Deep Learning for Natural Language Processing
Lecture 4: Pre-trained Language Model

Quan Thanh Tho
Faculty of Computer Science and Engineering
Back Khoa University

Agenda

- GPT-2 and what we have done
- Practical Situation: Huge Data, Few Labeled
- Language Model: General Architecture
- Case Studies
- GPT, Transformer and BERT

SYSTEM PROMPT (HUMAN-WRITTEN)	*In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.*
MODEL COMPLETION (MACHINE-WRITTEN, 10 TRIES)	The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science. Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved. Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

trình chuyên môn ngành công nghệ này anh cho rằng thành phố
ngàn thí sinh thực hiện việc nhận được các loại các chuyên g
học đào tạo nhưng nhiều ngành nghề nghiệp trong thời gian tớ

```
generate_text('một cô gái được các chàng trai yêu thích', 400
```

```
'một cô gái được các chàng trai yêu thích với các con của nhữ
ia sẻ để đánh giá vào ngày N cho thầy thí sinh có thể nhận đư
hút nhiều người từng thi đấu chỉ theo một cách cho thầy khách
trước đó thông tin trong hành trình thi thpt quốc gia cũng đã
```

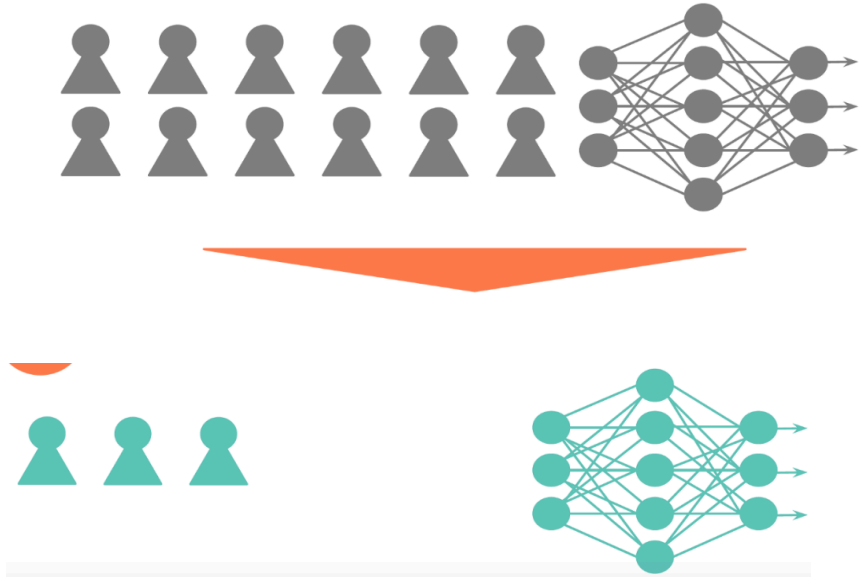
```
generate_text('andy murr', 400)
```

```
'andy murray chi tiết và có thể đánh chú ý cho những bức ảnh m
ng bài học chuyên nghiệp ngoài ra chúng tôi đang phải thực hiệ
bị và hiệu quả mới nhất giờ tay vợt người thái lan thế giới nh
ể tham gia công chúng cho thầy công ty thiết kế ngày N trong t
```

```
In [12]: generate_text('roger federer', 400)

Out[12]: 'roger federer là một trong những người đã có thể thấy được chiến thắng trong cuộc đua vô địch wimbledon N đã có mặ
         t tại trận chung kết giải quần vợt thế giới với tỷ số N N N N N N N trong trận đấu này đã có N lần thi đấu vào tứ k
         ết us open N trước khi giành chiến thắng N N N N N trước tay vợt người serbia vừa chiếm tỷ lệ N N N N N trong khi đ
         ó trận đấu ra sân của mình và đối thủ của federer đã có N lần gặp nha'

In [22]: generate_text('novak djo', 400)

Out[22]: 'novak djokovic và chuyển sang thi đấu trên sân cỏ những cú đánh thuận tay khi trở lại thi đấu chuyên nghiệp và đó
         là một chấn thương của nadal sau khi công bố người thắng cuộc N trước khi bước vào set N với tỷ số N N N N N N N N
         N N N N N N N N sam querrey và colombia N N N N N marin cilic N N N N N N N N N N N N N N N N N N N N N N N N N N N
         N N N N N N N N N N N N N N N N N N N N N N N N N N N N N N N N '

In [27]: generate_text('andy murr', 400)

Out[27]: 'andy murray đã có mặt tại các trận đấu như vậy thì chỉ cần thi đấu không thể trở lại trong năm nay khi thi đấu chỉ
         có thể trở lại thì mới có thể nhận thấy thực sự cho thấy sự cố gắng thi đấu của mình trong sự nghiệp của mình trước
         khi chia tay sau N set đấu này cũng có thể sẽ nhận được sự chú ý của giới chuyên môn và cho thấy sự thay đổi của cá
         c tay vợt trẻ như anh là chuyện thi đấu của mình trong khi đó a'

In [28]: generate_text('bitco', 400)

Out[28]: 'bitcoin theo đó việc chuyển đổi số thuế thu nhập của tập đoàn là N triệu usd trong tháng N theo báo cáo của uber c
         ho biết thêm các tổ chức tín dụng của hệ thống tài chính tại việt nam cho thấy trong tháng N cho thấy sự thay đổi c
         ủa thị trường trong nước và quốc tế và chính phủ trung quốc có thể thấy rằng những thách thức trong các nền kinh tế
         apec cho thấy thị trường chứng khoán việt nam có thể sẽ tăng'

In [29]: generate_text('apple', 400)

Out[29]: 'apple cho biết chúng tôi cần phải có thêm một số trường hợp thu thập thông tin cho người tiêu dùng cho các nhà mạn
         g này như thế nào các chuyên gia nhận định đây chính là một trong những nguyên nhân chính trong thị trường những ng
         ày qua và đó là một chiếc điện thoại thông minh của mình và các nhà phân tích cho rằng chúng ta có thể giúp các nhà
         cung cấp dịch vụ công nghệ thông tin trên thế giới và chúng '

In [40]: generate_text('micros', 400)

Out[40]: 'microsoft theo chuyên gia giáo dục việt nam cho biết họ đã chọn thời gian tham gia các hoạt động của các nhà trườn
         g trong trường học thuộc các ngành nghề theo học bạn có thể tham gia các hoạt động thi thpt quốc gia theo thời gian
         thực tế của nhà trường và chủ nhân của trường đại học tân tạo cho biết thêm vào đó những người làm công tác tuyển c
         họn các điều kiện thuận lợi cho các em học sinh tham gia các '
```

Introduction

Common situation at “big-data-based companies”
Cheap large dataset vs. expensive labeled data

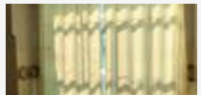


Bất động sản rao bán TP.HCM

Xem thêm >

Nơi để bạn phát triển tuong lai, Nhà giá 883
Thạnh Lộc 41, Phường Thạnh Lộc, Quận 12, TP.HCM
882 triệu 999 nghìn

Duy nhất chỉ có 1 căn, có thể chỉ dành cho bạn,
Đường TX 25, Phường Thạnh Xuân, Quận 12, TP.HCM
880 triệu



Đừng mua nhà quận 3 nếu chưa xem tin này -
Nguyễn Thượng Hiền, Phường 4, Quận 3, TP.HCM
ỷ 100 triệu

Bán nhà mặt phố Kim Mã, vị trí đẹp nhất phố (đang cho thuê 180 tr/th), 80m2, 8T, MT 4,5m, giá 33 tỷ

Giá : 33 tỷ
Diện tích : 80 m²
Quận/huyện : Ba Đình, Hà Nội 16/11/2018

Tôi cần bán CH Park 1 - Times City 458 Minh Khai, HN, 120m2, 3PN, căn góc, thiết kế đẹp, 35 tr/m2

Giá : 4.2 tỷ
Diện tích : 120 m²
Quận/huyện : Hai Bà Trưng, Hà Nội 16/11/2018

Asiana Capella Trần Văn Kiều, Q6, chỉ 1.6 tỷ sỡ hữu ngay CH, duplex, smark office. LH: 0901 090 228

Giá : 1.6 tỷ
Diện tích : 50 m²
Quận/huyện : Quận 6, Hồ Chí Minh 16/11/2018

Bất động sản rao bán TP.HCM

Xem thêm >

Nơi để bạn phát triển tuong lai, Nhà giá 883
Thạnh Lộc 41, Phường Thạnh Lộc, Quận 12, TP.HCM

Xin chào, bạn cần tìm bất động sản như thế nào?

