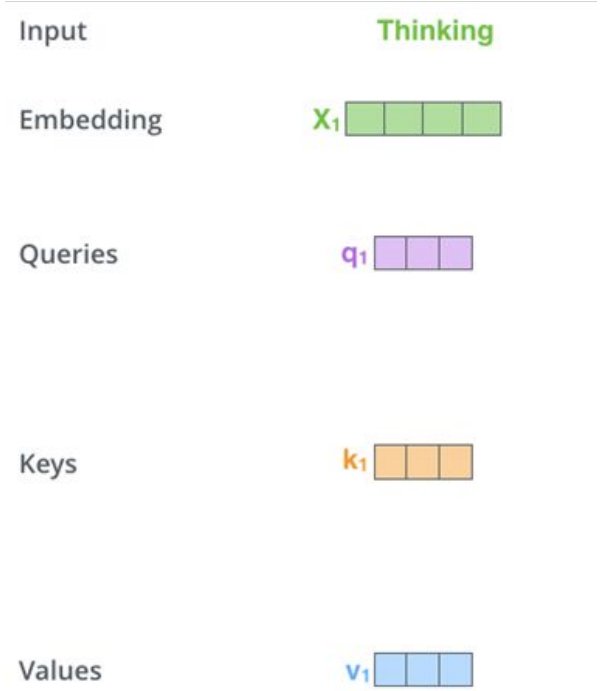
- **Consider the whole sequence**

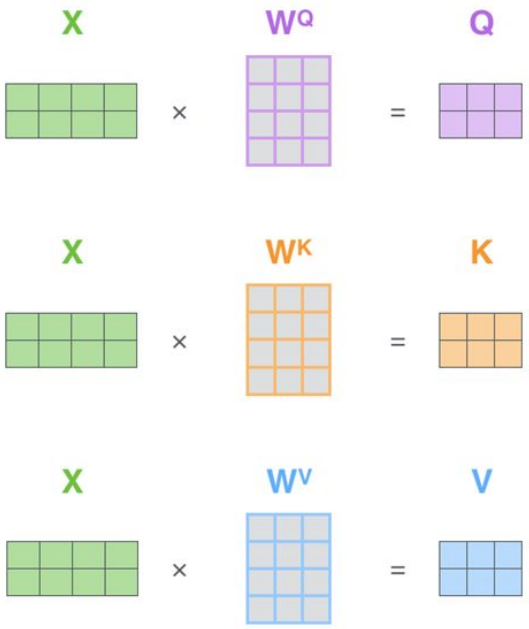


Attention visualization



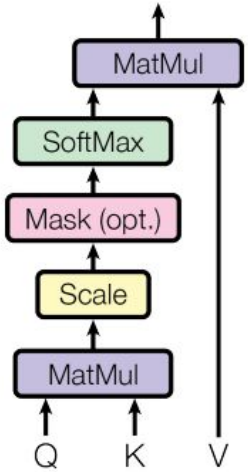
Attention visualization



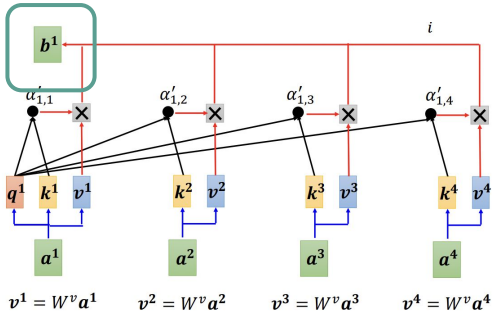


Attention visualization

Scaled Dot-Product Attention

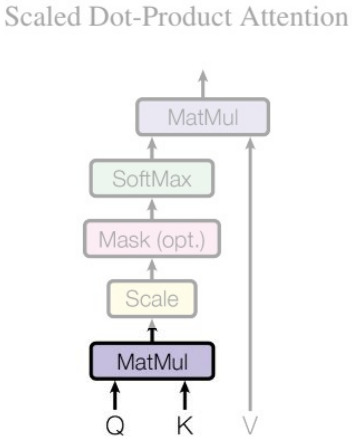


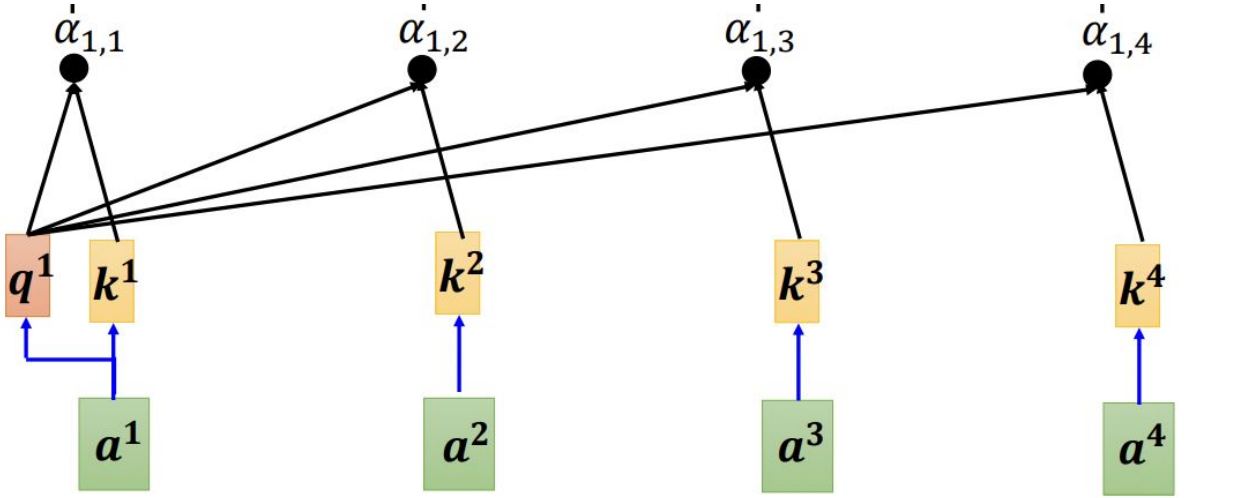
$$\text{softmax}\left(\frac{Q \times K^T}{\sqrt{d_k}}\right) V = Z$$



Self-attention

QK^T





$q^1 = W^q a^1$ $k^2 = W^k a^2$ $k^3 = W^k a^3$ $k^4 = W^k a^4$

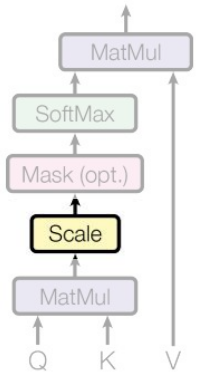
$k^1 = W^k a^1$

https://youtu.be/ugWDIIOHtPA

Scaled Dot-Product Attention

$$\frac{QK^T}{\sqrt{d_k}}$$

Scaled Dot-Product Attention

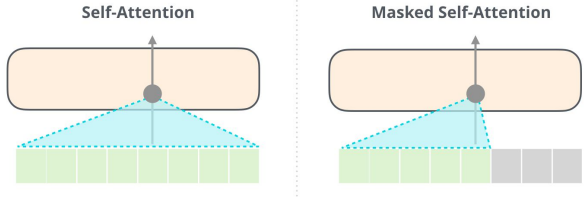


- **Problem**
  - large values of d_k, the dot products grow large in magnitude
- **queries and keys = d_k**
- **values = d_v**
- $d_k = d_v$
- **Solution**
  - scale the dot products by $\frac{1}{\sqrt{d_k}}$

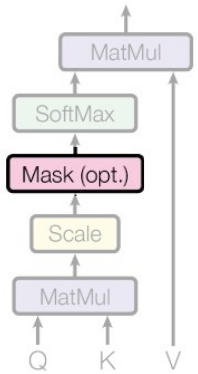
https://youtu.be/ugWDIIOHtPA

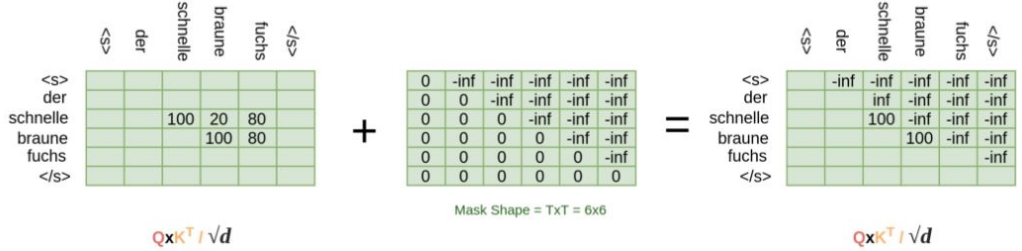
Masked self-attention

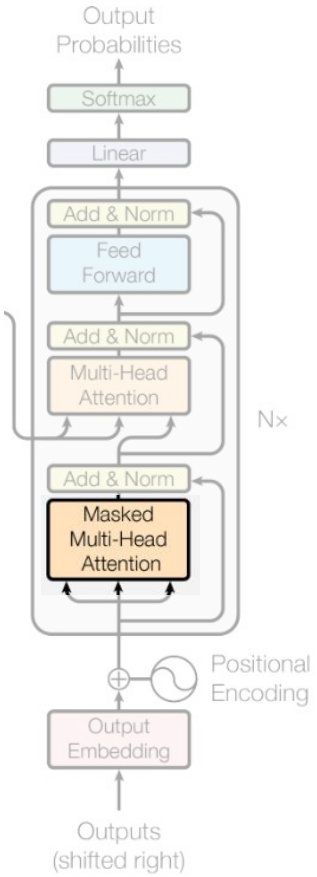
$$\frac{QK^T}{\sqrt{d_k}} + \text{attention Mask}$$



Scaled Dot-Product Attention







https://towardsdatascience.com/understanding-transformers-the-data-science-way-e4670a4ee076
https://jalammar.github.io/images/gpt2/self-attention-and-masked-self-attention.png

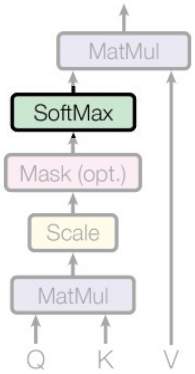
Self-attention

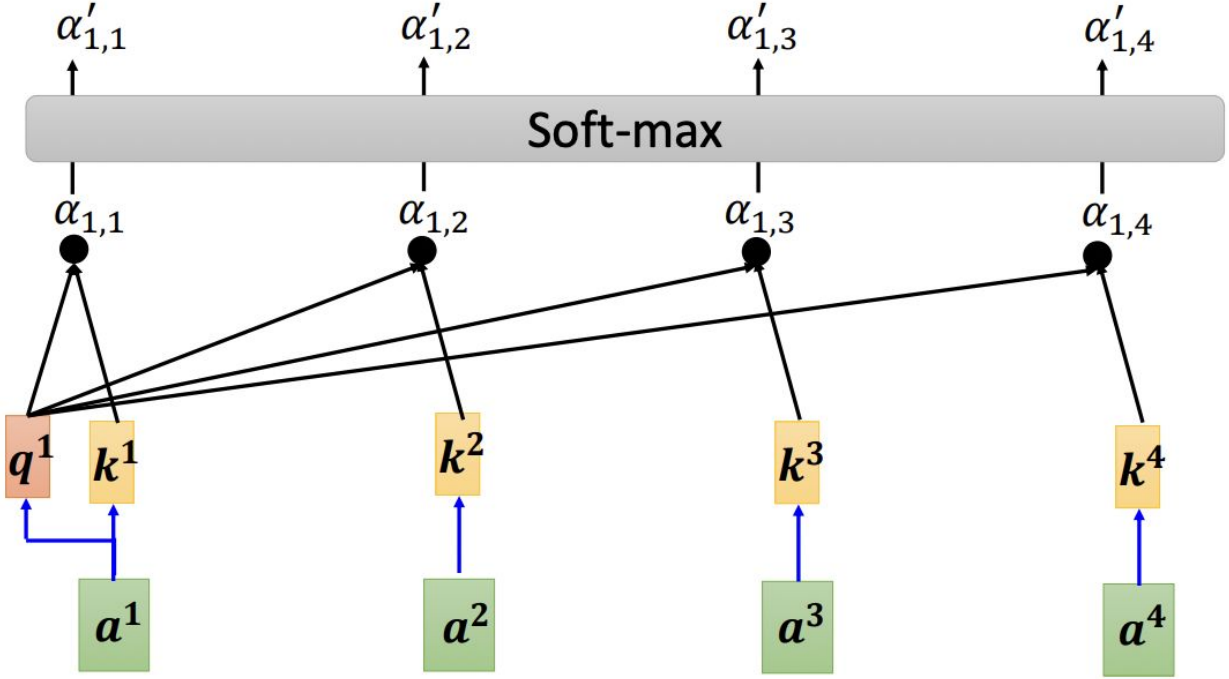
attention score / weight

$$\alpha'_{1,i} = exp(\alpha_{1,i}) / \sum_j exp(\alpha_{1,j})$$

$$\text{softmax}(\frac{QK^T}{\sqrt{d_k}})$$

Scaled Dot-Product Attention





$q^1 = W^q a^1$ $k^2 = W^k a^2$ $k^3 = W^k a^3$ $k^4 = W^k a^4$

$k^1 = W^k a^1$

https://youtu.be/ugWDIIOHtPA

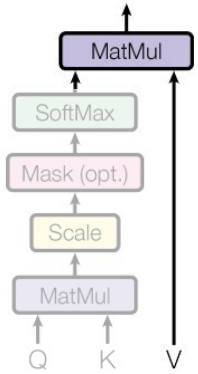
Self-attention

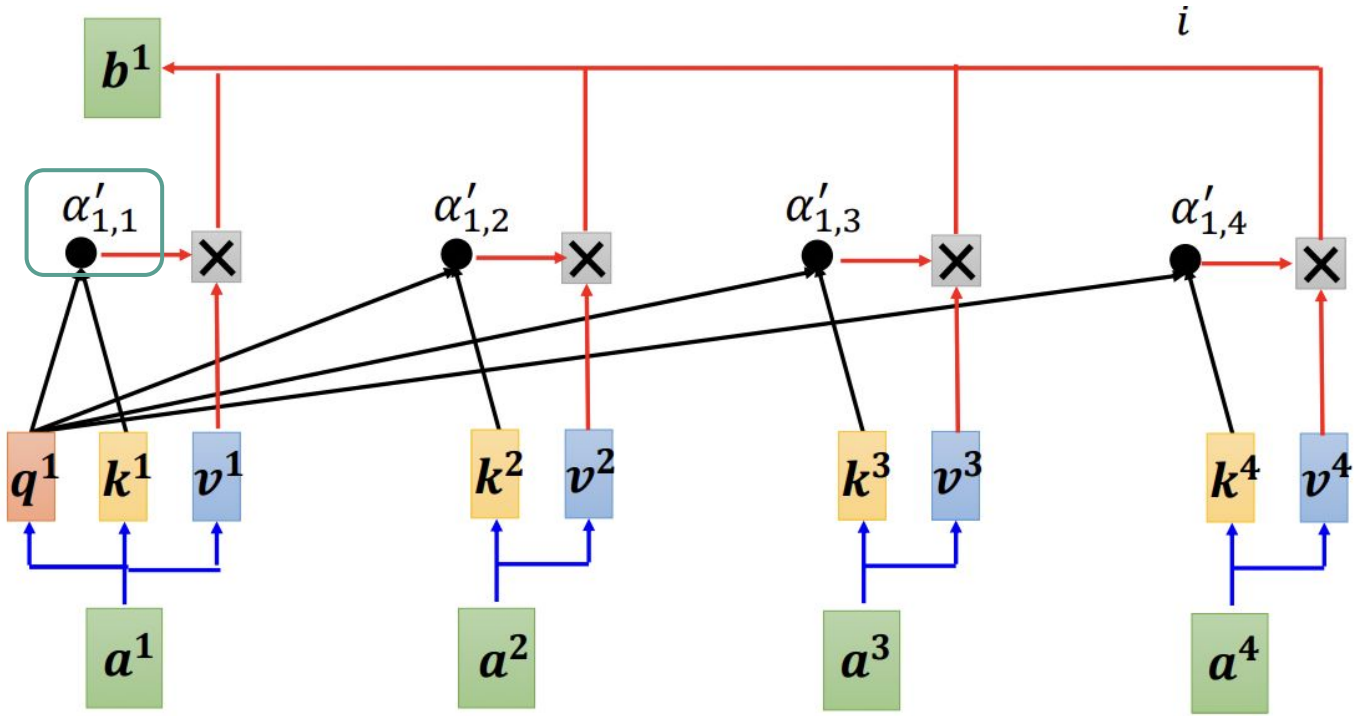
weighted sum

$$b^1 = \sum_i \alpha'_{1,i} v^i$$

$$\text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

Scaled Dot-Product Attention

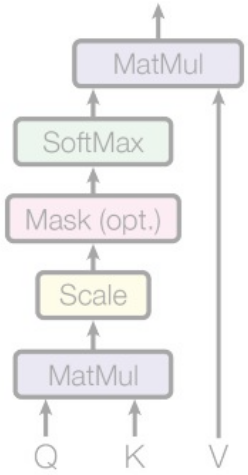




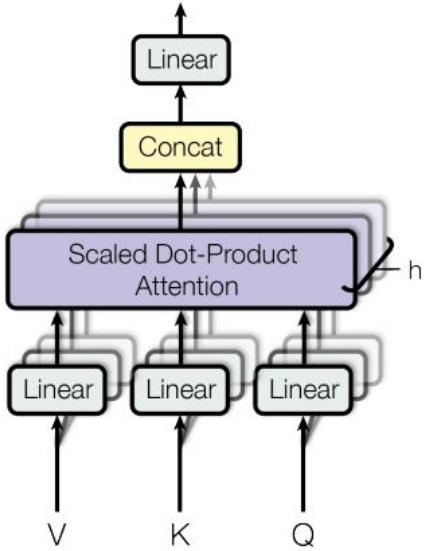
$$v^1 = W^v a^1 \qquad v^2 = W^v a^2 \qquad v^3 = W^v a^3 \qquad v^4 = W^v a^4$$

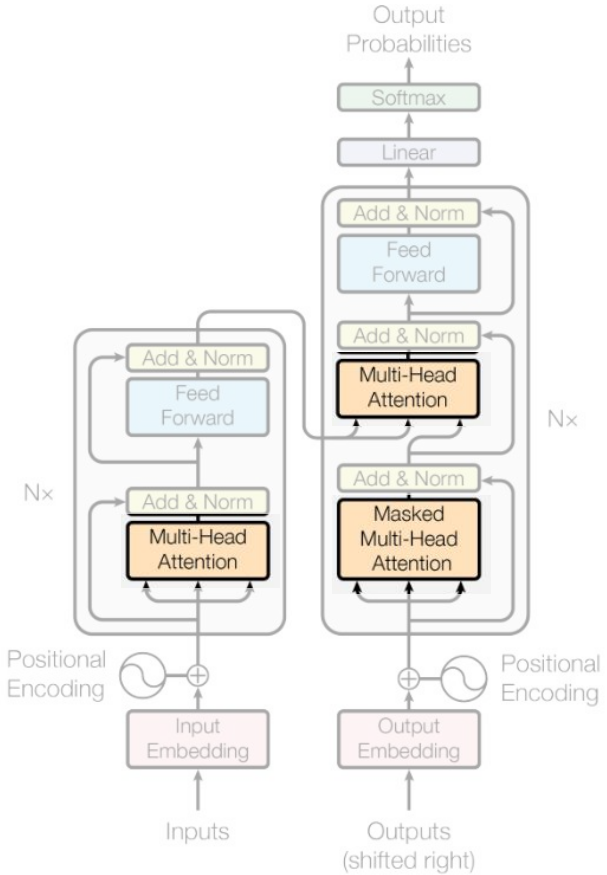
Multi-head Attention

Scaled Dot-Product Attention

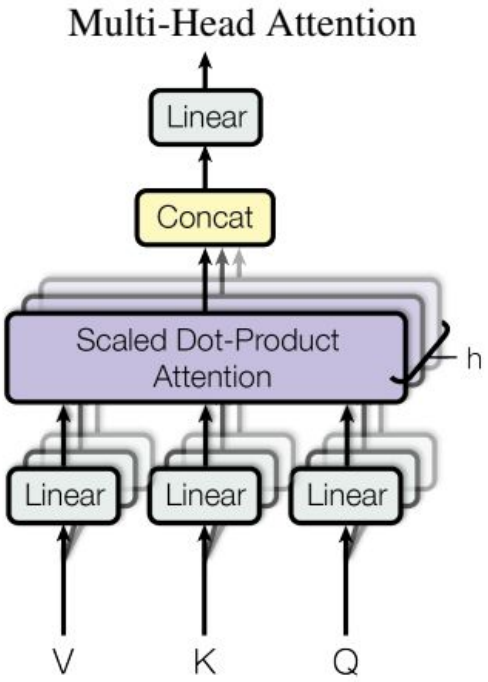


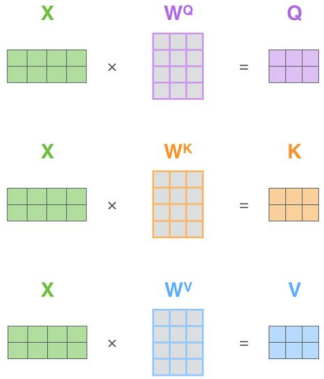
Multi-Head Attention





Multi-head Attention



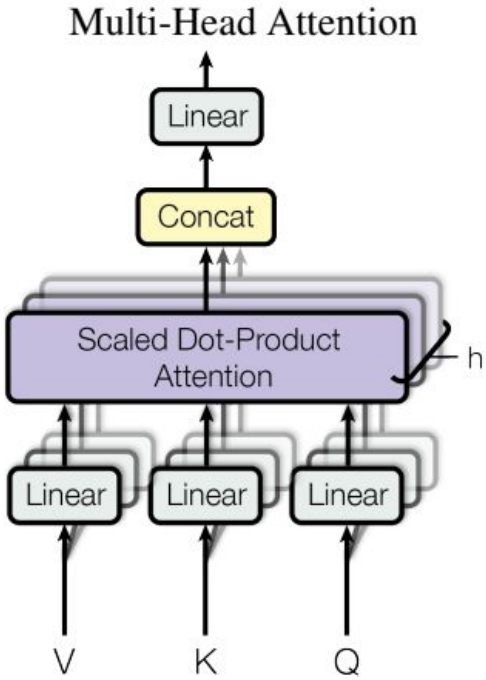


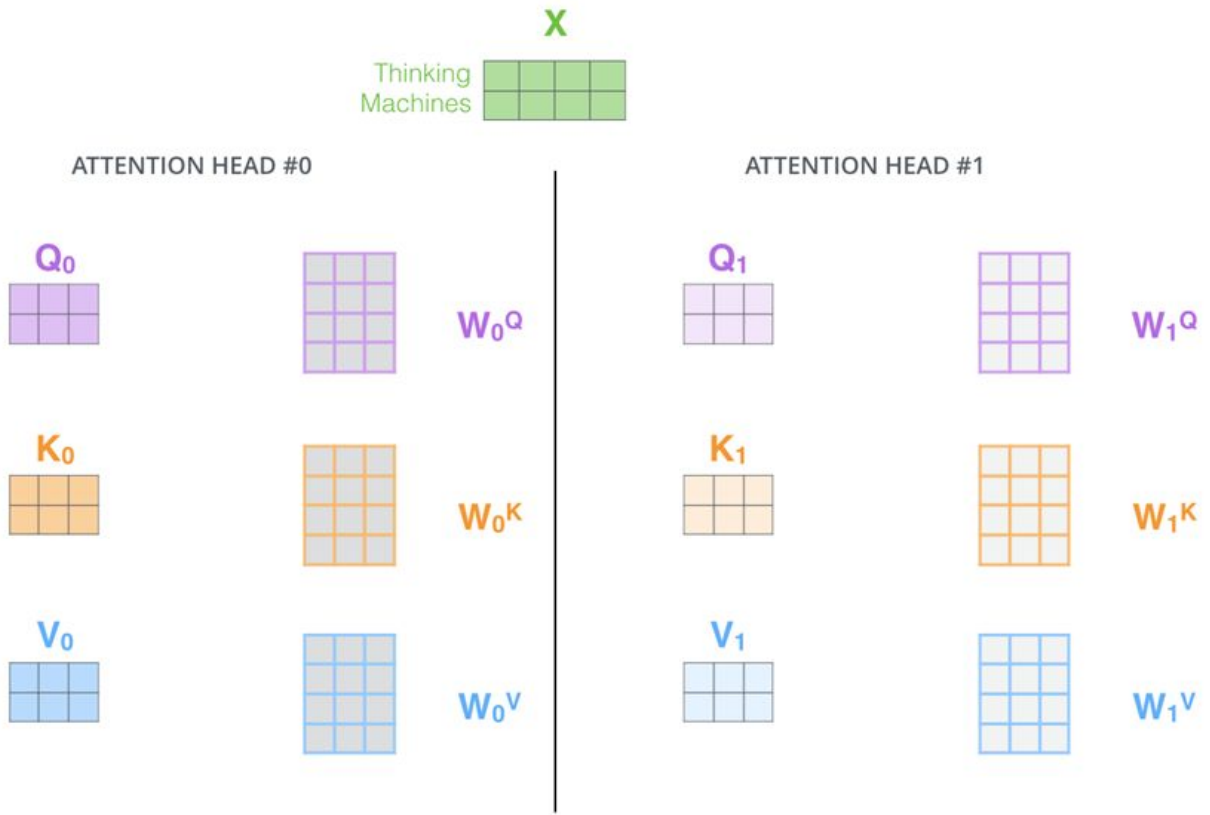
$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_\text{h})W^O$$

$$\text{where head}_\text{i} = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

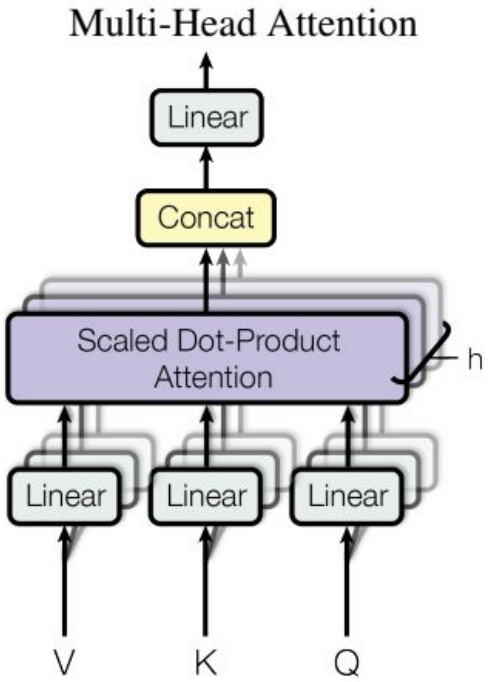
$$W_i^Q \in \mathbb{R}^{d_\text{model} \times d_k}, W_i^K \in \mathbb{R}^{d_\text{model} \times d_k}, W_i^V \in \mathbb{R}^{d_\text{model} \times d_v} \; W^O \in \mathbb{R}^{hd_v \times d_\text{model}}$$

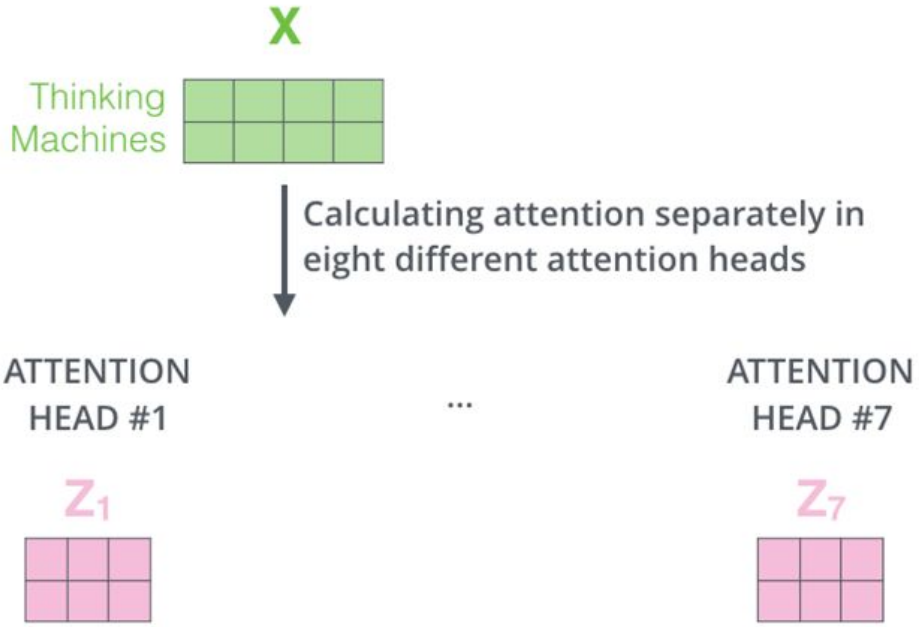
Attention visualization



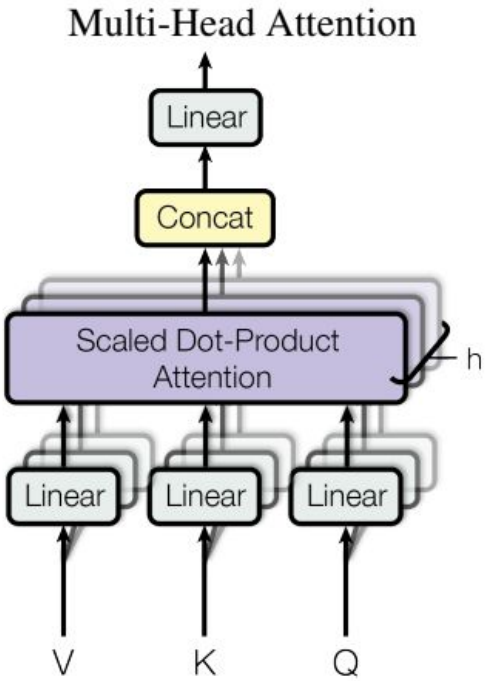


Attention visualization





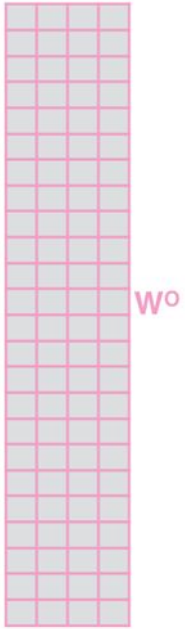
Attention visualization



1) Concatenate all the attention heads

2) Multiply with a weight matrix W^O that was trained jointly with the model

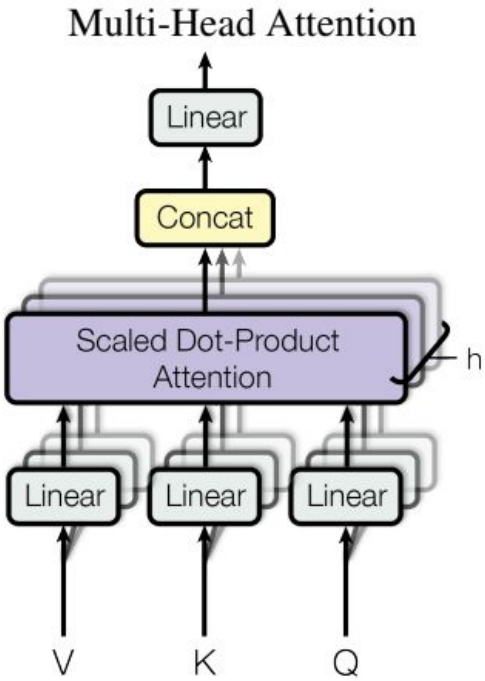
X



3) The result would be the Z matrix that captures information from all the attention heads. We can send this forward to the FFNN

=

Attention visualization



1) This is our input sentence*

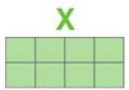
2) We embed each word*

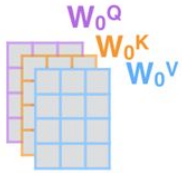
3) Split into 8 heads. We multiply X or R with weight matrices

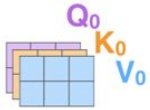
4) Calculate attention using the resulting Q/K/V matrices

5) Concatenate the resulting Z matrices, then multiply with weight matrix W^O to produce the output of the layer

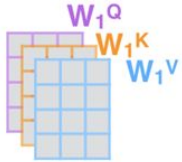
Thinking Machines

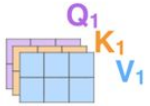




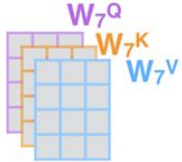


* In all encoders other than #0, we don't need embedding. We start directly with the output of the encoder right below this one





...

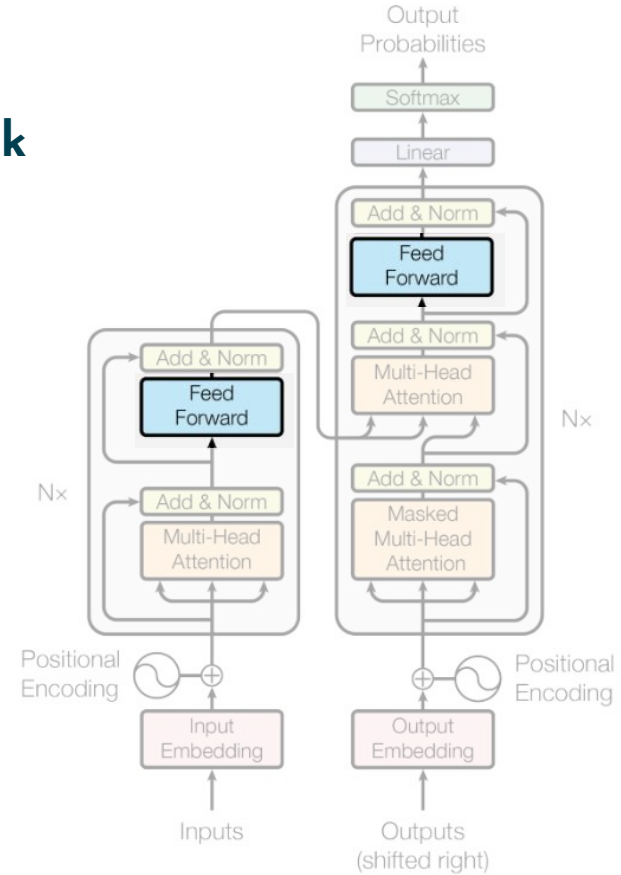


...

...