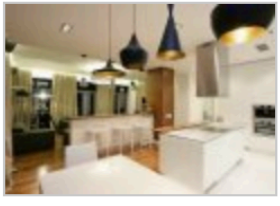
à tôi muốn mua một căn nhà mặt tiền ở Quận 10 để kinh doanh

Đây là 5 kết quả tìm được theo yêu cầu của bạn

Cần bán nhà mặt tiền đường Sư Vạn Hạnh, Phường 13, Quận 10. - Diện tích: 4m x 27m, TDT: 115m2. vuông vức, không lộ giới, lề rộng 7m, sổ hồng chính chủ. - Kết cấu: 1 trệt, 2 lầu. - Vị trí rất đắc địa gần ngay và 3 Tháng 2, gần khu Trung tâm thương mại An Phong đang xây dựng.khu kinh doanh buôn bán đa ngành nghề sầm uất. Tập trung nhiều trung tâm

CÓ | KHÔNG | Tư vấn

Bán nhà mặt phố
tr/th), 80m2, 8T,
Giá :
Diện tích :
Quận/huyện :



Tôi cần bán CH Pa
3PN, căn góc, thiế
Giá :
Diện tích :
Quận/huyện :

Asiana Capella Trà
duplex, smark off
Giá : 1.6 tỷ
Diện tích : 50 m²
Quận/huyện : Quận 6, Hồ Chí Minh 16/11/2018

Cần tiền bán gấp đất mặt tiền đường tam bửu tự - xã đa phước - bình chánh - tp . hcm
- diện tích : 2 , 000 m 2 (hai sổ đỏ , mỗi sổ là 1 , 000 m 2)
- đất trồng cây lâu năm , thích hợp làm nhà vườn , ao cá , trồng cây cảnh .
- đường bê tông xe hơi vào tận nơi ,
- đất cách đường quốc lộ 50 khoảng 200 m
- đất nằm gần đường dẫn cao tốc long thành dầu dây .
- cách bến xe quận 8 khoảng 6 km .

Predict

addr_street (S) | addr_district (D) | addr_city (C) | addr_ward (W) | position (P) | area (A) | price (M) | transaction_type (T) | realestate_type (E)

legal (L) | potential (O) | surrounding (U) | surrounding_characteristics (I) | surrounding_name (N) | interior_floor (F) | interior_room (R) | orientation (H)

project (J) | normal ( ) | Reset

Cần tiền bán gấp đất mặt tiền đường tam bửu tự - xã đa phước - bình chánh - tp . hcm
- diện tích : 2 , 000 m 2 (hai sổ đỏ , mỗi sổ là 1 , 000 m 2)
- đất trồng cây lâu năm , thích hợp làm nhà vườn , ao cá , trồng cây cảnh .
- đường bê tông xe hơi vào tận nơi ,
- đất cách đường quốc lộ 50 khoảng 200 m
- đất nằm gần đường dẫn cao tốc long thành dầu dây .
- cách bến xe quận 8 khoảng 6 km .
- cách khu dân cư phong phú khoảng 3 km

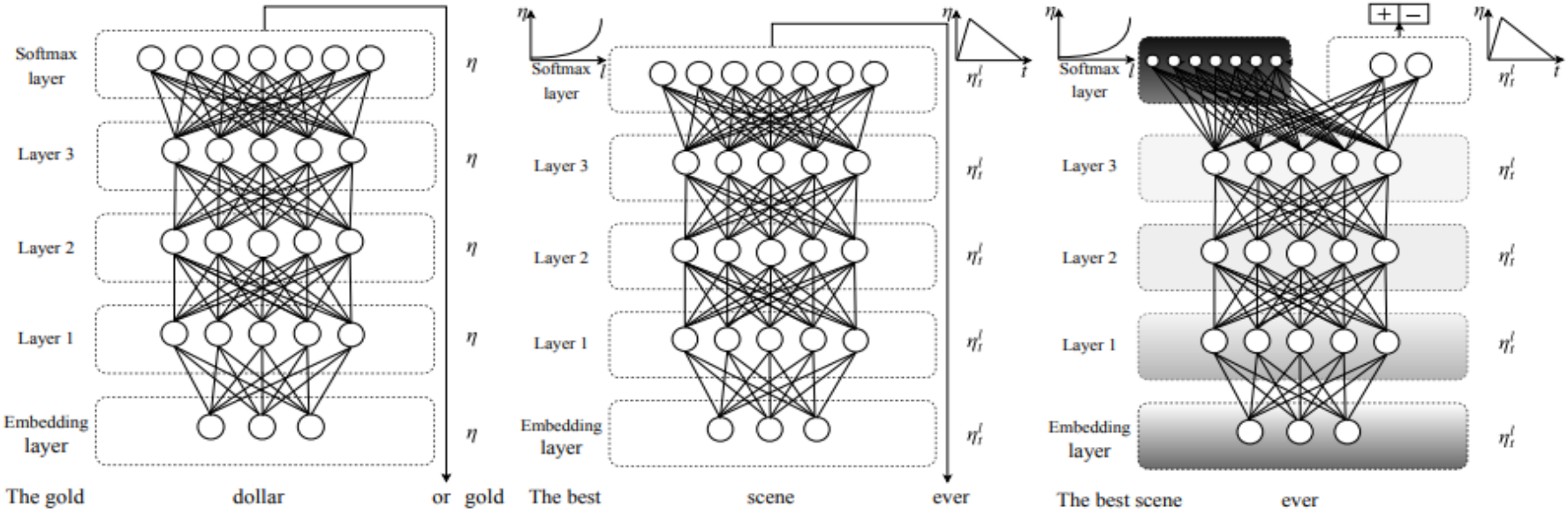
Neural-based Milestones for NLP

- 2001 - Neural language models
- 2008 - Multi-task learning
- 2013 - Word embeddings
- 2013 - Neural networks for NLP
- 2014 - Sequence-to-sequence models
- 2015 - Memory-based networks
- 2018 - Pretrained language models

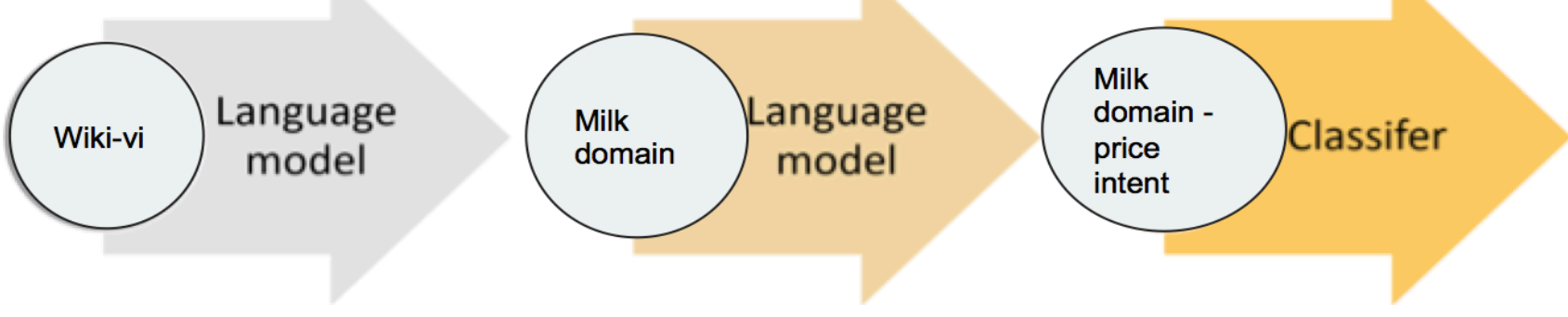
Character-based representations
Adversarial learning
Reinforcement learning

Conditional random fields
Bleu
Latent Dirichlet Allocation
Wikipedia exploitation

Pre-trained Neural Language Model



(a) LM pre-training (b) LM fine-tuning (c) Classifier fine-tuning



Linh Nguyen
Comment on the wall Bạch Hoàn. Post: Cuối cùng thì dù có - Nov 14, 2018 08:19 AM GMT+07:00

0 0

Ôi lại hội sữa người, sữa thú huhu. Con tới tuổi đã lớn, thì uống sữa thú chứ biết sao. Sữa bắp, gạo, hạt... nguyên chất nào đủ hàm lượng canxi, mà mua loại hữu cơ có bổ sung canxi thì đắt gấp 3 lần sữa thú... Sữa thủy phân đắt và khó tìm, thường chỉ dùng cho trẻ bị chàm và dị ứng thôi, trẻ khỏe mạnh chẳng tội gì phải uống.

3

Ngọc Khánh
Comment on the wall Bạch Hoàn. Post: Cuối cùng thì dù có - Nov 14, 2018 12:53 AM GMT+07:00

0 0

Các cháu có sữa uống là tốt rồi lại còn rẻ hơn thị trường nữa là có lợi cho túi tiền phụ huynh.
Còn tình trạng chung hoặc những nghi vấn về chất lượng sữa ở VN lại là một vấn đề rộng hơn

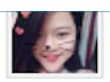
3

Kim Thuan
Comment on the wall Bạch Hoàn. Post: Cuối cùng thì dù có - Nov 13, 2018 11:31 PM GMT+07:00

0 0

Cuối tuần này mới có kết quả chính thức mà, nhiều khi đấu thầu giá rẻ hơn chưa chắc đã thắng vì còn tiêu chí khác nữa. Hy vọng vụ này xòng phẳng và Vinamilk thắng thầu.

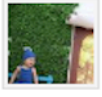
3



Comment on Hội Sữa Công Thức Sử Dụng Đúng Cách (Đọc Kỹ Nội Quy Ở Bài Ghim)'s group. Post: Bình thường e mua - Nov 08, 2018 08:16 PM GMT+07:00

😔 mua ở cửa hàng chính thức của th giá là nt đấy chị ơi . Cũng chả đắt đâu , chênh 20k mà đc cái đảm bảo cty , 2 nữa là hay có km nữa . Con e dùng vinamilk e cũng chỉ mua ở giấc mơ sữa việt thôi nè . Yên tâm hơn

2 1



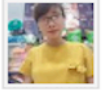
Hoàng Diệu

0 0

Comment on Hội Sữa Công Thức Sử Dụng Đúng Cách (Đọc Kỹ Nội Quy Ở Bài Ghim)'s group. Post: Bình thường e mua - Nov 08, 2018 08:03 PM GMT+07:00

Giá chính xác dao động từ 215 đến 218. Bạn mua 240k là quá đắt. Còn khác nhau là do 2 nhà phân phối ak. Ko phải sữa giả đâu. Mình bán nên mình biết ak. Mom yên tâm nha.

2 1



Hoan Nguyễn

0 0

Comment on Hội Sữa Công Thức Sử Dụng Đúng Cách (Đọc Kỹ Nội Quy Ở Bài Ghim)'s group. Post: Bình thường e mua - Nov 08, 2018 07:55 PM GMT+07:00

Tất cả các hãng sữa đều có 2 bao bì,do có 2 cty cung cấp bao bì đó ạ. Mn ko cần hoang mang đâu. Mua ở đại lý chuyên sữa nó sẽ bán đúng giá niêm yết,còn mua ở tạp hóa họ ăn số lượng nên bán rẻ hơn. Mua sữa lúc nào cũng gặp th như vậy đấy ạ.

2

Intuition

Sua pediasure giá bnhju vay

Intuition

Sua pediasure giá bnhju vay

Cái này giá bao nhiêu vậy
Giá nó bao nhiêu

Bao nhiêu một hộp sữa vậy
Giá hộp sữa bao nhiêu
Giá hộp này bnhju
Bnhju thì bán được

Language model
Intent classification
NER tagging

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Devlin et al., 2018 (Google AI Language)

NLP course

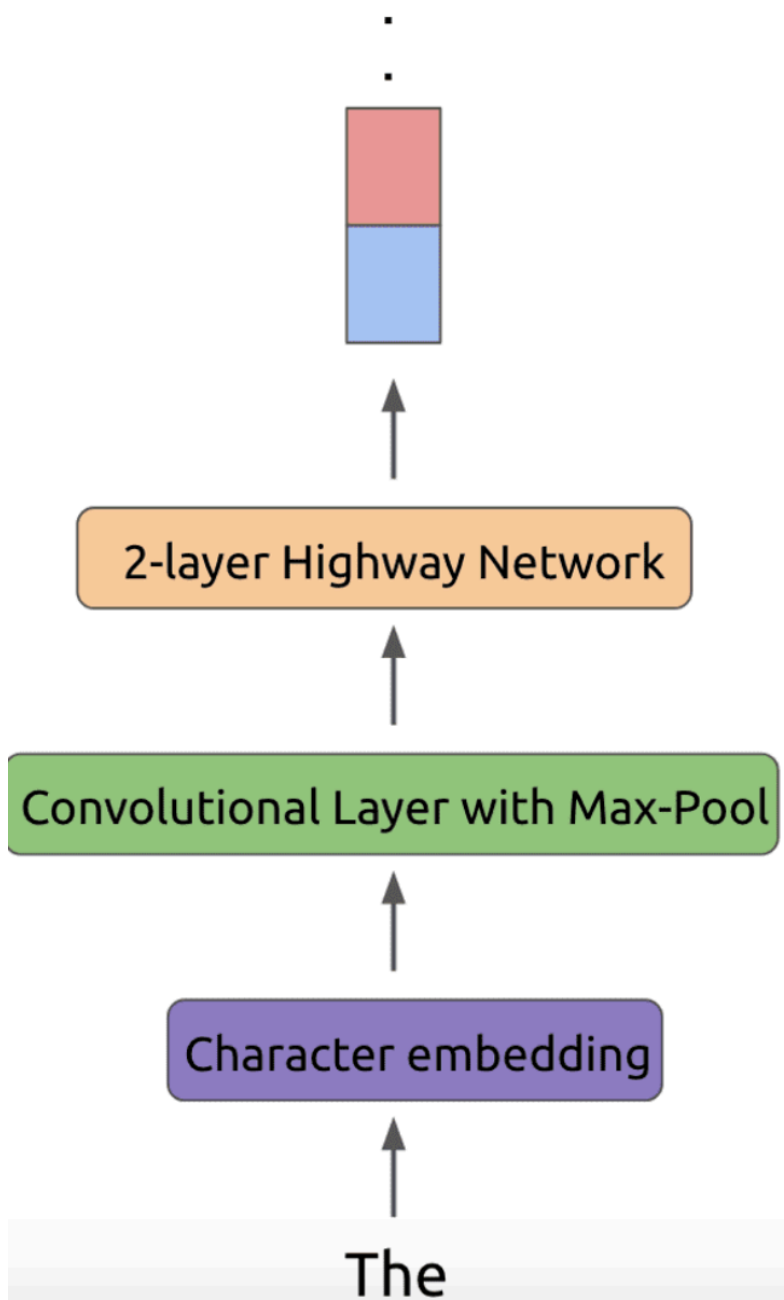
Outline

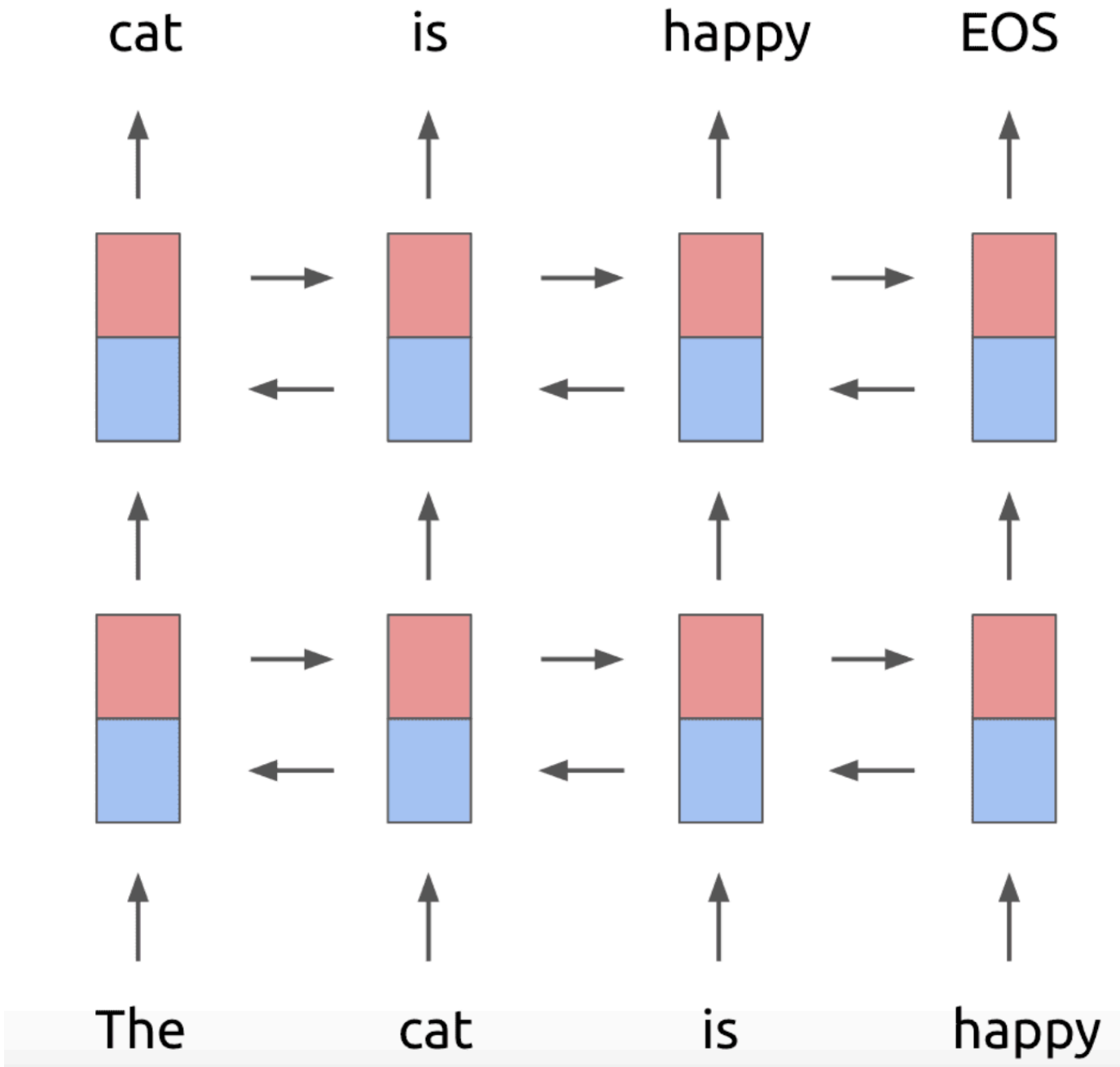
- Research context
- Main ideas
- BERT
- Experiments
- Conclusions