Feed Forward

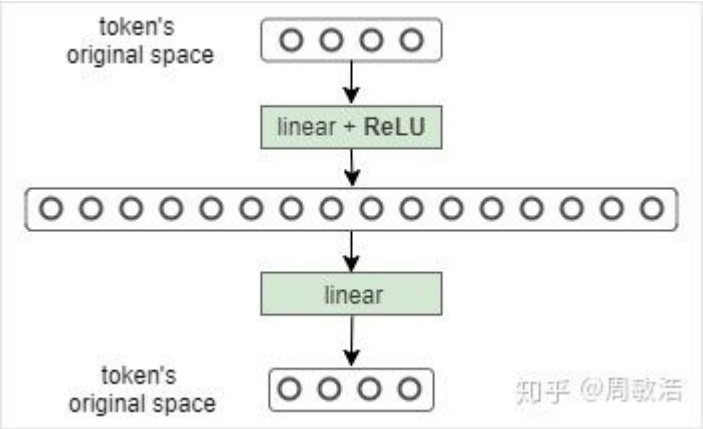
- **Fully connected feed-forward network**



Feed Forward

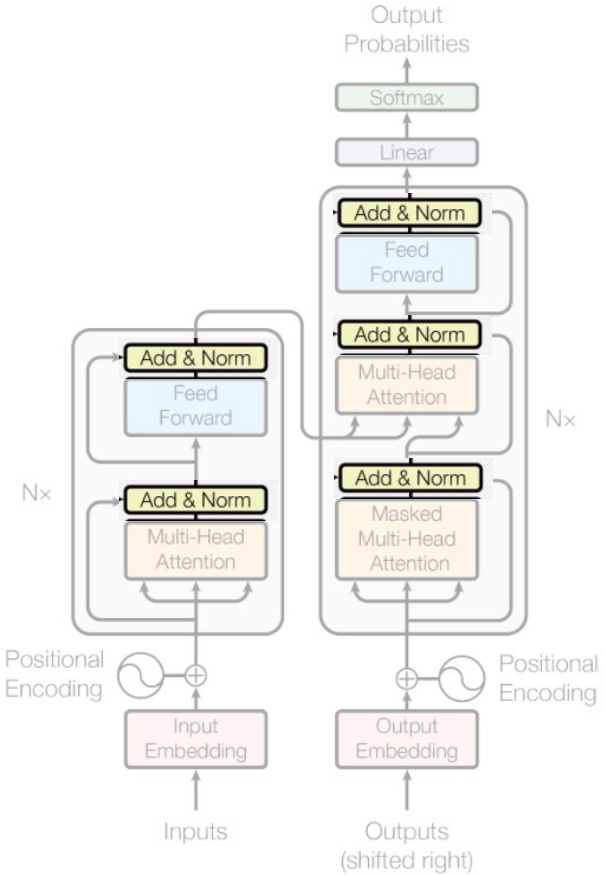
- **To extract the required information**

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



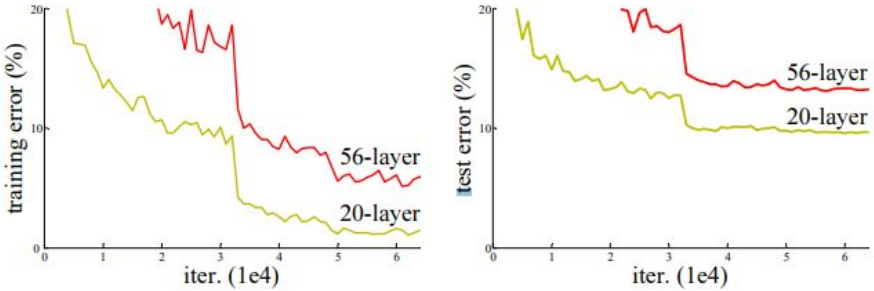
Add & Norm

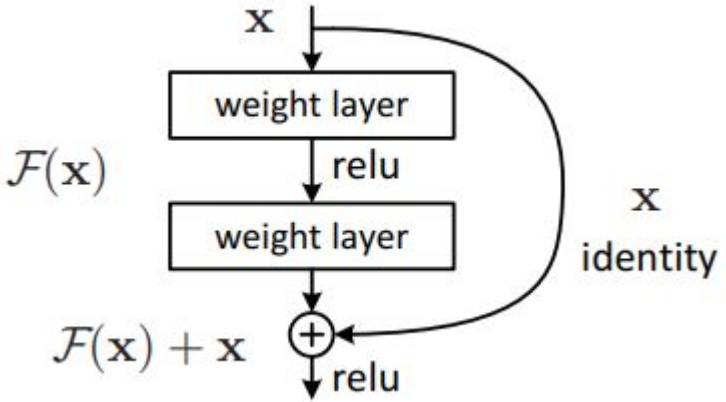
- **LayerNorm(x + Sublayer(x))**
- **Add**
  - Residual Connection
  - x + Sublayer(x)
- **Norm**
  - Layer Normaliztion



Residual Connection

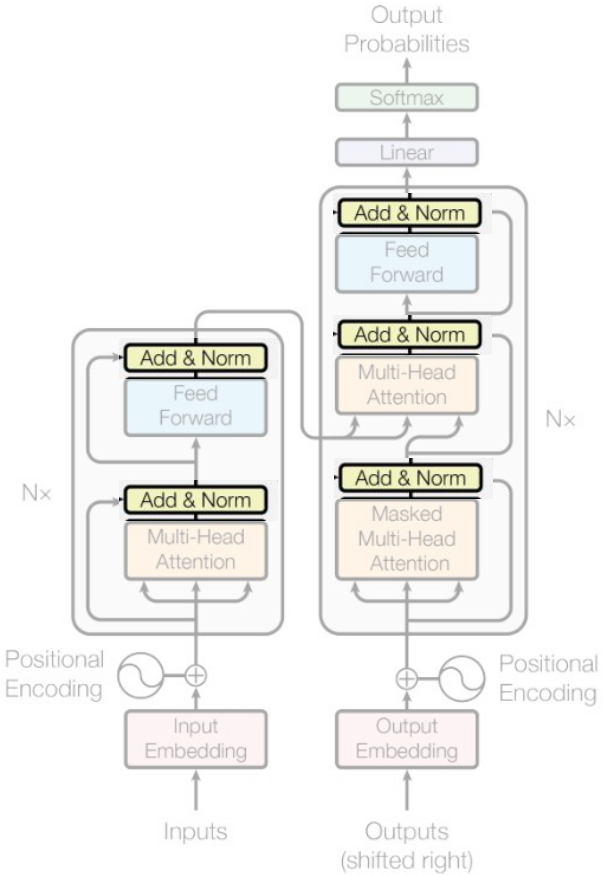
- **Degradation problem**



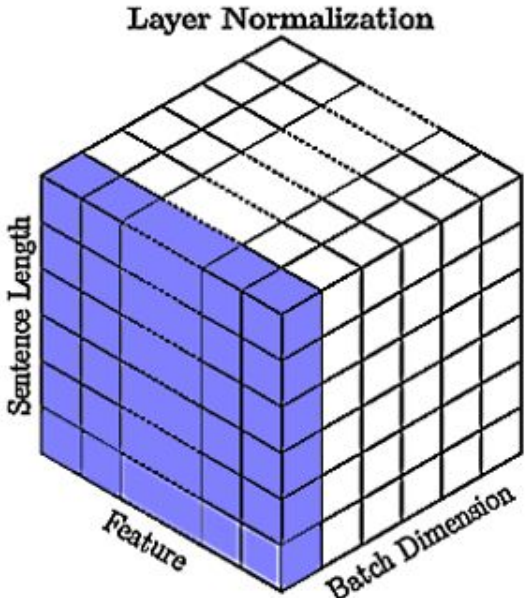


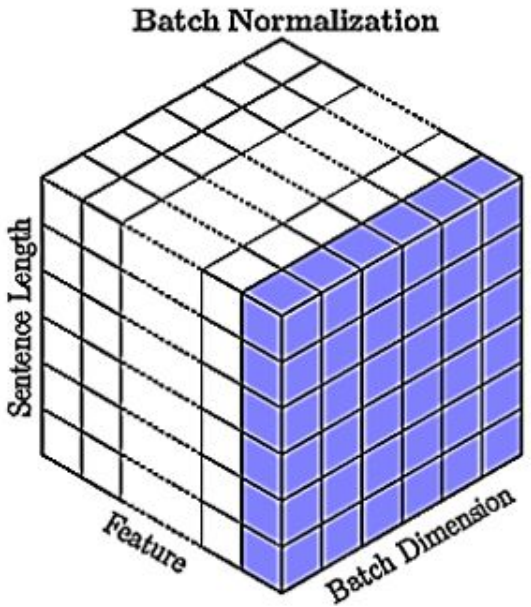
Add & Norm

- LayerNorm(x + Sublayer(x))
- Add
  - Residual Connection
  - x + Sublayer(x)
- **Norm**
  - Layer Normaliztion



Layer Normaliztion

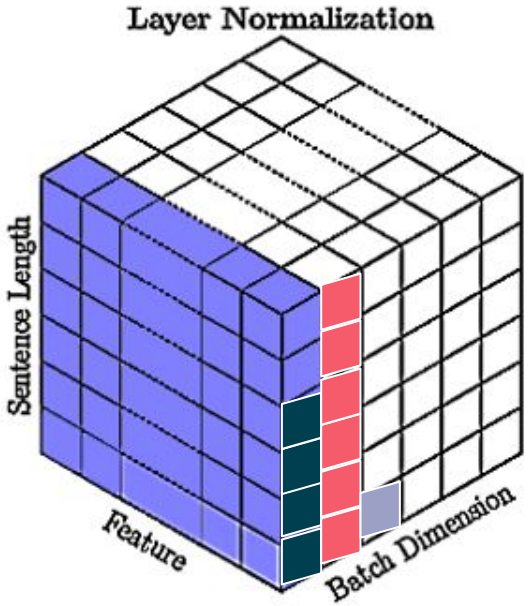




Transformer use this

https://www.kaggle.com/code/halflingwizard/how-does-layer-normalization-work/notebook

Layer Normaliztion



1. This is a book
2. To be or not to be
3. Yummy

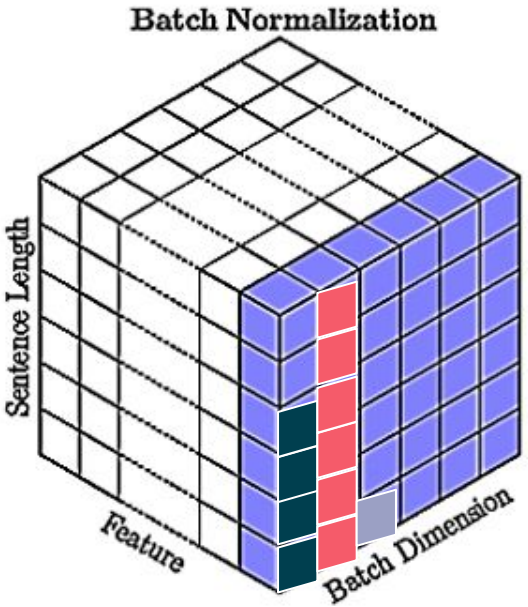
This	0.51	0.36
is	0.25	0.44
a	0.88	0.96
book	0.13	0.27

To	0.31	0.18
be	0.45	0.44
or	0.78	0.15
not	0.13	0.72
to	0.31	0.18
be	0.45	0.44

$$x_{norm} = \frac{x - avg(x)}{\sqrt{var(x)}}$$

https://www.kaggle.com/code/halflingwizard/how-does-layer-normalization-work/notebook

Batch Normaliztion



1. This is a book
2. To be or not to be
3. Yummy

Yummy	0.38	0.11

This	0.51	0.36
is	0.25	0.44
a	0.88	0.96
book	0.13	0.27

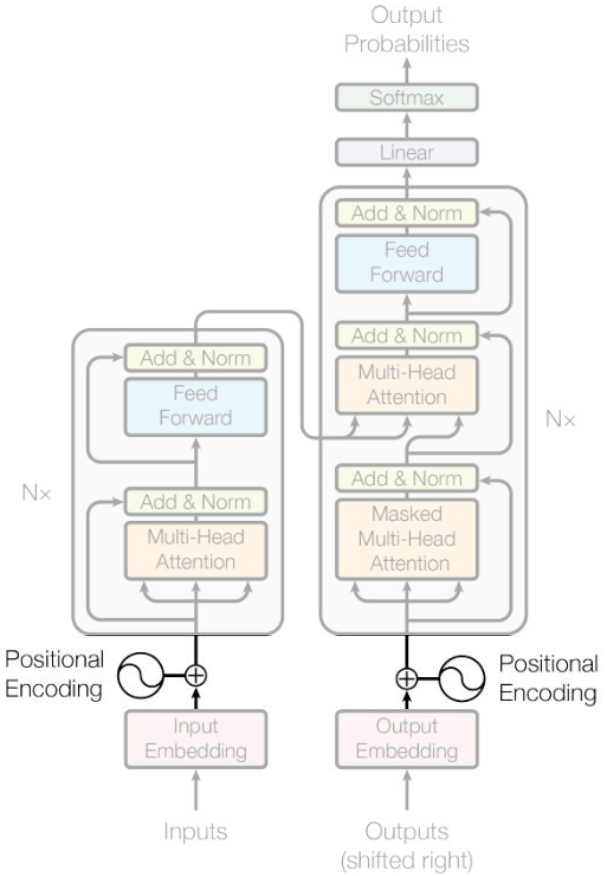
To	0.31	0.18
be	0.45	0.44
or	0.78	0.15
not	0.13	0.72
to	0.31	0.18
be	0.45	0.44

$$x_{norm} = \frac{x - avg(x)}{\sqrt{var(x)}}$$

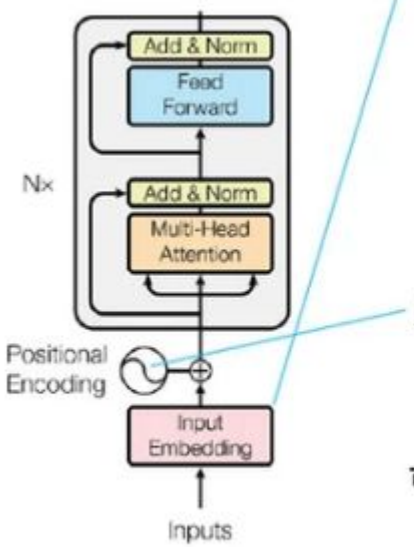
https://www.kaggle.com/code/halflingwizard/how-does-layer-normalization-work/notebook

Positional Encoding

- **Lack of positional information**



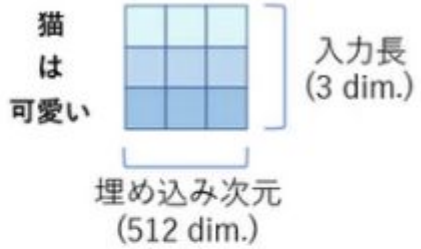
Positional Encoding



入力例：[“猫”, “は”, “可愛い”]

Embedding Layer　単語をベクトル化(単語列なので行列)

[“猫”, “は”, “可愛い”] ➡



Positional Encoding　時系列情報の付与

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

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

https://towardsdatascience.com/understanding-transformers-the-data-science-way-e4670a4ee076

Positional Encoding

Sequence	Index of token, k	Positional Encoding Matrix with d=4, n=100			
		i=0	i=0	i=1	i=1
I	0	P_{00}=sin(0) = 0	P_{01}=cos(0) = 1	P_{02}=sin(0) = 0	P_{03}=cos(0) = 1
am	1	P_{10}=sin(1/1) = 0.84	P_{11}=cos(1/1) = 0.54	P_{12}=sin(1/10) = 0.10	P_{13}=cos(1/10) = 1.0
a	2	P_{20}=sin(2/1) = 0.91	P_{21}=cos(2/1) = -0.42	P_{22}=sin(2/10) = 0.20	P_{23}=cos(2/10) = 0.98
Robot	3	P_{30}=sin(3/1) = 0.14	P_{31}=cos(3/1) = -0.99	P_{32}=sin(3/10) = 0.30	P_{33}=cos(3/10) = 0.96

Positional Encoding Matrix for the sequence 'I am a robot'

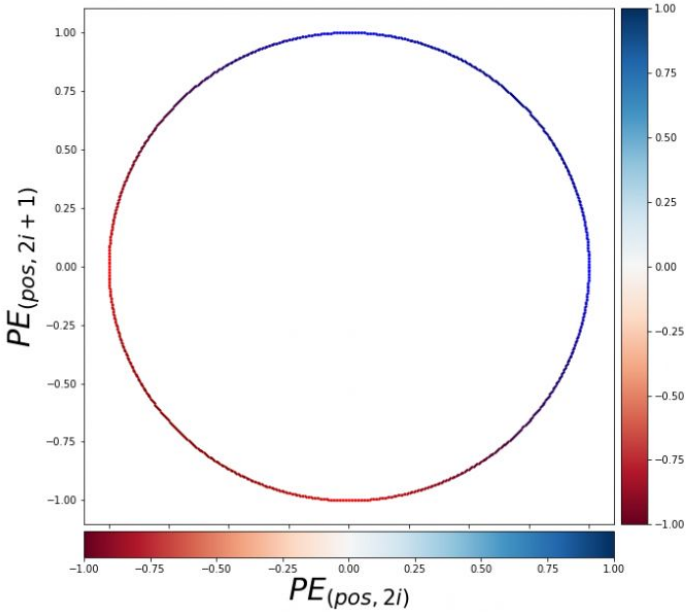
$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{\text{model}}})$$

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

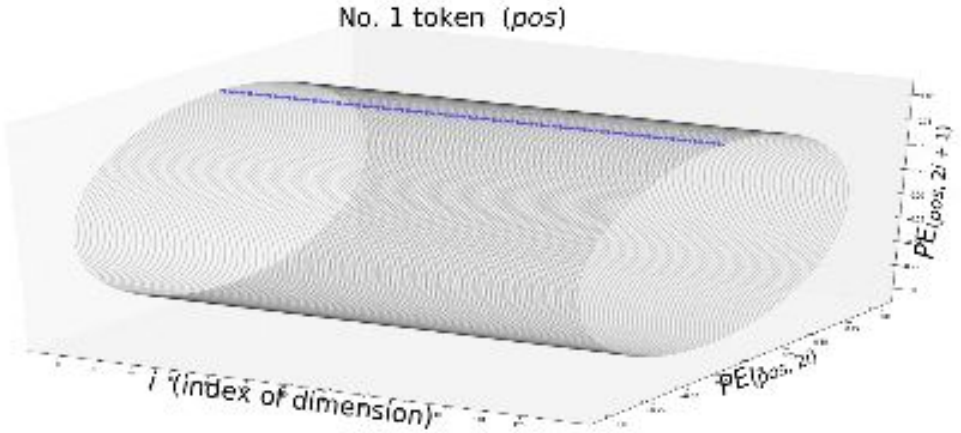
https://towardsdatascience.com/understanding-transformers-the-data-science-way-e4670a4ee076

(additional Info.)

Positional Emb. visualization



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



https://data-science-blog.com/blog/2021/04/22/positional-encoding-residual-connections-padding-masks-all-the-details-of-transformer-model/

Improving Language Understanding
by Generative Pre-Training

GPT
2018/06

3.7K

Language Models are
Unsupervised Multitask Learners

GPT-2
2019/02

3.5K

Language Models are Few-Shot Learners

GPT-3
2020/05

4.9K

Transformer
2017/06

49K

Attention Is All You Need

BERT
2018/10

46K

Pre-training of Deep Bidirectional
Transformers for Language Understanding

RoBERTa
2019/07

4.1K

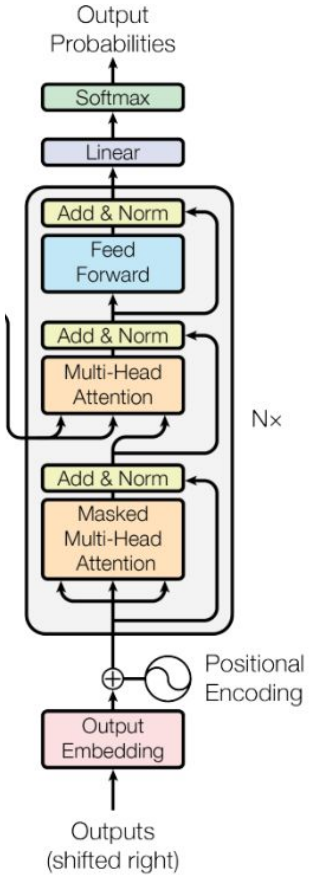
RoBERTa: A Robustly Optimized
BERT Pretraining Approach

Source

- **GPT**
  - Improving Language Understanding by Generative Pre-Training
  - https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf
- **GPT-2**
  - Language Models are Unsupervised Multitask Learners
  - https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
- **GPT-3**
  - Language Models are Few-Shot Learners
  - https://arxiv.org/pdf/2005.14165.pdf

GPT

- **Problem**
  - Large unlabeled text corpora are abundant, labeled data for learning these specific tasks is scarce
- **Goal**
  - Realize generative pre-training of a language model on a diverse corpus of unlabeled text
  - Discriminative fine-tuning on each specific task
- **Approach**
  - Semi-supervised approach for language understanding tasks
    - unsupervised pre-training
    - supervised fine-tuning.
  - Transformer decoder architecture

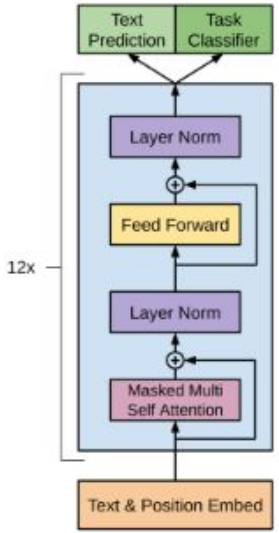


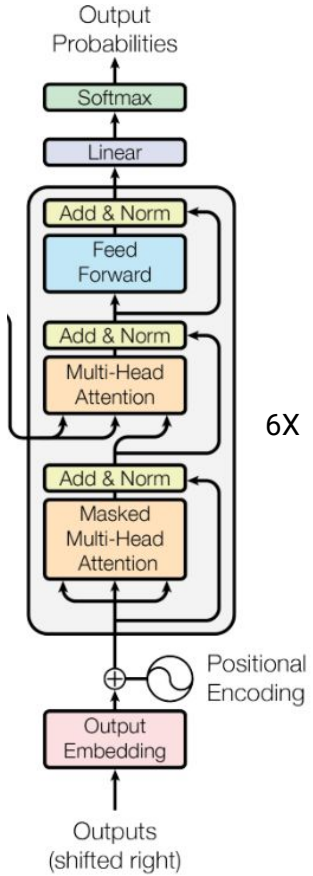
Framework

1. **Unsupervised pre-training**
2. **Supervised fine-tuning**
3. **Task-specific input transformations**

Training task

- predict next word





1. Unsupervised pre-training

- **Unsupervised corpus of tokens** $\mathcal{U} = \{u_1, \ldots, u_n\}$
- **Maximize the likelihood of** $L1(\mathcal{U})$

$$L_1(\mathcal{U}) = \sum \log P(u_i | u_{i-k}, \ldots, u_{i-1}; \Theta)$$

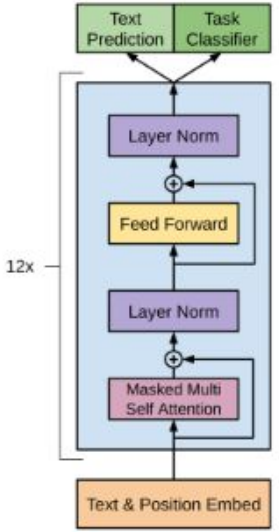
k : context window size

1. Unsupervised pre-training

- **Context vector of tokens** $U = (u{-}k, \ldots, u{-}1)$

- **Number of layers** n
- **Token embedding matrix** We
- **Position embedding matrix** Wp

$$h_0 = UW_e + W_p$$
$$h_l = \texttt{transformer\_block}(h_{l-1}) \forall i \in [1, n]$$
$$P(u) = \texttt{softmax}(h_n W_e^T)$$

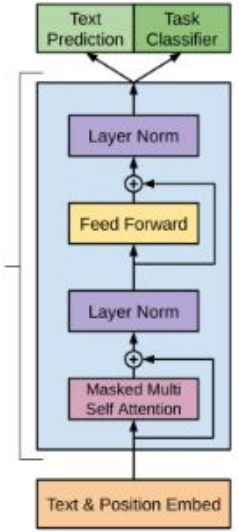


2. Supervised fine-tuning

- **Adapt the parameters to the supervised <u>target task</u>**

- **<u>Final</u> transformer block's activation** h^m_l
- **Parameters** Wy
- **Sequence of input tokens** $x^1, \ldots, x^m$
- **Label** y

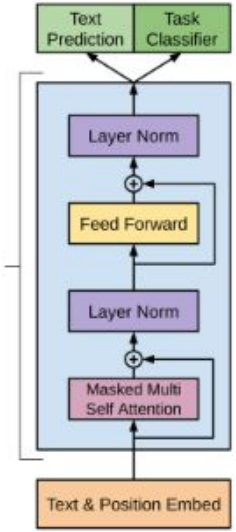
$$P(y|x^1, \ldots, x^m) = \texttt{softmax}(h^m_l W_y)$$



2. Supervised fine-tuning

- **Maximize the likelihood of** $L_2(C)$

$$L_2(\mathcal{C}) = \sum_{(x,y)} \log P(y|x^1, \ldots, x^m)$$

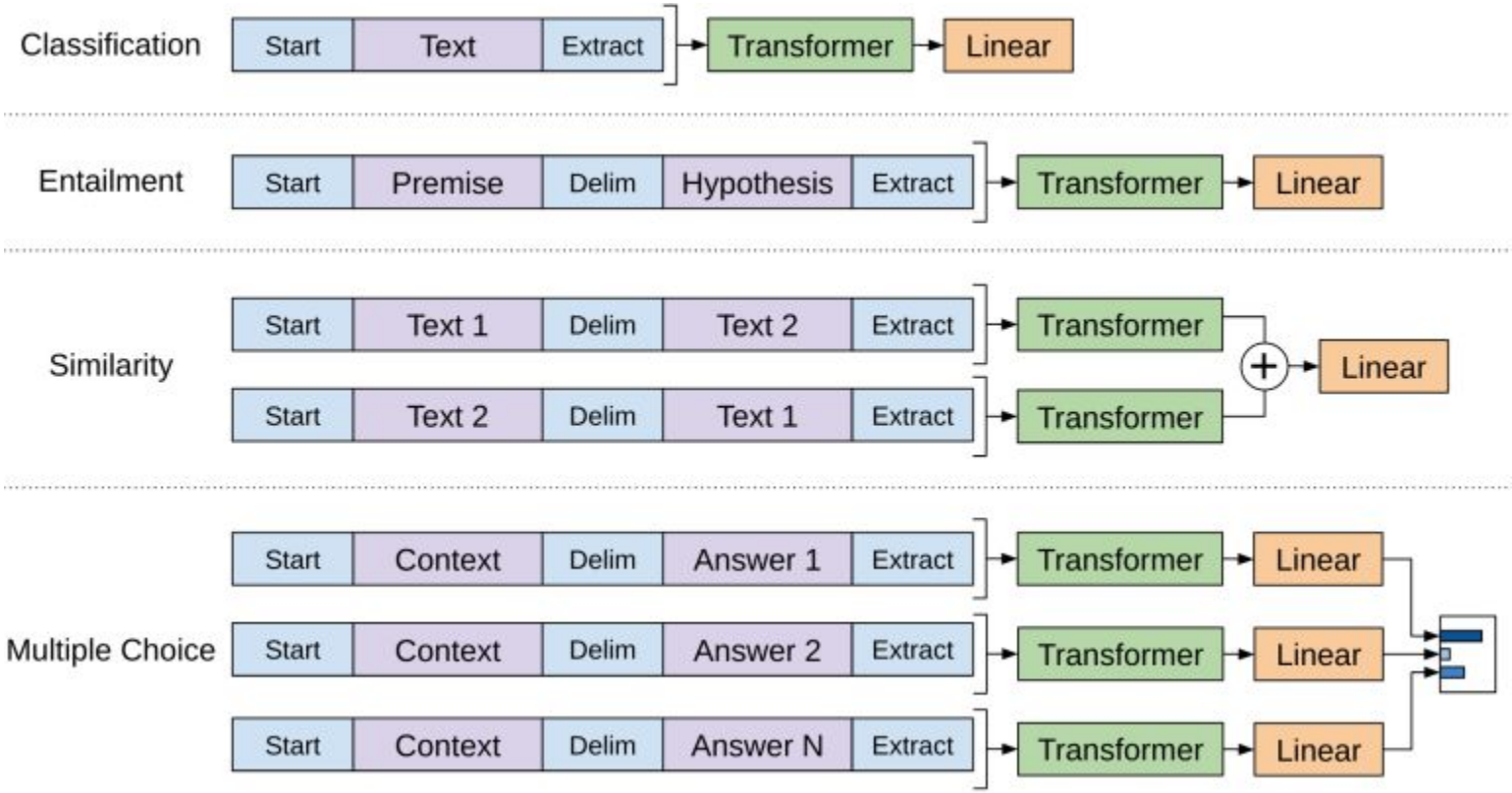


2. Supervised fine-tuning

- **Optimize** $L_3(C)$
- **Weight λ**

$$L_3(C) = L_2(C) + \lambda * L_1(C)$$

3. Task-specific input transformations



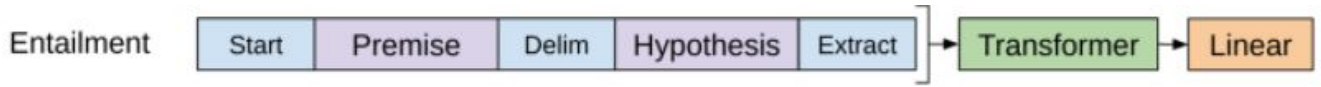
Text classification

- **Fine-tune model as described in Supervised fine-tuning**

Textual entailment

- **Premise p and hypothesis h token sequences, with a delimiter token ($) in between**

Premise	Label	Hypothesis
A man inspects the uniform of a figure in some East Asian country.	contradiction	The man is sleeping.
An older and younger man smiling.	neutral	Two men are smiling and laughing at the cats playing on the floor.
A soccer game with multiple males playing.	entailment	Some men are playing a sport.

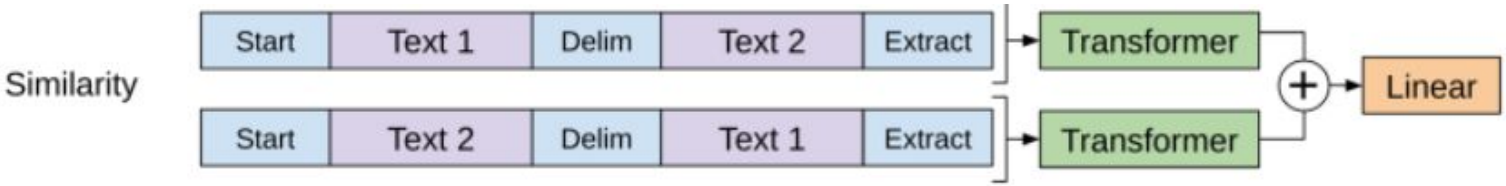


https://paperswithcode.com/task/natural-language-inference

Similarity

- No inherent ordering of the two sentences
- Independently to produce two sequence representations h^m_l
- Add two h^m_l element-wise

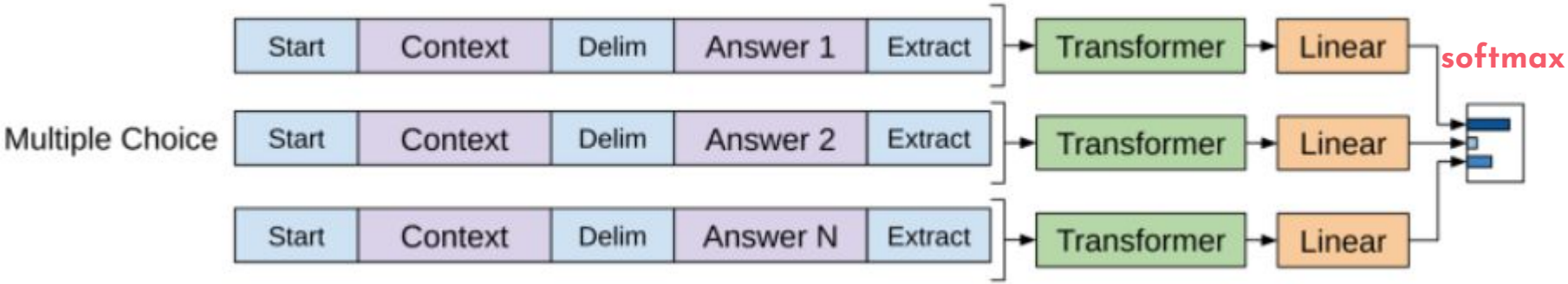
	sent_1	sent_2
0	A girl is styling her hair.	A girl is brushing her hair.
1	A group of men play soccer on the beach.	A group of boys are playing soccer on the beach.
2	One woman is measuring another woman's ankle.	A woman measures another woman's ankle.
3	A man is cutting up a cucumber.	A man is slicing a cucumber.
4	A man is playing a harp.	A man is playing a keyboard.



https://nlp.town/blog/sentence-similarity/

Question Answering & Commonsense Reasoning

- **[z; q; $; ak]**
  - Context document z
  - Question q
  - Set of possible answers $\{ak\}$
- **Sequences are processed independently**
- **Output a distribution over possible answers**



PsyQA:Question

- **Chinese dataset of Psychological health support in the form of Question-Answer pair**

Question (Post Title)

为什么有些事情越想心越闷?
The more I think about some things, the more upset I feel. Why?

Description (Post Content)

比如家里人说我和别人相亲的事，我是不太主动比较被动的人……又怪自己，又恨别人到处说，搞得心里很难受很郁闷，这该怎么办?
For example, my family asked me to go on a blind date with others. I am not an active but passive person. ... I blame myself and blame others for speaking ill of me everywhere, making me very uncomfortable and depressed. What should I do?

Keywords

情绪 表达情绪 情绪调节 情绪智力
Emotion, Emotion Expression, Emotion Regulation, EQ

你好呀～事情越想越闷可能是陷入了反刍思维中。反刍式思考是指 …… 反刍思维作为一种认知，对情绪也有重要的影响。在这种情况下，你首先要冷静下来 ……比如自己闷在家里没出去相亲，家人就说自己是不是想打光棍儿。其实你仔细看这两件事情并没有因果关系。……但这样的逻辑也是不太合情理的。当然，在这种情况下，你也可以使用转移注意力的方式，让自己的情绪稍微平复下来。比如做一次冥想练习，或者出去做运动。

Hi ~ The more you think about it, the more depressed you feel. This is maybe because you are trapped in ruminant thinking. Ruminant thinking means that ..... Ruminant thinking, as a form of cognition, also has an important effect on emotion. In this case, you need to calm down first… For example, you stayed at home and didn't go out for a blind date, and your family said that you just wanted to be single. When you look at it carefully, there is no causal relationship between the two events. ..... But this logic doesn't work. Of course, in this case, you can also distract your attention to calm yourself down a bit. Take a meditation practice, or go outside to exercise.