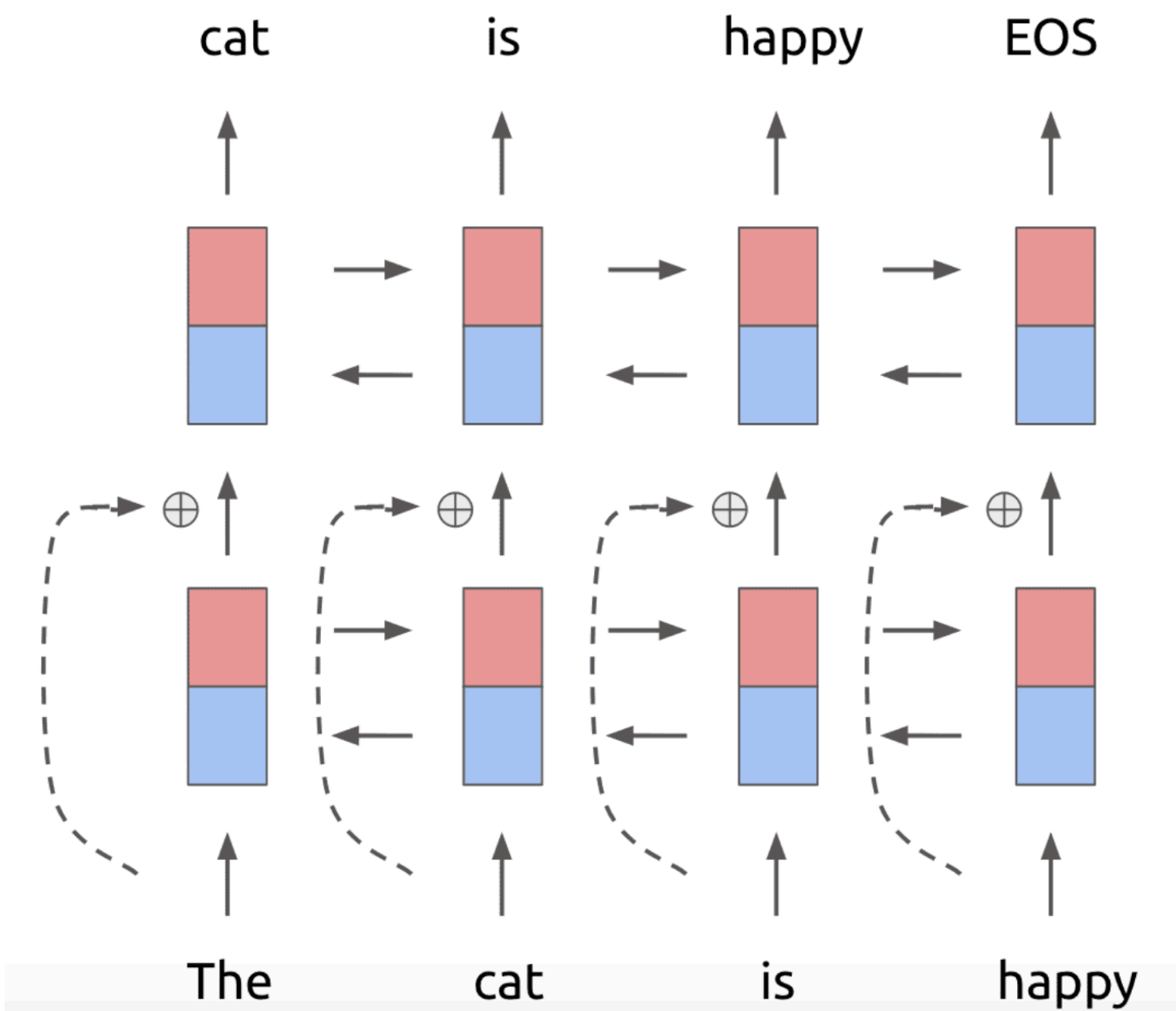
Research context

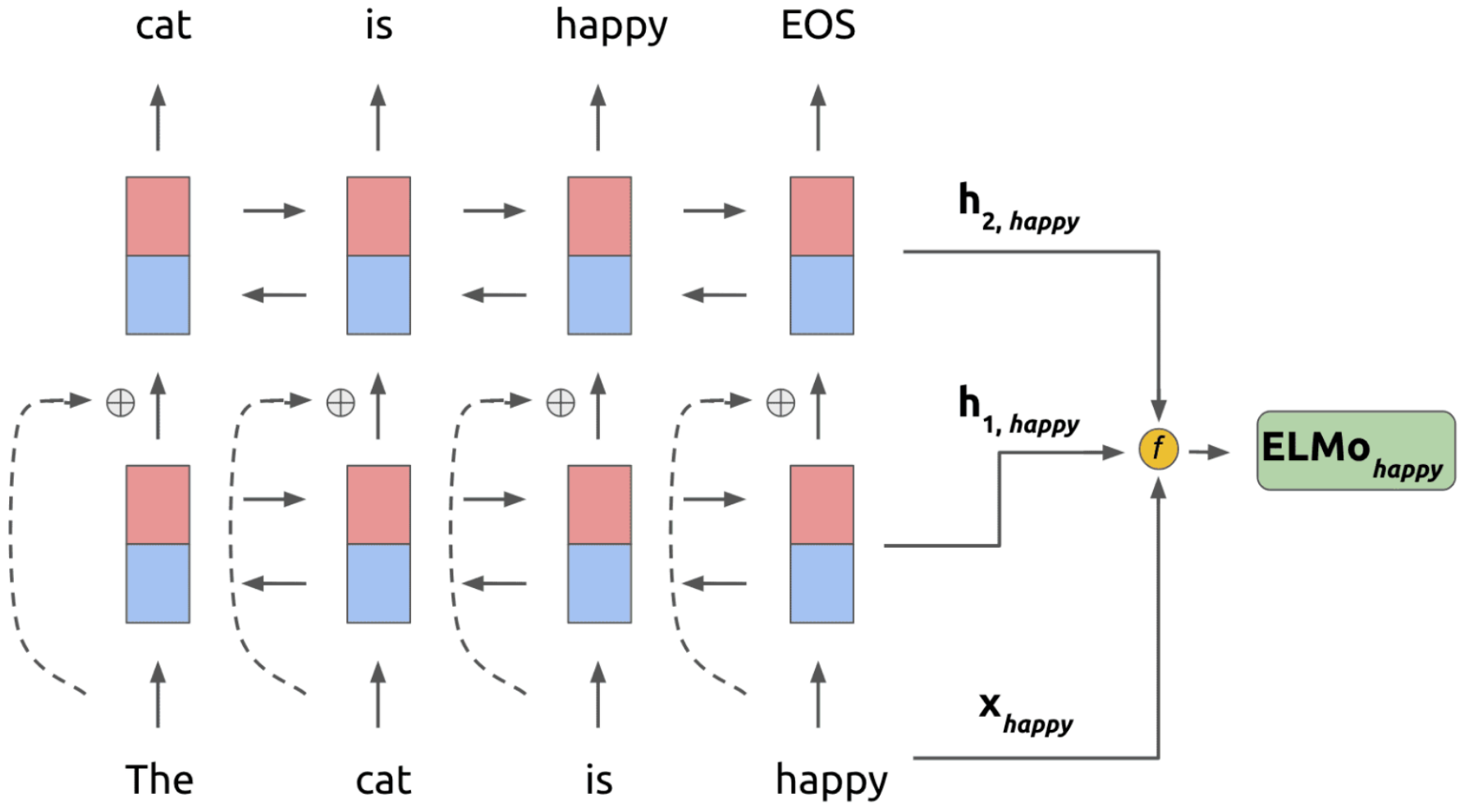
- Language model pre-training has been used to improve many NLP tasks
  - ELMo (Peters et al., 2018)
  - OpenAI GPT (Radford et al., 2018)
  - ULMFit (Howard and Rudder, 2018)
- Two existing strategies for applying pre-trained language representations to downstream tasks
  - *Feature-based*: include pre-trained representations as additional features (e.g., ELMo)
  - *Fine-tunning*: introduce task-specific parameters and fine-tune the pre-trained parameters (e.g., OpenAI GPT, ULMFit)