Improving Language Understanding
by Generative Pre-Training

GPT
2018/06

3.7K

Language Models are
Unsupervised Multitask Learners

GPT-2
2019/02

3.5K

Language Models are Few-Shot Learners

GPT-3
2020/05

4.9K

Transformer
2017/06

49K

Attention Is All You Need

BERT
2018/10

46K

Pre-training of Deep Bidirectional
Transformers for Language Understanding

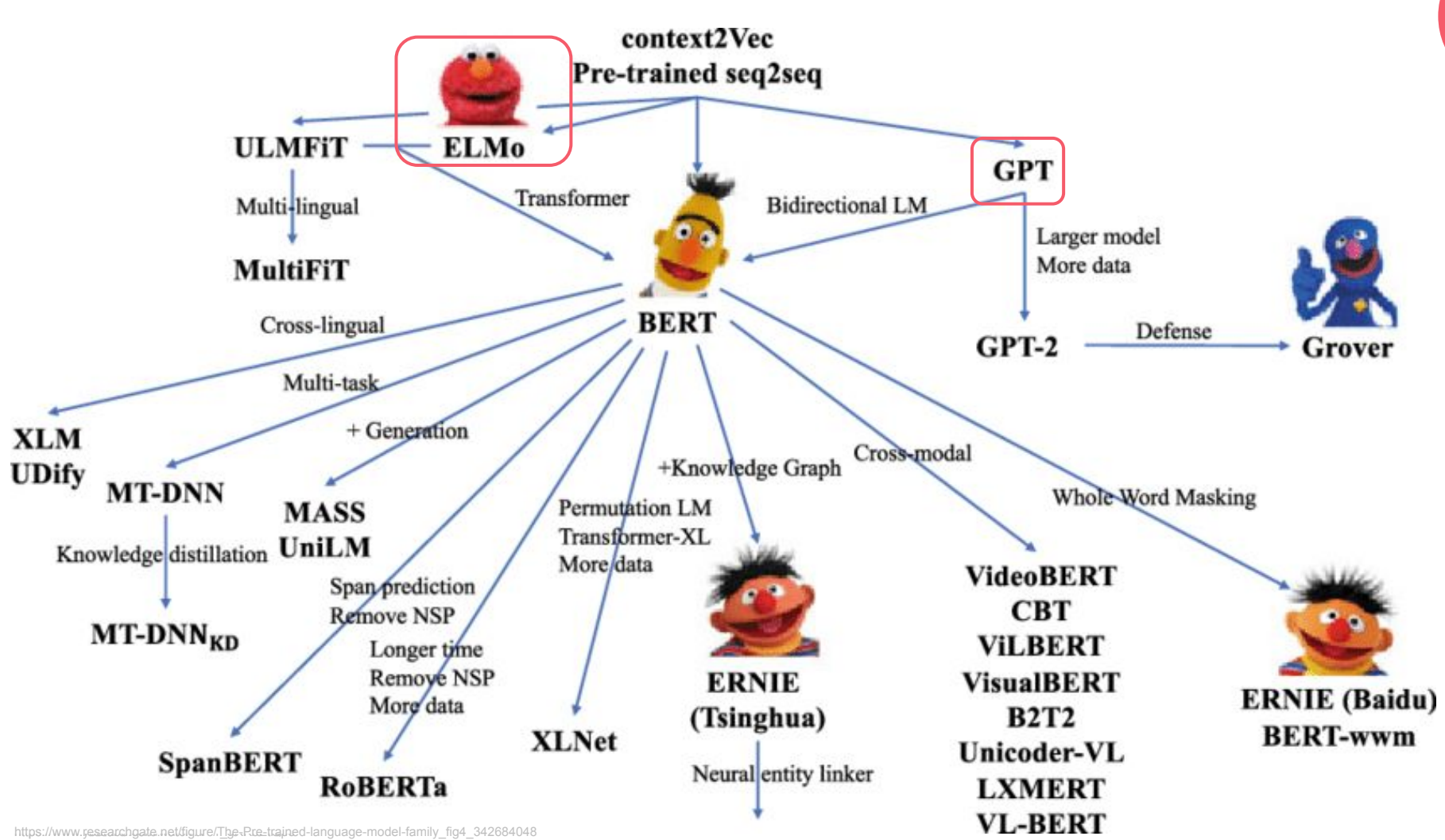
RoBERTa
2019/07

4.1K

RoBERTa: A Robustly Optimized
BERT Pretraining Approach

Source

- **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**
  - https://arxiv.org/pdf/1810.04805.pdf

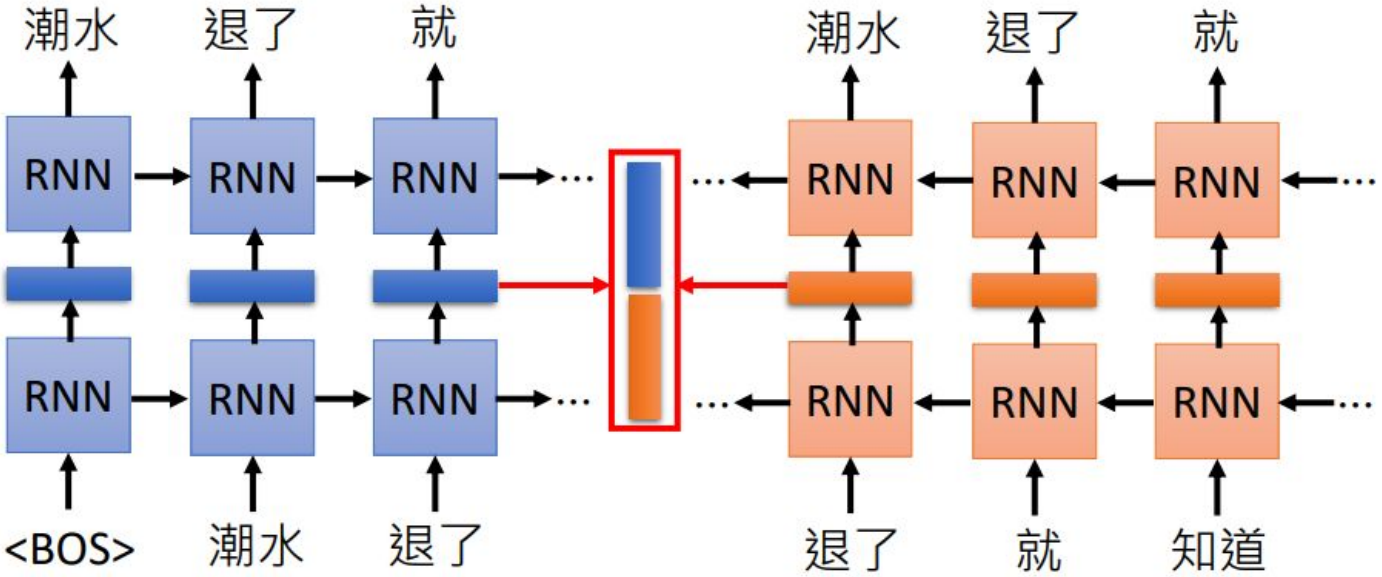


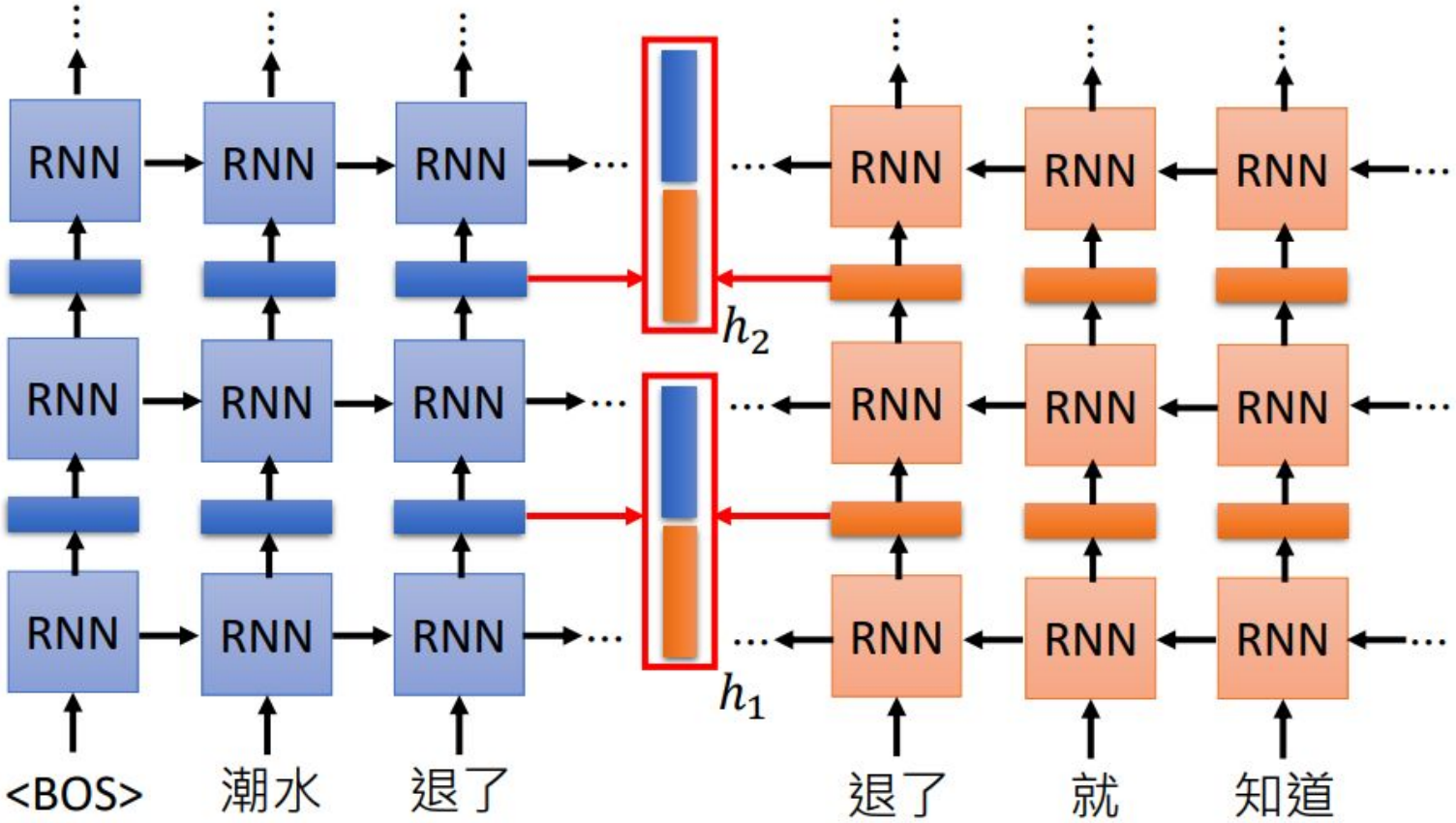
Embeddings from Language Model (ELMO)

https://arxiv.org/abs/1802.05365

- RNN-based language models (trained from lots of sentences)

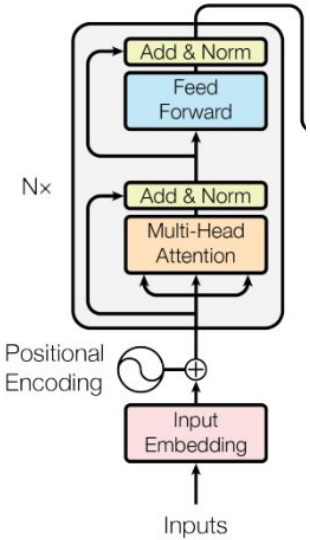
e.g. given “潮水 退了 就 知道 誰 沒穿 褲子”





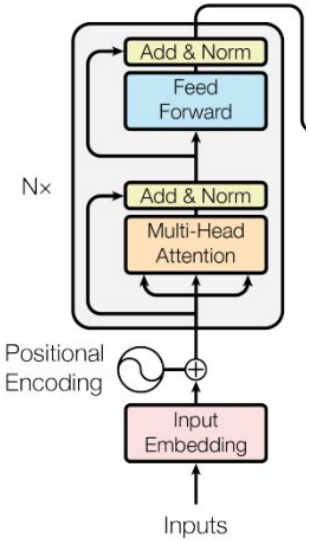
BERT

- **Problem**
  - Standard LM are unidirectional, missing context from both directions
  - RNN-based models hard to parallel
- **Goal**
  - Improve the fine-tuning based approaches(e.g. GPT)
- **Approach**
  - Semi-supervised
    - unsupervised pre-training
    - supervised fine-tuning
  - Transformer encoder architecture



Model Architecture

- **Multi-layer bidirectional Transformer encoder**
- **$\text{BERT}_{\text{BASE}}$**
  - L=12, H=768, A=12, Total Parameters=110M
- **$\text{BERT}_{\text{LARGE}}$**
  - L=24, H=1024, A=16, Total Parameters=340M



Framework

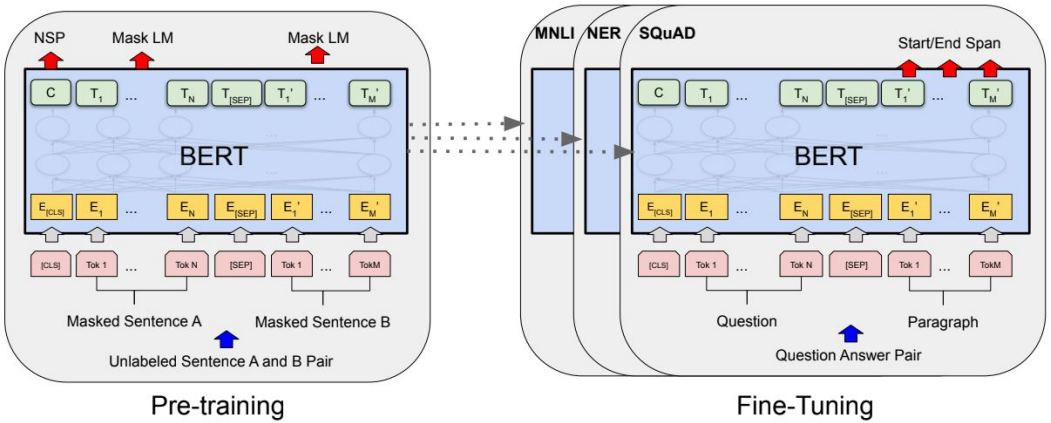
1. Unsupervised Pre-Training

a. Model is trained on unlabeled data over different pre-training tasks

2. Supervised Fine-Tuning

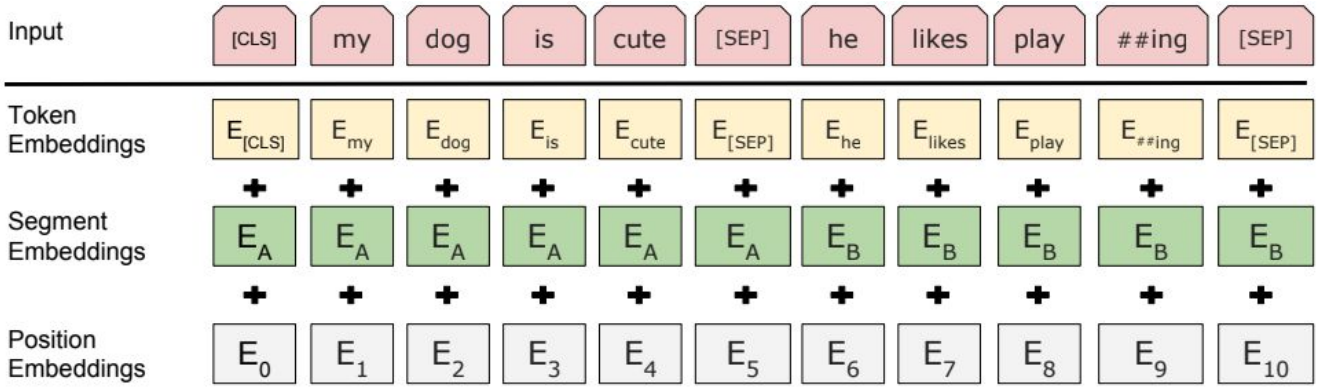
a. Initialized with the pre-trained parameters

b. All of the parameters are fine-tuned using labeled data from the downstream tasks



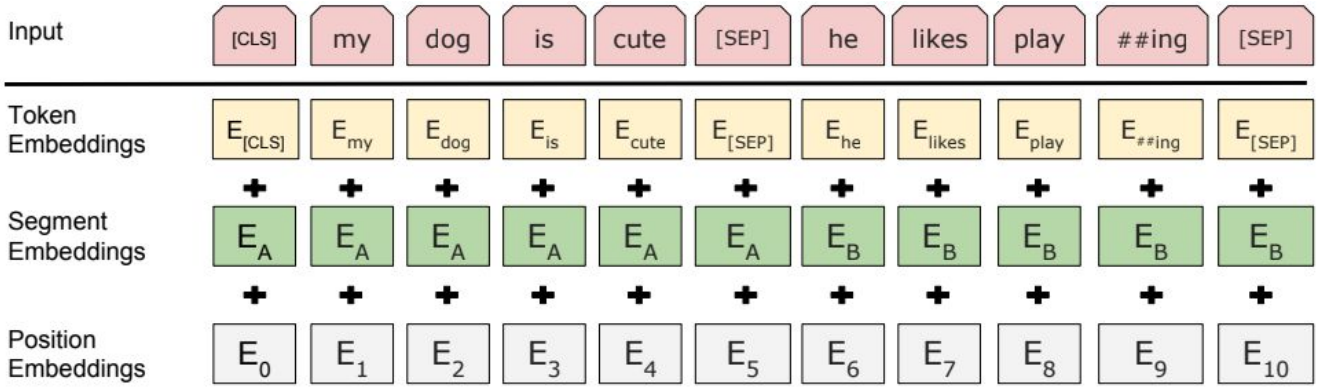
Input

- **Sequence**
  - single sentence
  - a pair of sentences(e.g.〈 Question, Answeri〉)
- **Special token**
  - [CLS] **Classification token** at the first of every sequence
  - [SEP] **Separate token** to separate sentence A and sentence B



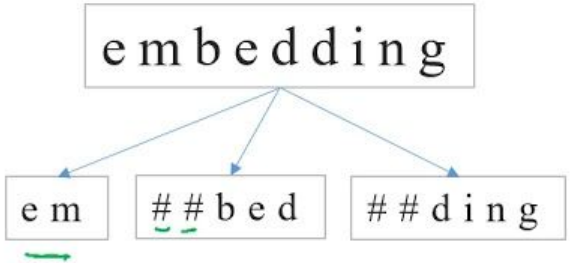
Input

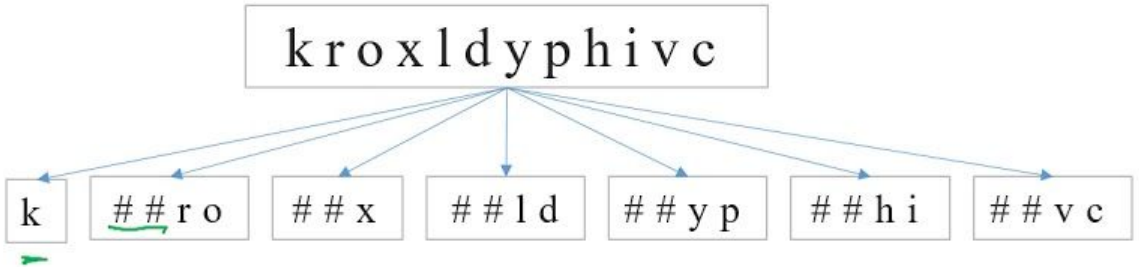
- **Sum token, segment, position embeddings**
  - Token Embedding
    - WordPiece embeddings with a 30,000 token vocabulary
  - Segment Embedding
    - Learned embeddings belong to sentence A or sentence B
  - Position Embedding
    - Learned positional embedding



WordPiece embeddings

All subwords start with “##” …





Framework

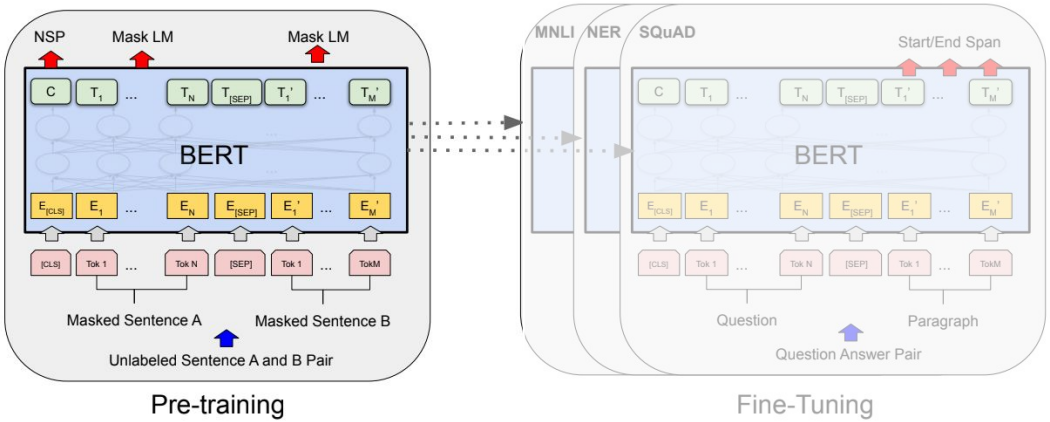
1. Unsupervised Pre-Training

a. Model is trained on unlabeled data over different pre-training tasks

2. Supervised Fine-Tuning

a. Initialized with the pre-trained parameters

b. All of the parameters are fine-tuned using labeled data from the downstream tasks



1. Unsupervised Pre-Training

- **Corpus**
  - BooksCorpus、English Wikipedia
- **Task #1: Masked LM (MLM)**
- **Task #2: Next sentence prediction (NSP)**

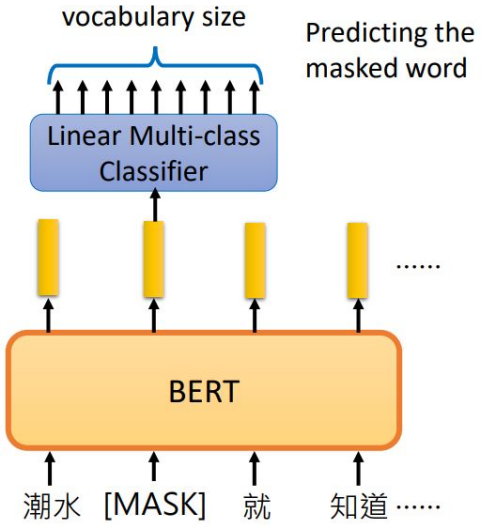
#1 Masked Language Models (MLM)

- **Train a deep bidirectional representation**
- **Mask 15% of all WordPiece tokens in each sequence at random for prediction**
- **[MASK] token does not appear during fine-tuning**
  - Replace the token with
    - [MASK] token 80% of the time

      my dog is hairy → my dog is [MASK]
    - a random token 10% of the time

      my dog is hairy → my dog is apple
    - the unchanged i-th token 10% of the time

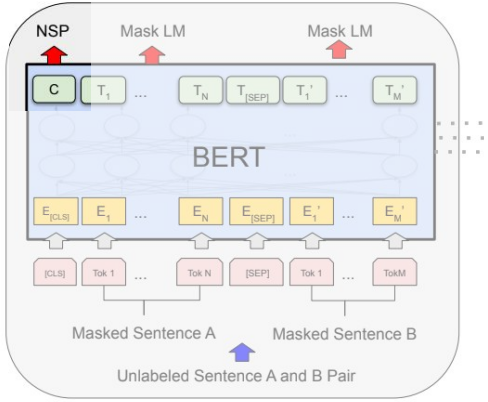
      my dog is hairy → my dog is hairy



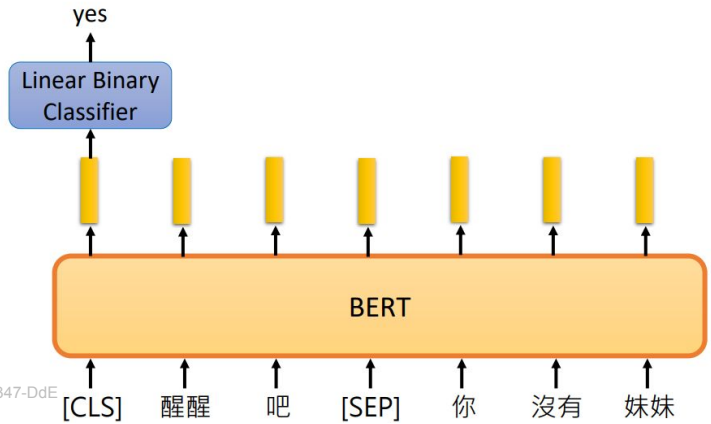
https://www.youtube.com/watch?v=UYPa347-DdE

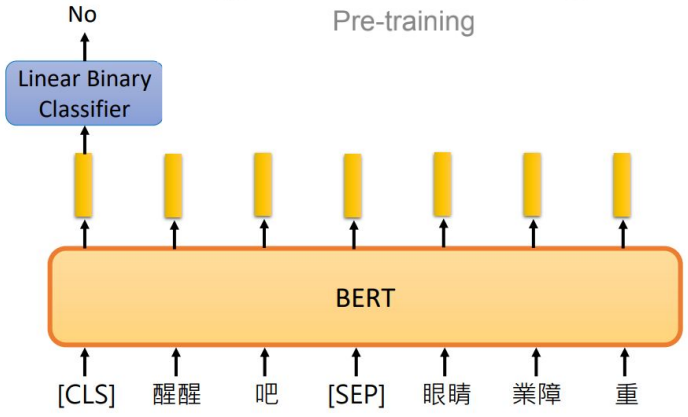
#2 Next sentence prediction(NSP)

- **Understands sentence relationships**
- **50% of the time *IsNext***
  - B is the actual next sentence that follows A
- **50% of the time *NotNext***
  - a random sentence from the corpus



Pre-training





Framework

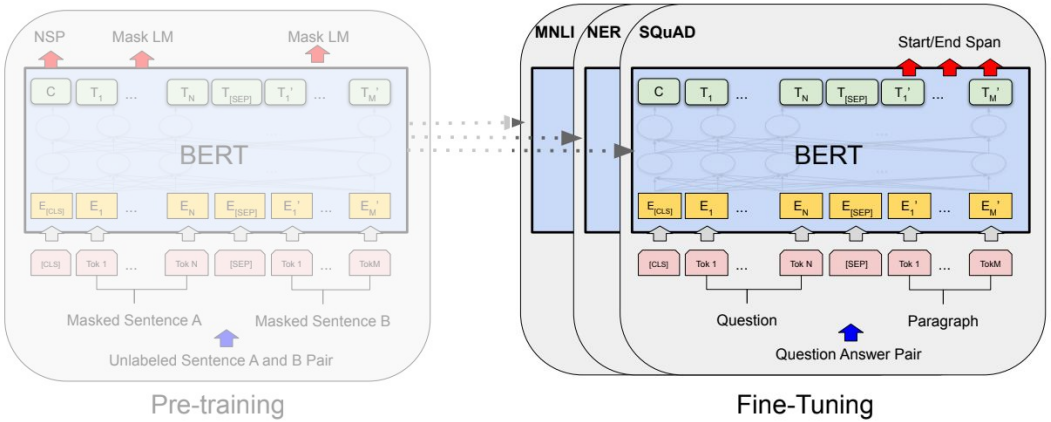
1. Unsupervised Pre-Training

a. Model is trained on unlabeled data over different pre-training tasks

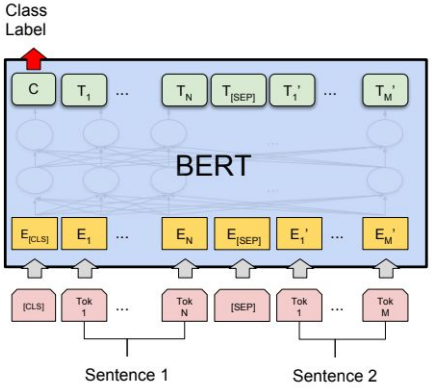
2. Supervised Fine-Tuning

a. Initialized with the pre-trained parameters

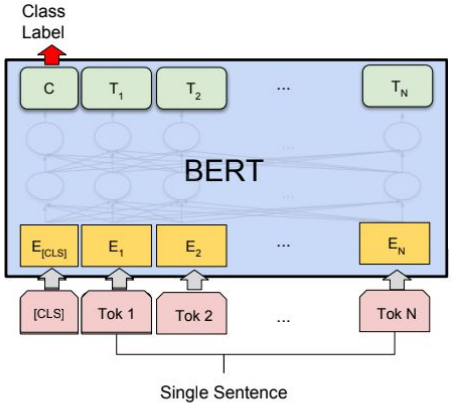
b. All of the parameters are fine-tuned using labeled data from the downstream tasks



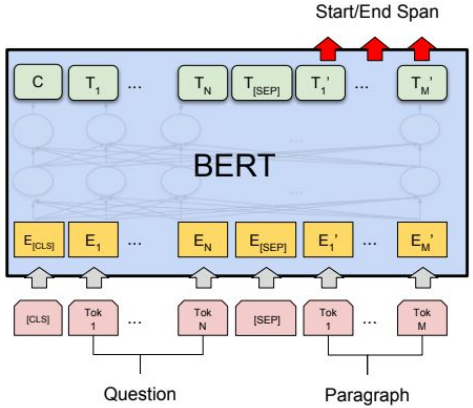
2. Supervised Fine-Tuning



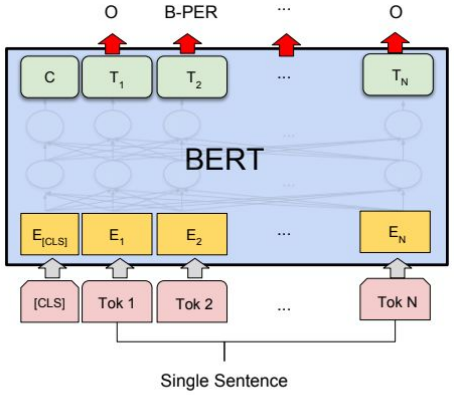
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG



(b) Single Sentence Classification Tasks: SST-2, CoLA



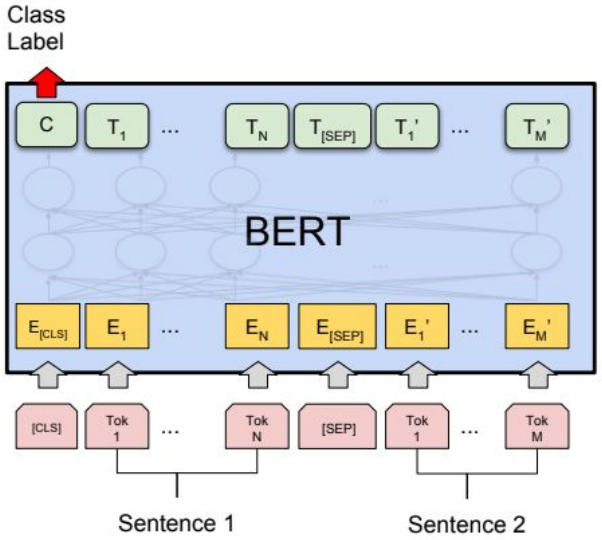
(c) Question Answering Tasks: SQuAD v1.1



(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

(a)Sentence Pair Classification Tasks

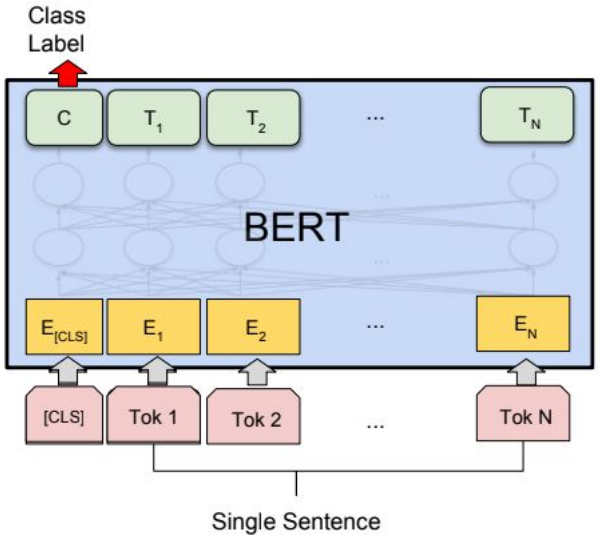
- **Input**
  - two sentences
- **Output**
  - class
- **Example**
  - Natural Language Inference



(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG

(b)Single Sentence Classification Tasks

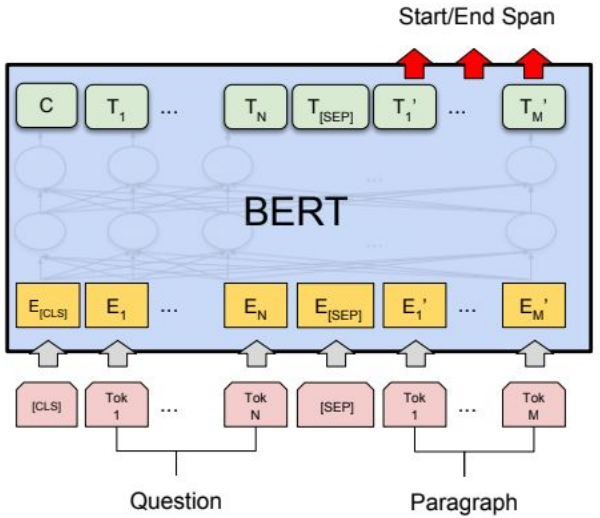
- **Input**
  - single sentence
- **Output**
  - class
- **Example**
  - Sentiment analysis
  - Document Classification



(b) Single Sentence Classification Tasks: SST-2, CoLA

(c)Question Answering Tasks

- **Input**
  - two sentences
- **Output**
  - two integers (s, e)
- **Example**
  - QA



(c) Question Answering Tasks: SQuAD v1.1

(c)Question Answering Tasks

In meteorology, precipitation is any product of the condensation of 17 spheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain 77 atte 79 cations are called "showers".