ELMO


2-layer Highway Network
Convolutional Layer with Max-Pool
Character embedding
The


cat
is
happy
EOS
The
cat
is
happy


cat
is
happy
EOS
The
cat
is
happy


cat
is
happy
EOS
The
cat
is
happy
$h_{2,\ happy}$
$h_{1,\ happy}$
x_{happy}
f
$ELMo_{happy}$

ULMFit

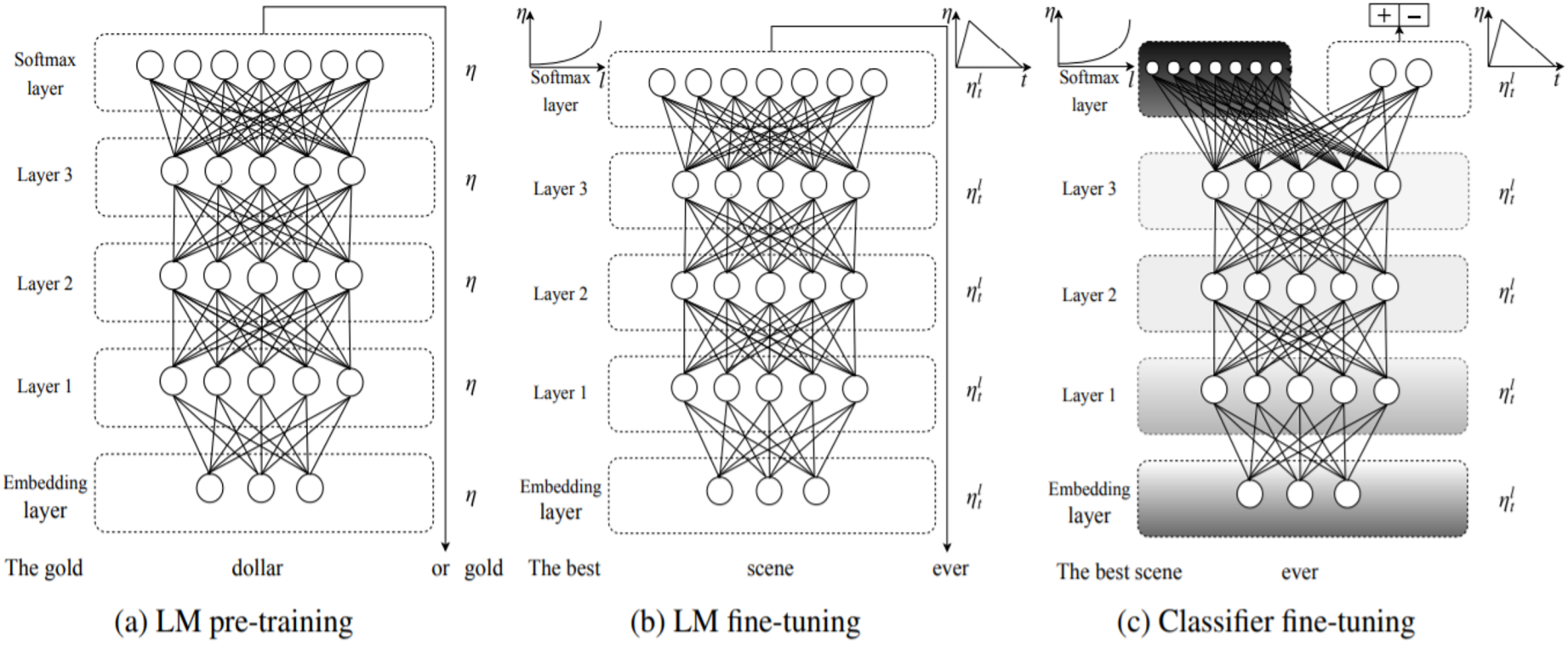


Figure 1: ULMFiT consists of three stages: a) The LM is trained on a general-domain corpus to capture general features of the language in different layers. b) The full LM is fine-tuned on target task data using discriminative fine-tuning ('*Discr*') and slanted triangular learning rates (STLR) to learn task-specific features. c) The classifier is fine-tuned on the target task using gradual unfreezing, '*Discr*', and STLR to preserve low-level representations and adapt high-level ones (shaded: unfreezing stages; black: frozen).

Limitations of current techniques

- Language models in pre-training are unidirectional, they restrict the power of the pre-trained representations
  - OpenAI GPT used left-to-right architecture
  - ELMo concatenates *forward* and *backward* language models
- **Solution** BERT: Bidirectional Encoder Representations from Transformers

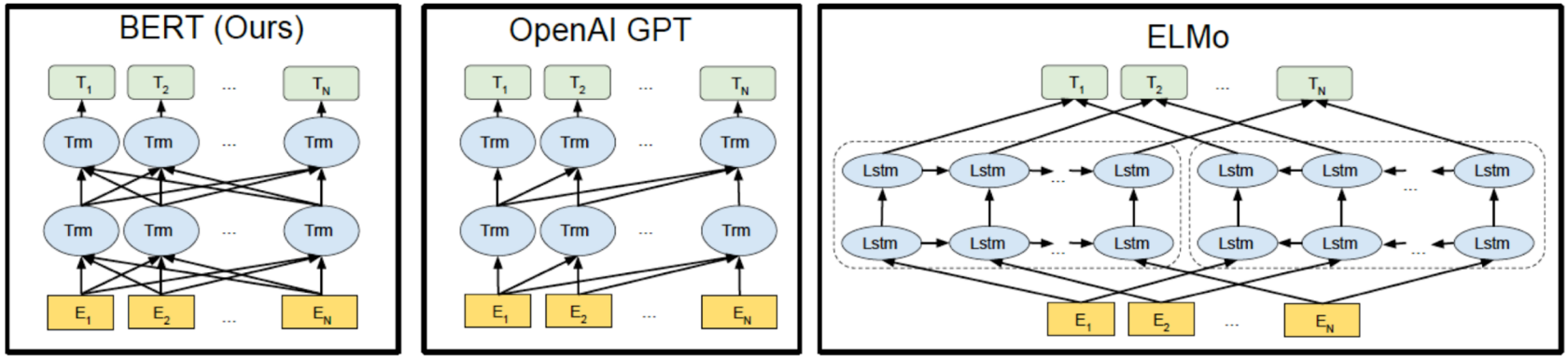
BERT: Bidirectional Encoder Representations from Transformers

- Main ideas
  - Propose a new pre-training objective so that a deep bidirectional Transformer can be trained
    - **The "masked language model" (MLM):** the objective is to predict the original word of a masked word based only on its context
    - **"Next sentence prediction"**
- Merits of BERT
  - Just fine-tune BERT model for specific tasks to achieve state-of-the-art performance
  - BERT advances the state-of-the-art for eleven NLP tasks

BERT: Bidirectional Encoder Representations from Transformers

Model architecture

- BERT's model architecture is a multi-layer **bidirectional Transformer encoder**
  - (Vaswani et al., 2017) "Attention is all you need"
- Two models with different sizes were investigated
  - $\text{BERT}_{\text{BASE}}$: L=12, H=768, A=12, Total Parameters=110M
    - (*L*: number of layers (Transformer blocks), *H* is the hidden size, *A*: the number of self-attention heads)
  - $\text{BERT}_{\text{LARGE}}$: L=24, H=1024, A=16, Total Parameters=340M

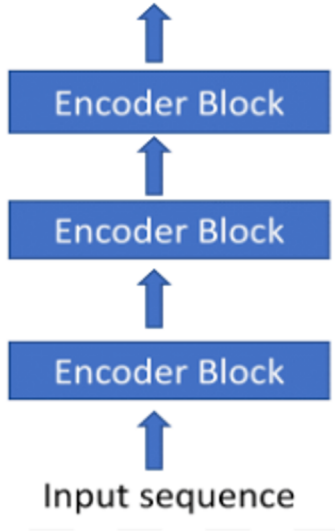


Differences in pre-training model architectures:
BERT, OpenAI GPT, and ELMo

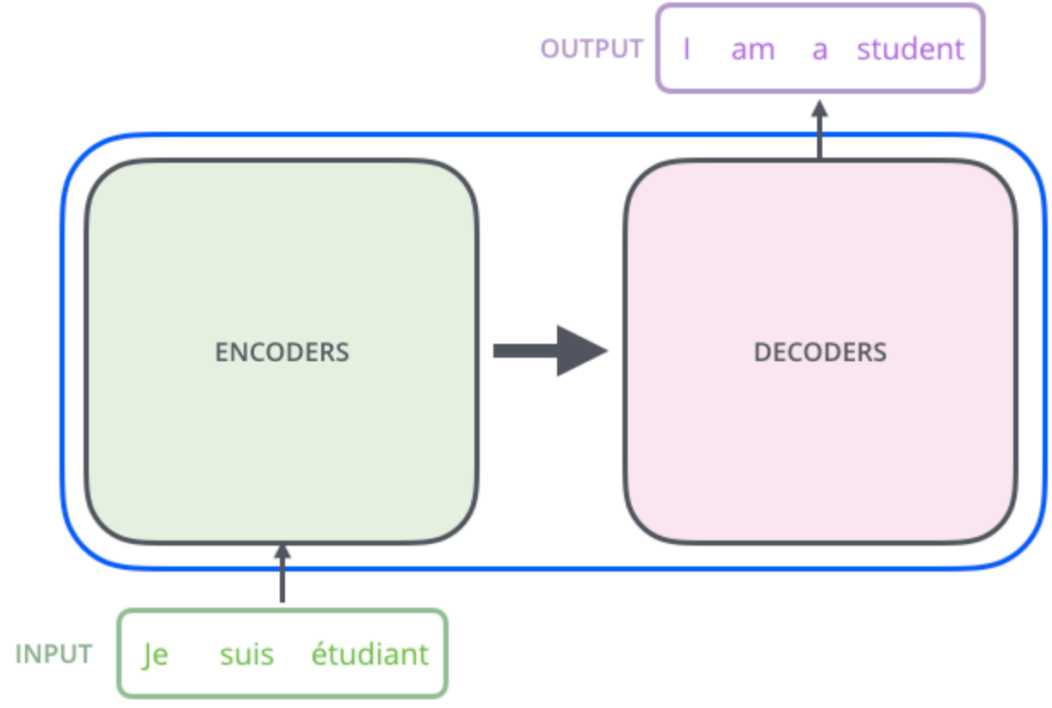
Transformer Encoders

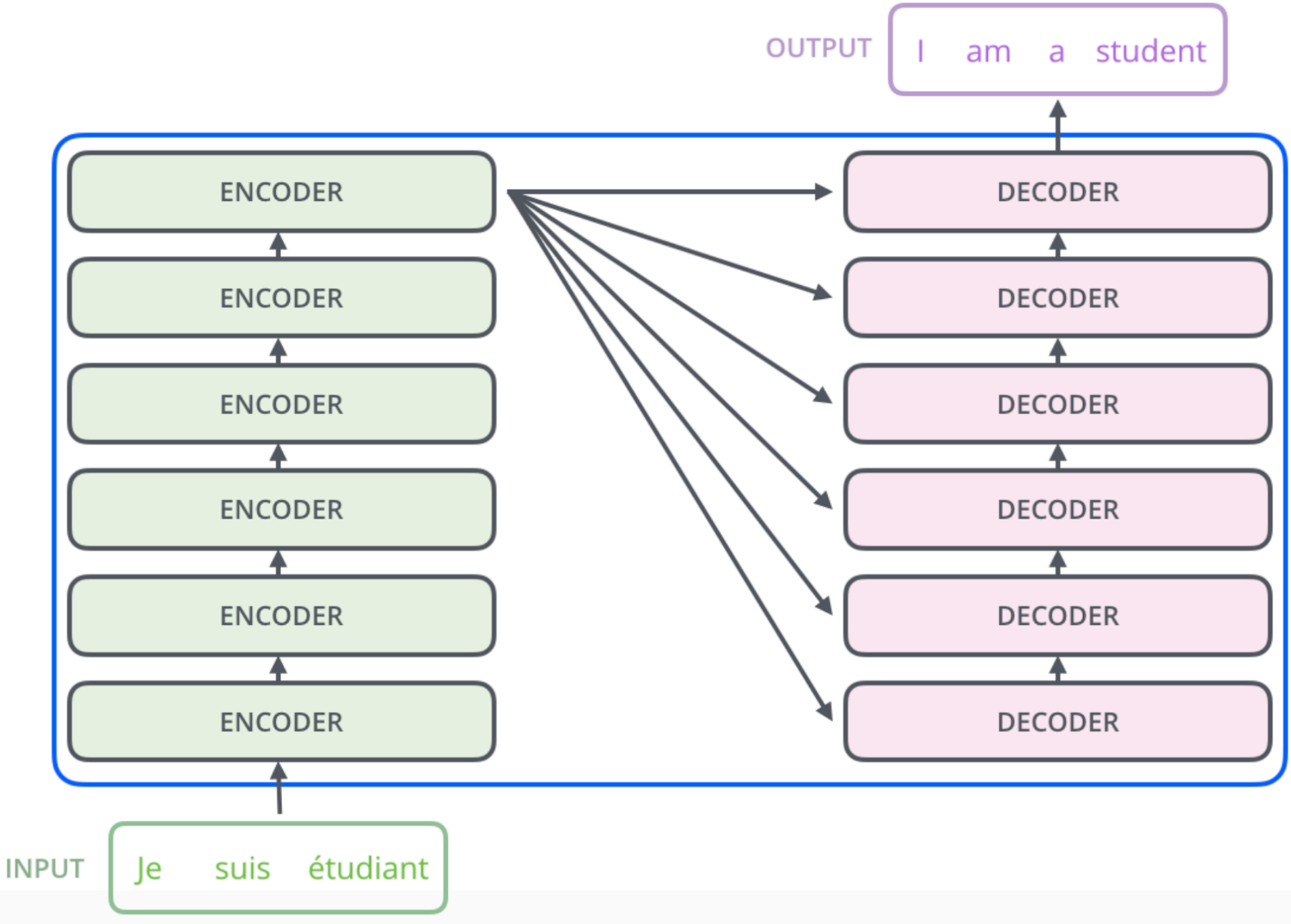
Vaswani et al. (2017) *Attention is all you need*

- Transformer is an attention-based architecture for NLP
- Transformer composed of two parts: **Encoding** component and **Decoding** component
- BERT is a multi-layer bidirectional Transformer **encoder**

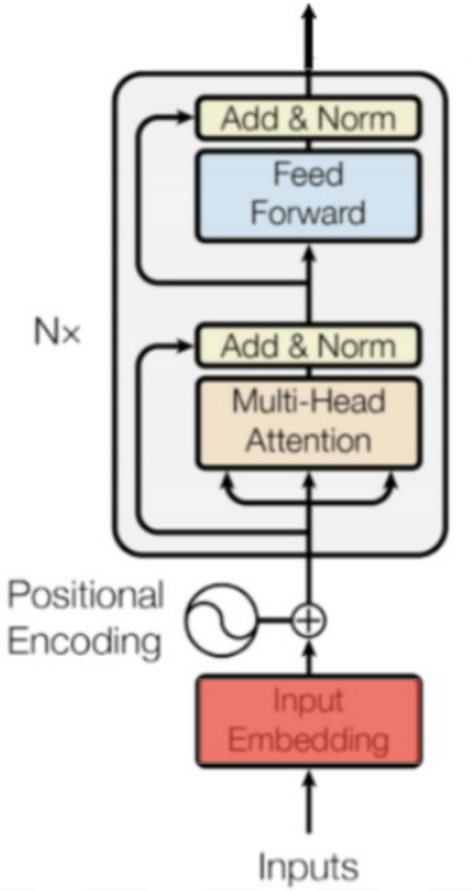



INPUT
Je suis étudiant
THE TRANSFORMER
OUTPUT
I am a student


OUTPUT
I am a student
ENCODERS
DECODERS
INPUT
Je suis étudiant


OUTPUT
I am a student
ENCODER
ENCODER
ENCODER
ENCODER
ENCODER
ENCODER
DECODER
DECODER
DECODER
DECODER
DECODER
DECODER
INPUT
Je suis étudiant