What causes precipitation to fall?
gravity
$s = 17, e = 17$

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

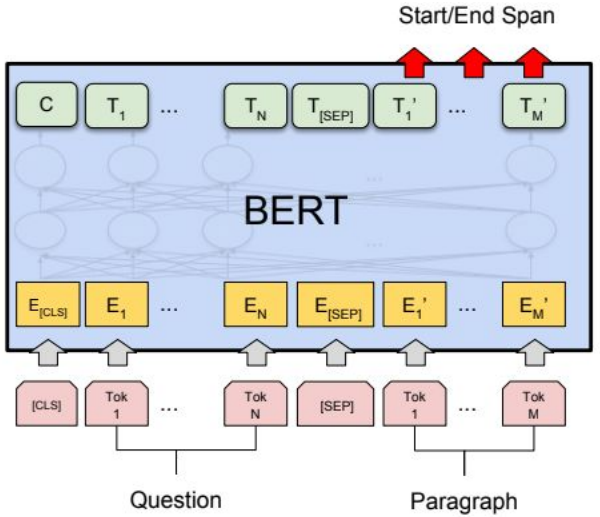
Where do water droplets collide with ice crystals to form precipitation?
within a cloud
$s = 77, e = 79$

Document: $D = \{d_1, d_2, \cdots, d_N\}$

Query: $Q = \{q_1, q_2, \cdots, q_N\}$

output: two integers (s, e)

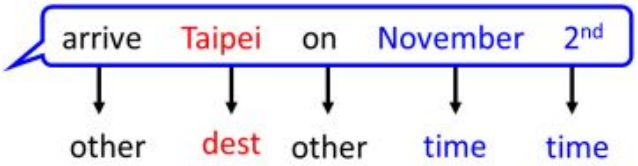
Answer: $A = \{q_s, \cdots, q_e\}$

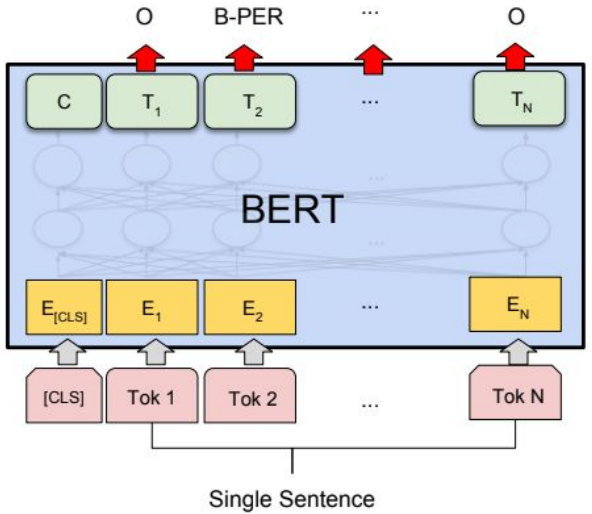


(c) Question Answering Tasks: SQuAD v1.1

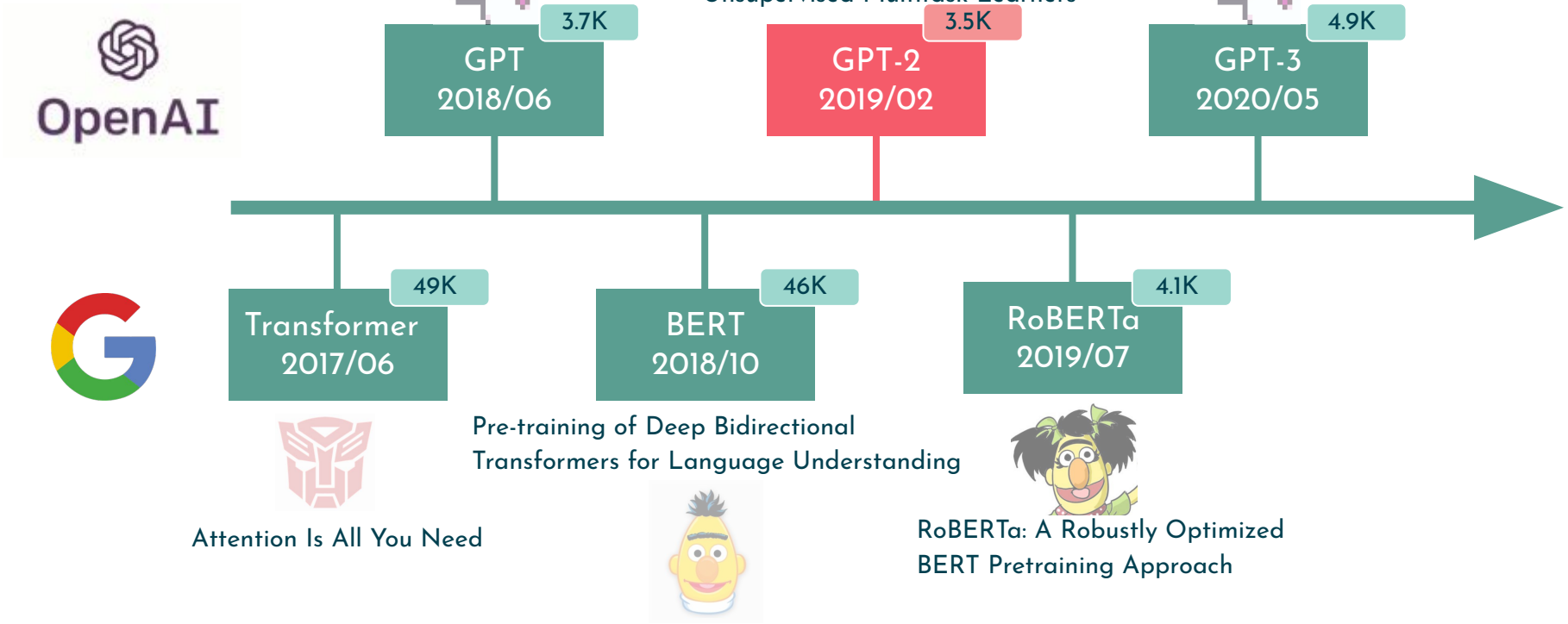
(d)Single Sentence Tagging Tasks

- **Input**
  - single sentence
- **Output**
  - class of each word
- **Example**
  - Slot filling





(d) Single Sentence Tagging Tasks: CoNLL-2003 NER



https://www.youtube.com/watch?v=nzqIFIcCSWQ

BERT GLUE Test results

- GLUE
  - A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding

System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	**Average** -
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**92.7**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**82.1**

(additional Info.)GLUE Tasks

Corpus	|Train|	|Test|	Task	Metrics	Domain
			Single-Sentence Tasks		
CoLA	8.5k	**1k**	acceptability	Matthews corr.	misc.
SST-2	67k	1.8k	sentiment	acc.	movie reviews
			Similarity and Paraphrase Tasks		
MRPC	3.7k	1.7k	paraphrase	acc./F1	news
STS-B	7k	1.4k	sentence similarity	Pearson/Spearman corr.	misc.
QQP	364k	**391k**	paraphrase	acc./F1	social QA questions
			Inference Tasks		
MNLI	393k	**20k**	NLI	matched acc./mismatched acc.	misc.
QNLI	105k	5.4k	QA/NLI	acc.	Wikipedia
RTE	2.5k	3k	NLI	acc.	news, Wikipedia
WNLI	634	**146**	coreference/NLI	acc.	fiction books

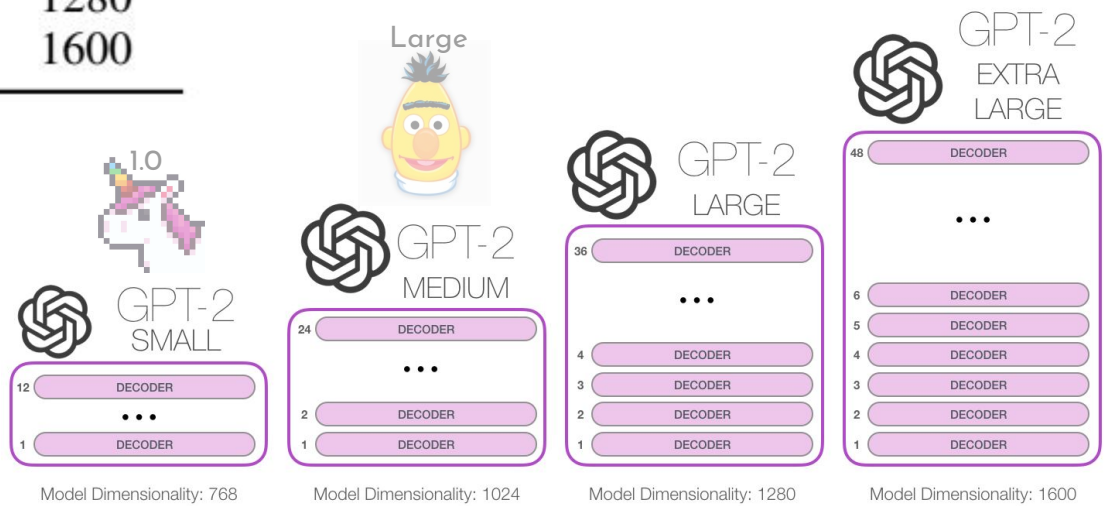
Table 1: Task descriptions and statistics. All tasks are single sentence or sentence pair classification, except STS-B, which is a regression task. MNLI has three classes; all other classification tasks have two. Test sets shown in bold use labels that have never been made public in any form.

GPT -2 Characteristic

- **Training on a million dataset, WebText**
  - 40 GB of text
- **More parameters and layers than ever**
- **Perform down-stream tasks in a zero-shot setting, without any parameter or archi-tecture modification**
  - Reading Comprehension
  - Summarization
  - Translation

Model sizes

Parameters	Layers	d_{model}
117M	12	768
345M	24	1024
762M	36	1280
1542M	48	1600



https://jalammar.github.io/illustrated-gpt2/

Zero-shot results on many datasets

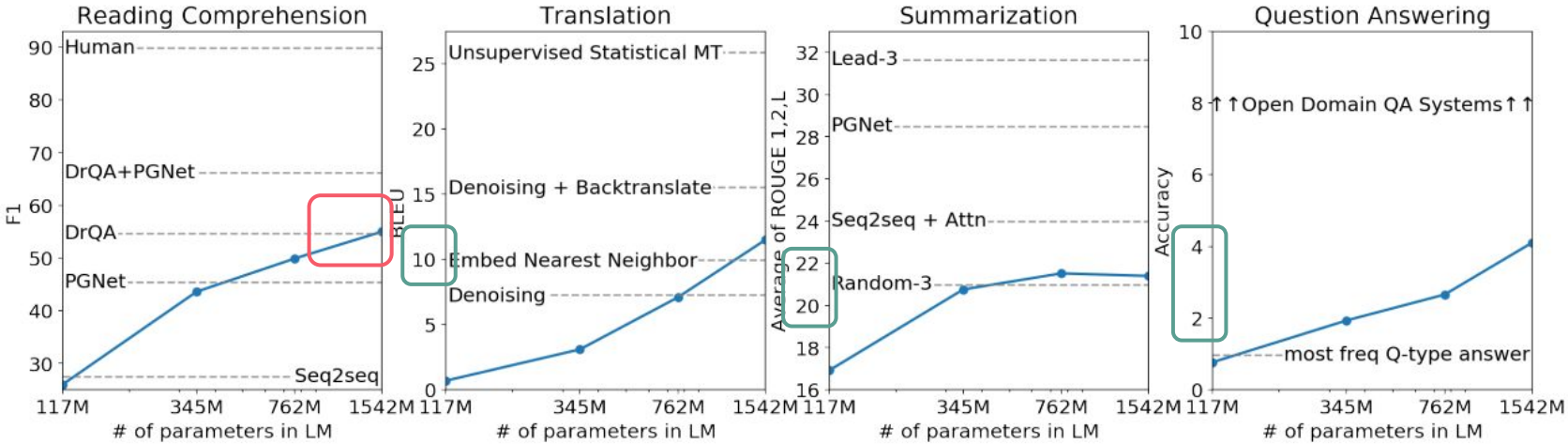
	LAMBADA (PPL)	LAMBADA (ACC)	CBT-CN (ACC)	CBT-NE (ACC)	WikiText2 (PPL)	PTB (PPL)	enwik8 (BPB)	text8 (BPC)	WikiText103 (PPL)	1BW (PPL)
SOTA	99.8	59.23	85.7	82.3	39.14	46.54	0.99	1.08	18.3	**21.8**
117M	**35.13**	45.99	**87.65**	**83.4**	**29.41**	65.85	1.16	1.17	37.50	75.20
345M	**15.60**	55.48	**92.35**	**87.1**	**22.76**	47.33	1.01	**1.06**	26.37	55.72
762M	**10.87**	**60.12**	**93.45**	**88.0**	**19.93**	**40.31**	**0.97**	**1.02**	22.05	44.575
1542M	**8.63**	**63.24**	**93.30**	**89.05**	**18.34**	**35.76**	**0.93**	**0.98**	**17.48**	42.16

SOTA state-of-the-art

(additional Info.)Datasets task

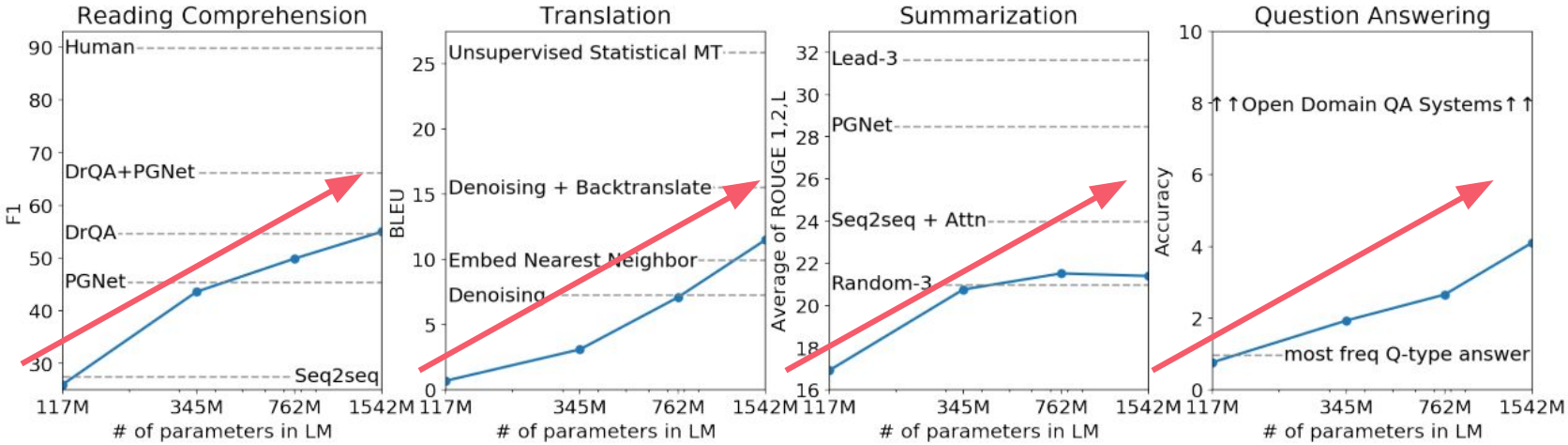
- **Evaluation of models for sequence labelling.**
  - Penn Treebank (PTB)
- **Measure long-term dependencies**
  - LAMBADA
  - WikiText-2
  - WikiText-103
- **Measure how well language models can exploit wider linguistic context**
  - CBT
  - NE - answers to the questions are named entities
  - CN - answers to the questions are common nouns
- **Measure the model's ability to compress data**
  - enWik8
  - Text8
- **Measuring progress in statistical language modeling**
  - 1BW

Zero-shot task performance on NLP tasks

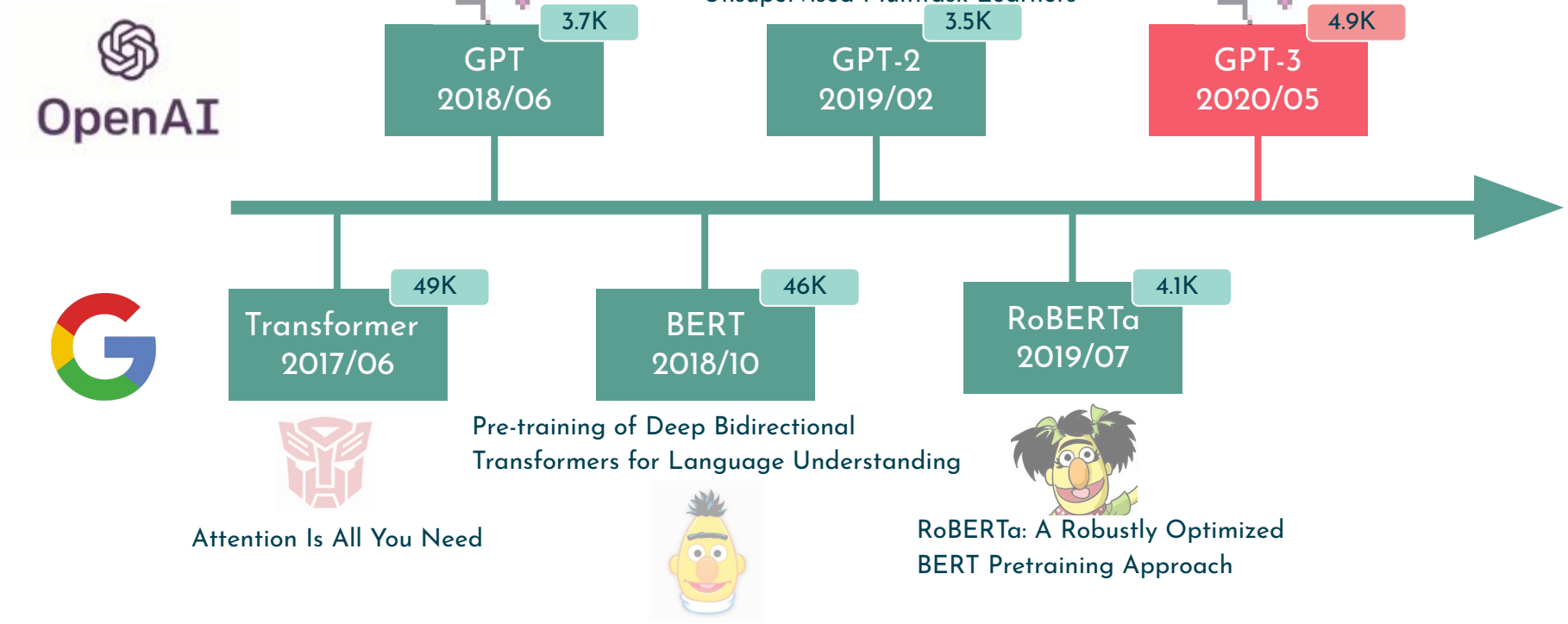


- **Reading Comprehension**
  - $d1, d2, \cdots, dN$,"Q:", $q1, q2, \cdots, qN$,"A:"
- **Summarization**
  - $d1, d2, \cdots, dN$,"TL;DR:"
- **Translation**
  - English sentence 1 = French sentence 1
  - English sentence 2 = French sentence 2
  - English sentence 3 = ?