Inside an Encoder Block

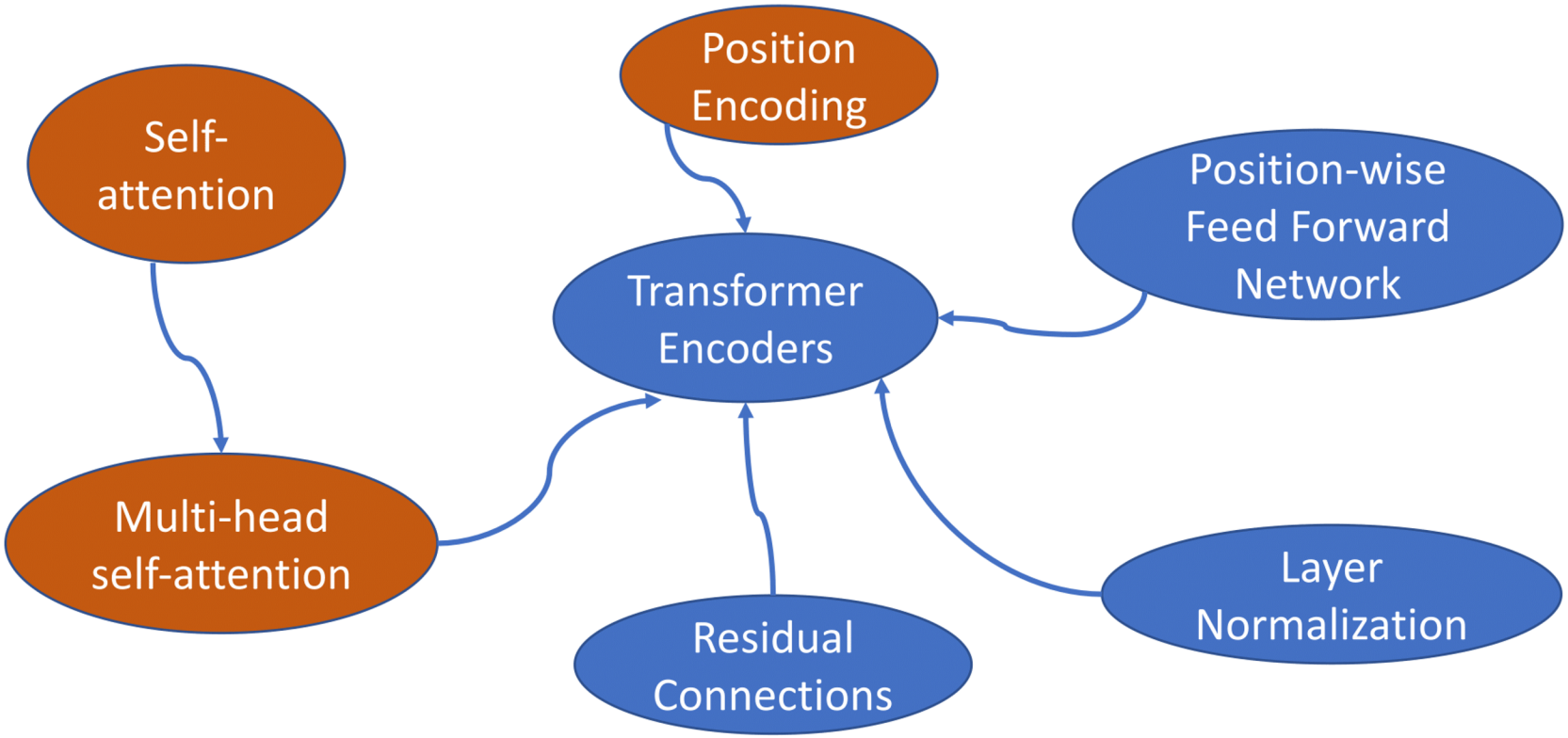


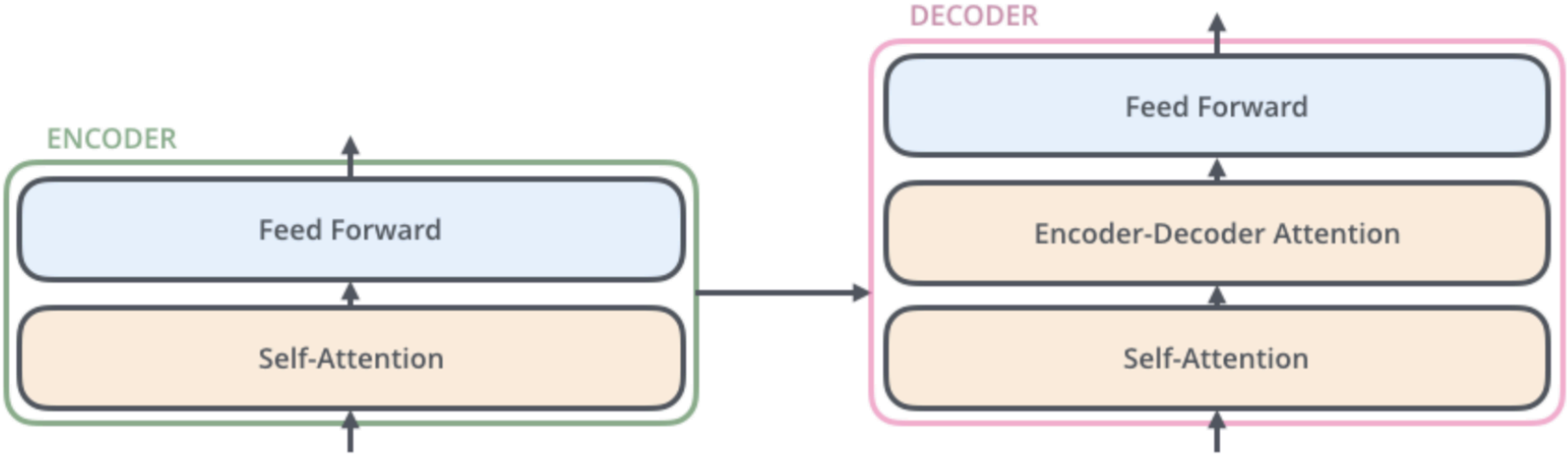
In BERT experiments, the number of blocks N was chosen to be 12 and 24.

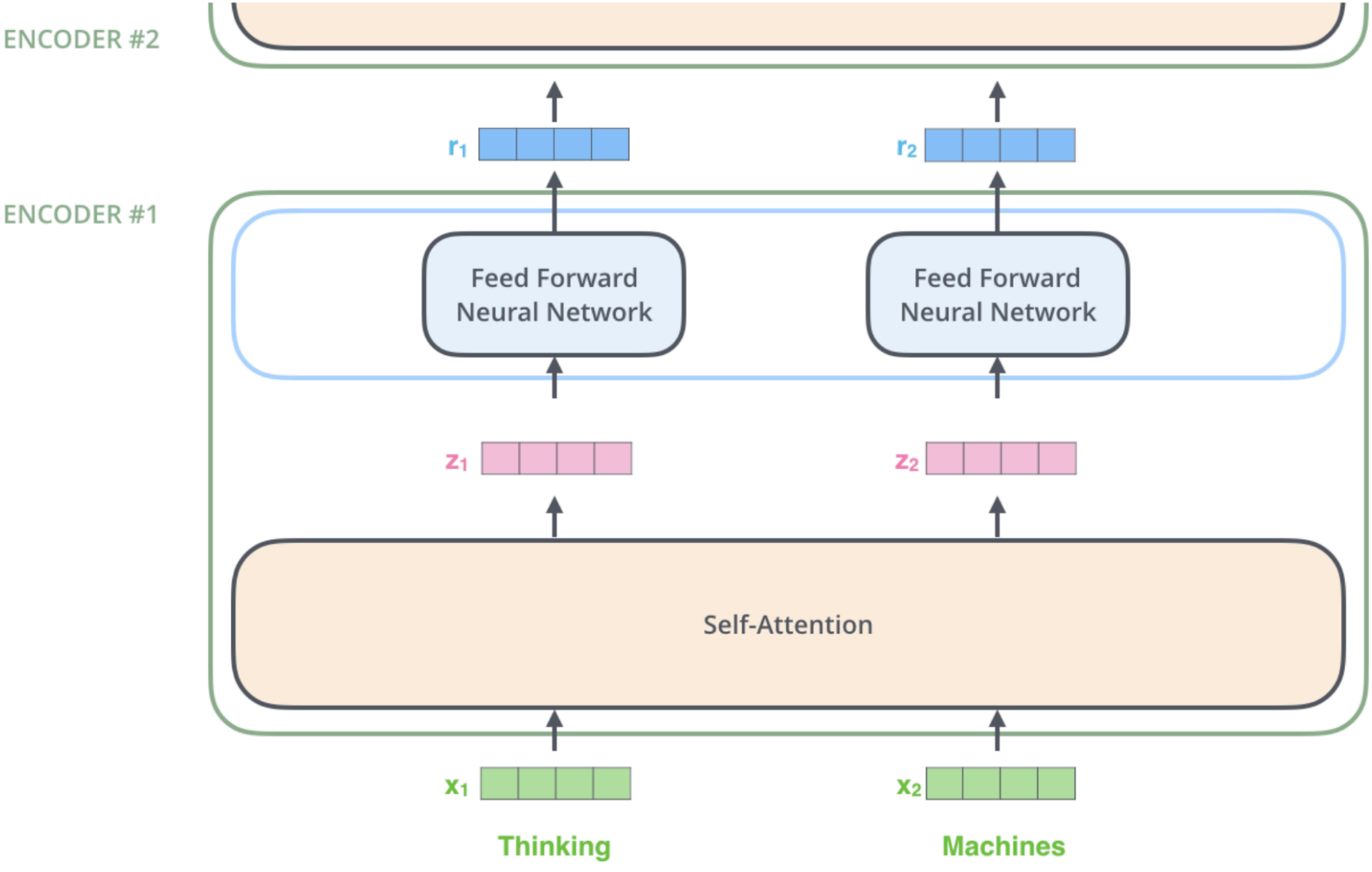
Blocks do not share weights with each other

Source: https://medium.com/dissecting-bert/dissecting-bert-part-1-d3c3d495cdb3

Transformer Encoders: Key Concepts




DECODER
Feed Forward
Encoder-Decoder Attention
Self-Attention
ENCODER
Feed Forward
Self-Attention