Zero-shot task performance on NLP tasks



- **Reading Comprehension**
  - $d1, d2, \cdots, dN$,"Q:", $q1, q2, \cdots, qN$,"A:"
- **Summarization**
  - $d1, d2, \cdots, dN$,"TL;DR:"
- **Translation**
  - English sentence 1 = French sentence 1
  - English sentence 2 = French sentence 2
  - English sentence 3 = ?



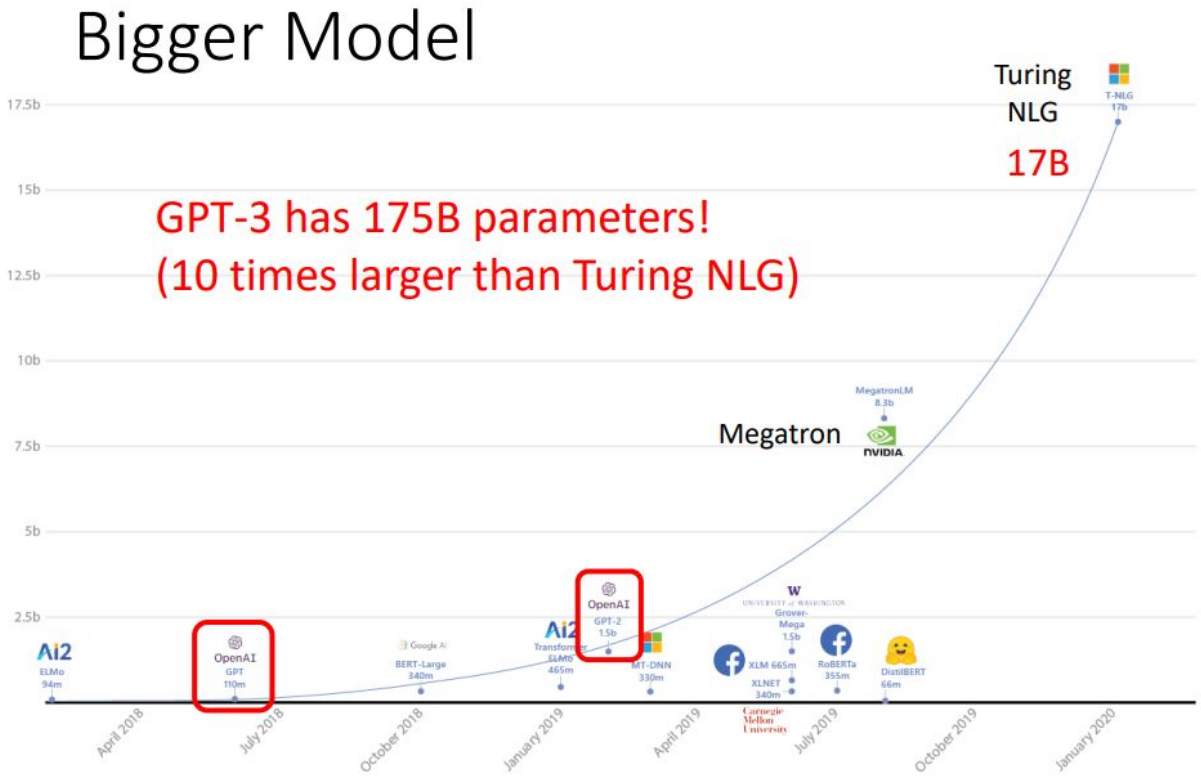
https://www.youtube.com/watch?v=nzqIFIcCSWQ

GPT-3 Characteristic

- **175 billion parameters**
- **Applied without any gradient updates or fine-tuning**
- **In-context learning**

Model size

https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/



(additional Info.) Model size

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or "GPT-3"	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

Parameters	Layers	d_{model}
117M	12	768
345M	24	1024
762M	36	1280
1542M	48	1600

GPT-3 Characteristic

- 175 billion parameters
- Applied without any gradient updates or fine-tuning
- **In-context learning**
  - Using the text input of a pretrained language model as a form of task specification
    - the model is conditioned on a natural language instruction and/or a few demonstrations of the task
    - and is then expected to complete further instances of the task simply by predicting what comes next

In-context learning

The three settings we explore for in-context learning

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  cheese =>                           ← prompt
```

One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← example
3  cheese =>                           ← prompt
```

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

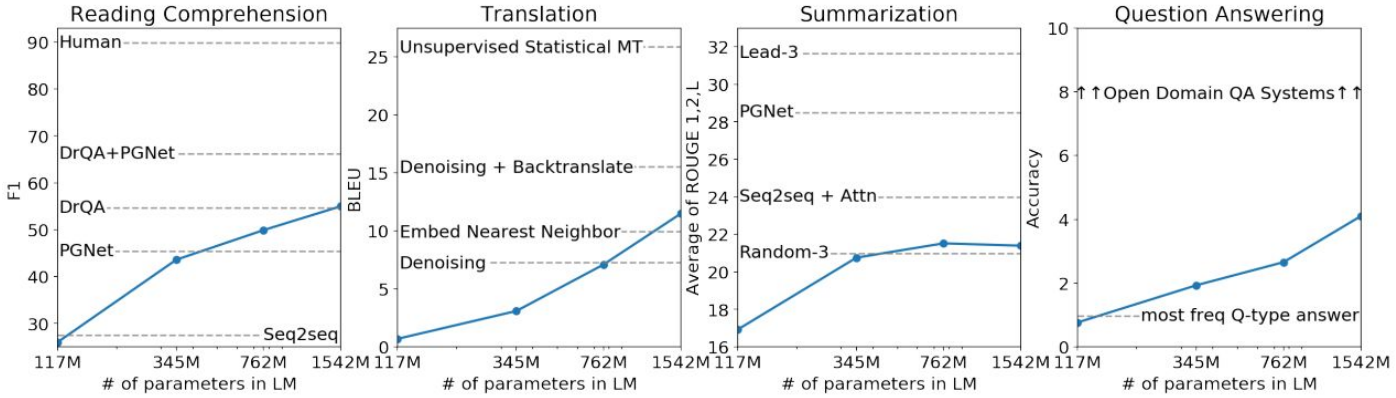
```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← examples
3  peppermint => menthe poivrée        ←
4  plush girafe => girafe peluche      ←
5  cheese =>                           ← prompt
```

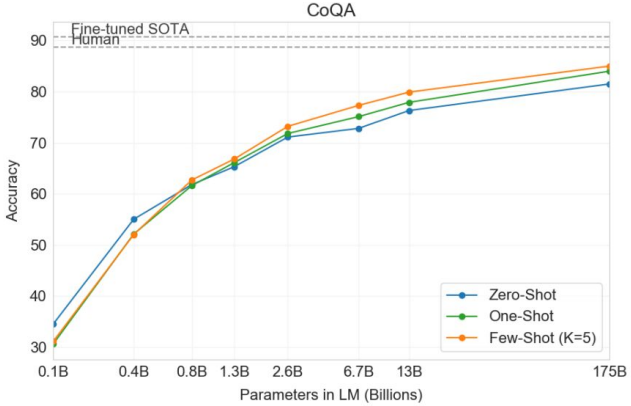
(additional Info.) Training dataset

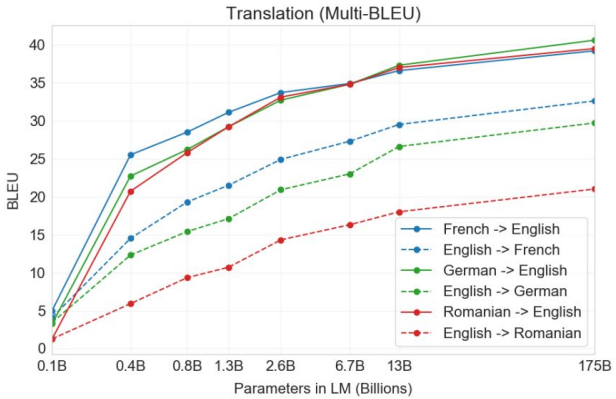
- **Common Crawl dataset**
  a. constituting nearly a trillion words.
  b. low-quality
- **Data clean**
  a. Filtered a version of CommonCrawl based on similarity to a range of high-quality reference corpora(WebText)
  b. Deduplication at the document level
  c. Added known high-quality reference corpora

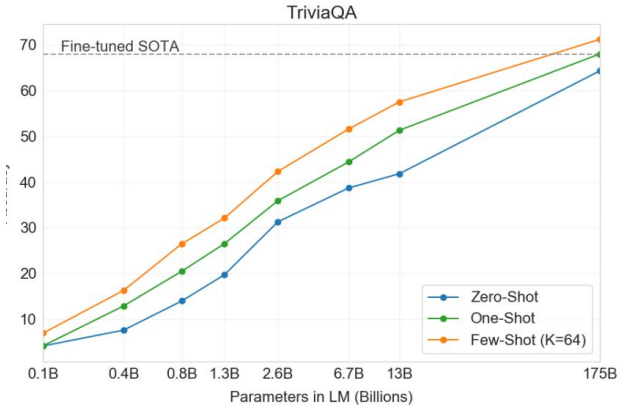
Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

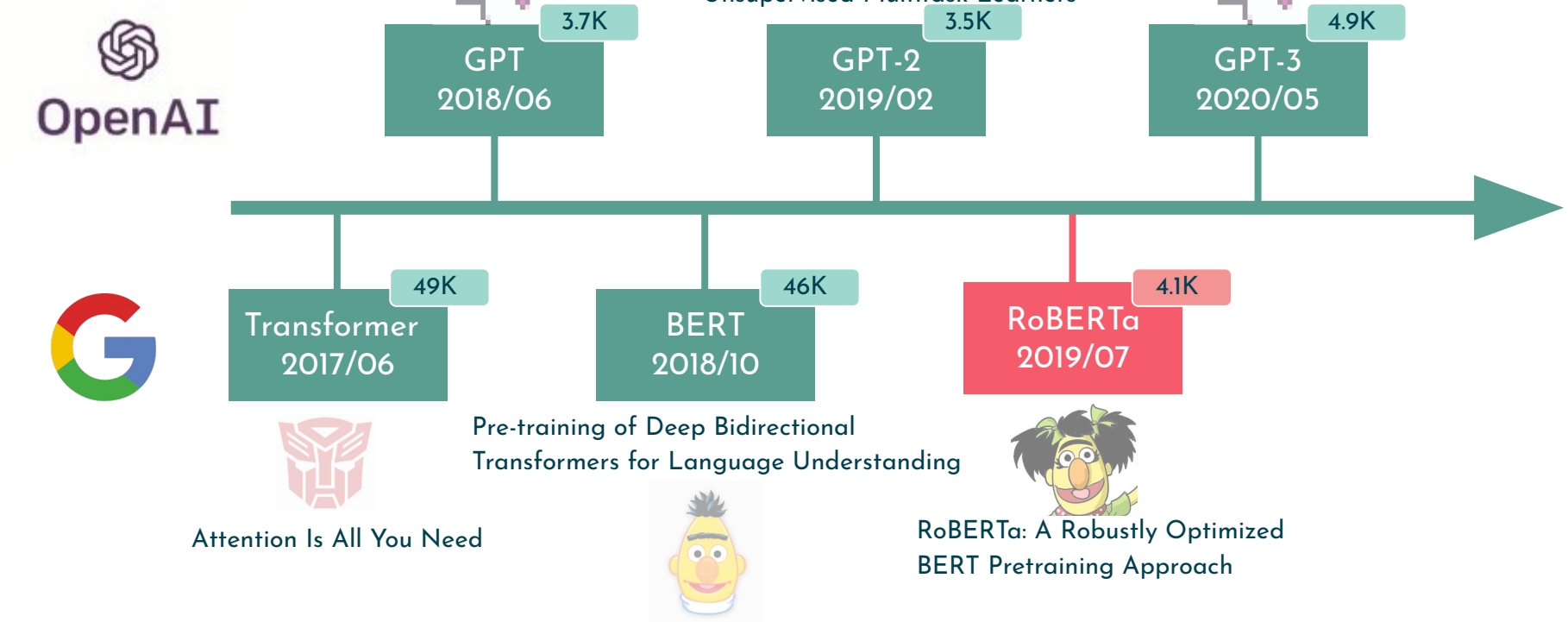
tasks performance on NLP tasks











https://www.youtube.com/watch?v=nzqIFIcCSWQ

Source

- **RoBERTa: A Robustly Optimized BERT Pretraining Approach**
  - https://arxiv.org/pdf/1907.11692.pdf

RoBERTa

- **Problem**
  - BERT was significantly undertrained
- **Modifications**
  - Training the model longer, with bigger batches,over more data
  - Training on longer sequences
  - Removing the next sentence prediction(NSP) objective
  - Dynamically changing the masking pattern applied to the training data
- **Large new dataset (CC-NEWS)**
  - Comparable size to other privately used datasets, to better control for training set size effects

Modifications

1. **Training the model longer, with bigger batches,over more data**
   - BPE vocabulary of size =50K (Wordpiece size =30K)
   - minibatches containing B=8K (B = 256)
   - CC-NEWS(160GB) (16GB)
2. **Training on longer sequences**
   - sequences of maximum length T=512 tokens (≤ 512 tokens)

Modifications

3. Removing the next sentence prediction(NSP) objective

- FULL-SENTENCES
  - Each input is packed with full sentences sampled contiguously from one or more documents
  - Total length is at most 512 tokens
- DOC-SENTENCES
  - May not cross document boundaries
  - May be shorter than 512 tokens

Model	SQuAD 1.1/2.0	MNLI-m	SST-2	RACE
Our reimplementation (without NSP loss):				
FULL-SENTENCES	90.4/79.1	84.7	92.5	64.8
DOC-SENTENCES	90.6/79.7	84.7	92.7	65.6
$\text{BERT}_{\text{BASE}}$	88.5/76.3	84.3	92.8	64.3
$\text{XLNet}_{\text{BASE}}$ (K = 7)	–/81.3	85.8	92.7	66.1
$\text{XLNet}_{\text{BASE}}$ (K = 6)	–/81.0	85.6	93.4	66.7

Modifications

- **SEGMENT-PAIR+NSP**
  - input has a pair of segments, which can each contain multiple natural sentence
- **SENTENCE-PAIR+NSP**
  - Each input contains a pair of natural sentences,
- **Conclusion**
  - using individual sentences hurts performance on downstream tasks

Model	SQuAD 1.1/2.0	MNLI-m	SST-2	RACE
Our reimplementation (with NSP loss):				
SEGMENT-PAIR	90.4/78.7	84.0	92.9	64.2
SENTENCE-PAIR	88.7/76.2	82.9	92.1	63.0

Modifications

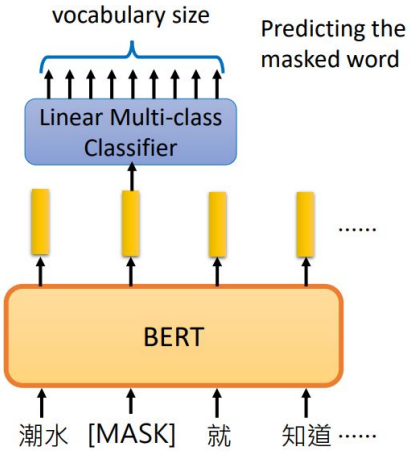
4. Dynamically changing the masking pattern applied to the training data

- Static masking (BERT)
  - masking once during data preprocessing
- Dynamic Masking
  - duplicated 10 times so that each sequence is masked in 10 different ways over the 40 epochs of training

XLnet

$BERT_{BASE}$

Masking	SQuAD 2.0	MNLI-m	SST-2
reference	76.3	84.3	92.8
Our reimplementation:			
static	78.3	84.3	92.5
dynamic	78.7	84.0	92.9



Large new dataset (CC-NEWS)

- **BOOKCORPUS (16G)**
  - Original data used to train BERT
- **CC-NEWS (76G)**
  - Collected from the English portion of the CommonCrawl News dataset
- **OPENWEBTEXT (38G)**
  - Open-source recreation of the WebText (GPT)
- **STORIES (31G)**
  - containing a subset of CommonCrawl data filtered to match the story-like style of Winograd schemas
- **Total 160G**

3

Conclusion

Why PsyQA use these model?

- **GPT - 2**
  - Pretrained Chinese GPT-2 available **does not** train on any corpus related to psychology or mental health support
  - Train a GPT-2 from scratch based on the corpus.
- **RoBERTa**
  - The task requires to assign a strategy label to each sentence in a **long answer**

Picture Source

- **Google**
  - https://seeklogo.com/vector-logo/268022/google-2015-icon
- **Transformer**
  - https://www.kindpng.com/imgv/TJoJmbi_transformers-generations-combiner-wars-leader-logo-transformers-hd/
- **Unicorn**
  - https://play.google.com/store/apps/details?id=com.appcraft.unicorn&hl=zh_TW&gl=CN
- **Bert**
  - https://www.kindpng.com/imgv/iTxobRi_bert-png-bert-png-cartoon-bert-sesame-street/
- **RoBERTa**
  - https://toughpigs.com/afd-mural-art/