ENCODER #2
r1
r2
ENCODER #1
Feed Forward Neural Network
Feed Forward Neural Network
z1
z2
Self-Attention
x1
x2
Thinking
Machines

Self-Attention

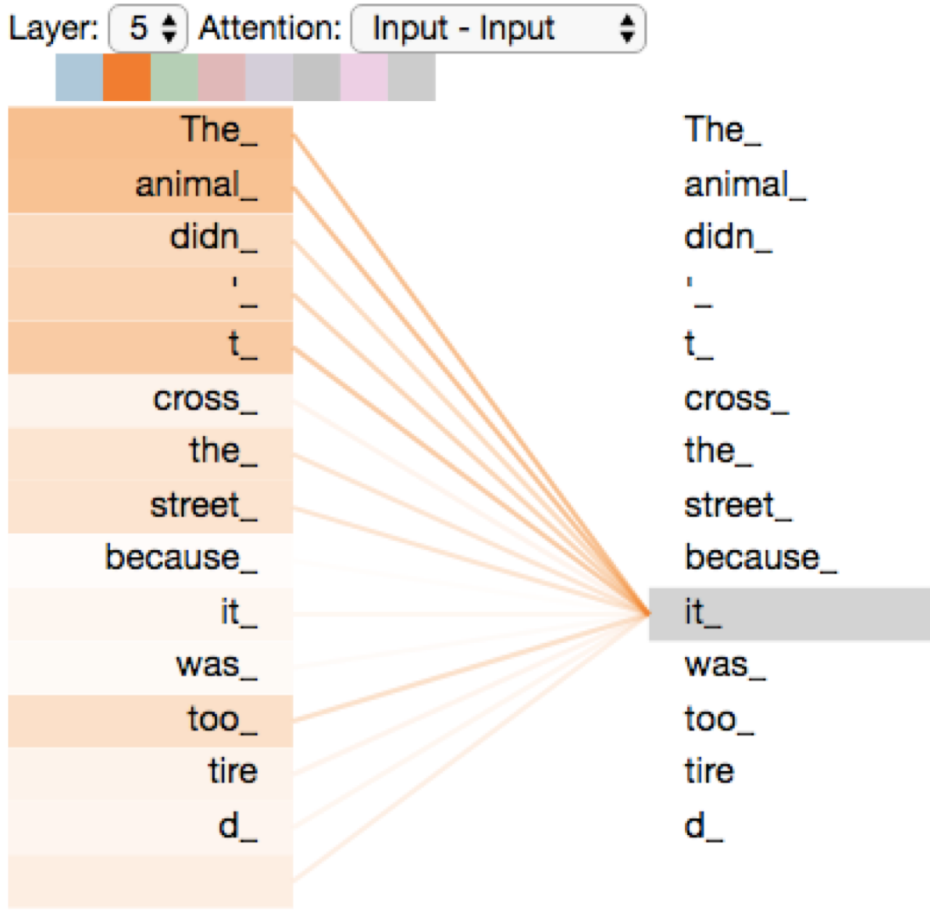
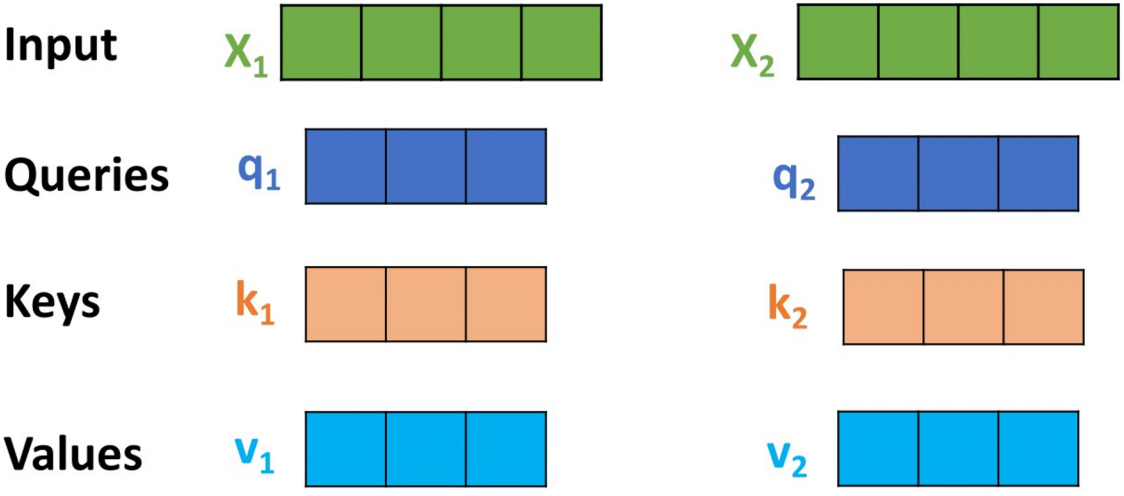


Image source: https://jalammar.github.io/illustrated-transformer/

Self-Attention in Detail

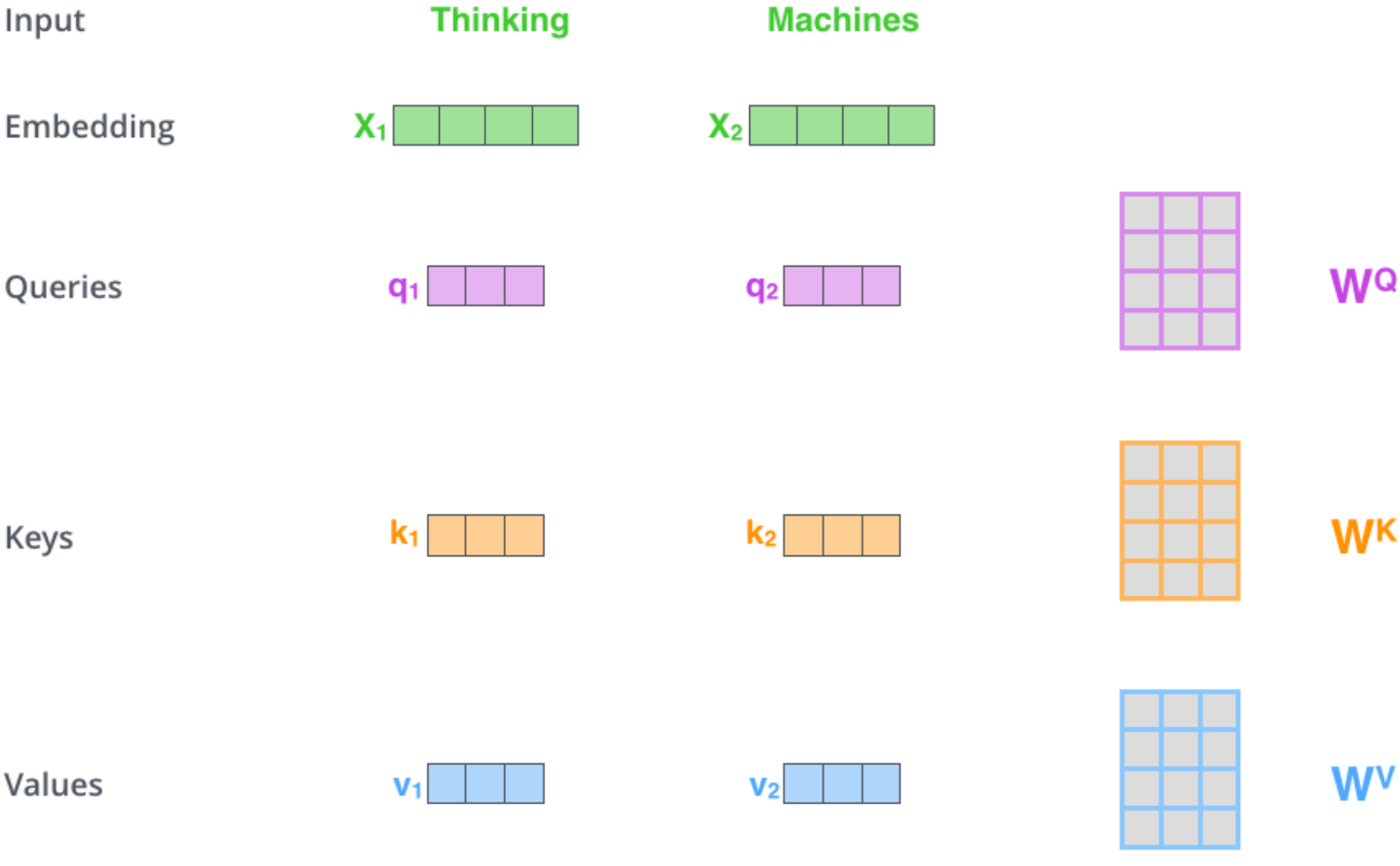
- Attention maps a query and a set of key-value pairs to an output
  - query, keys, and output are all vectors

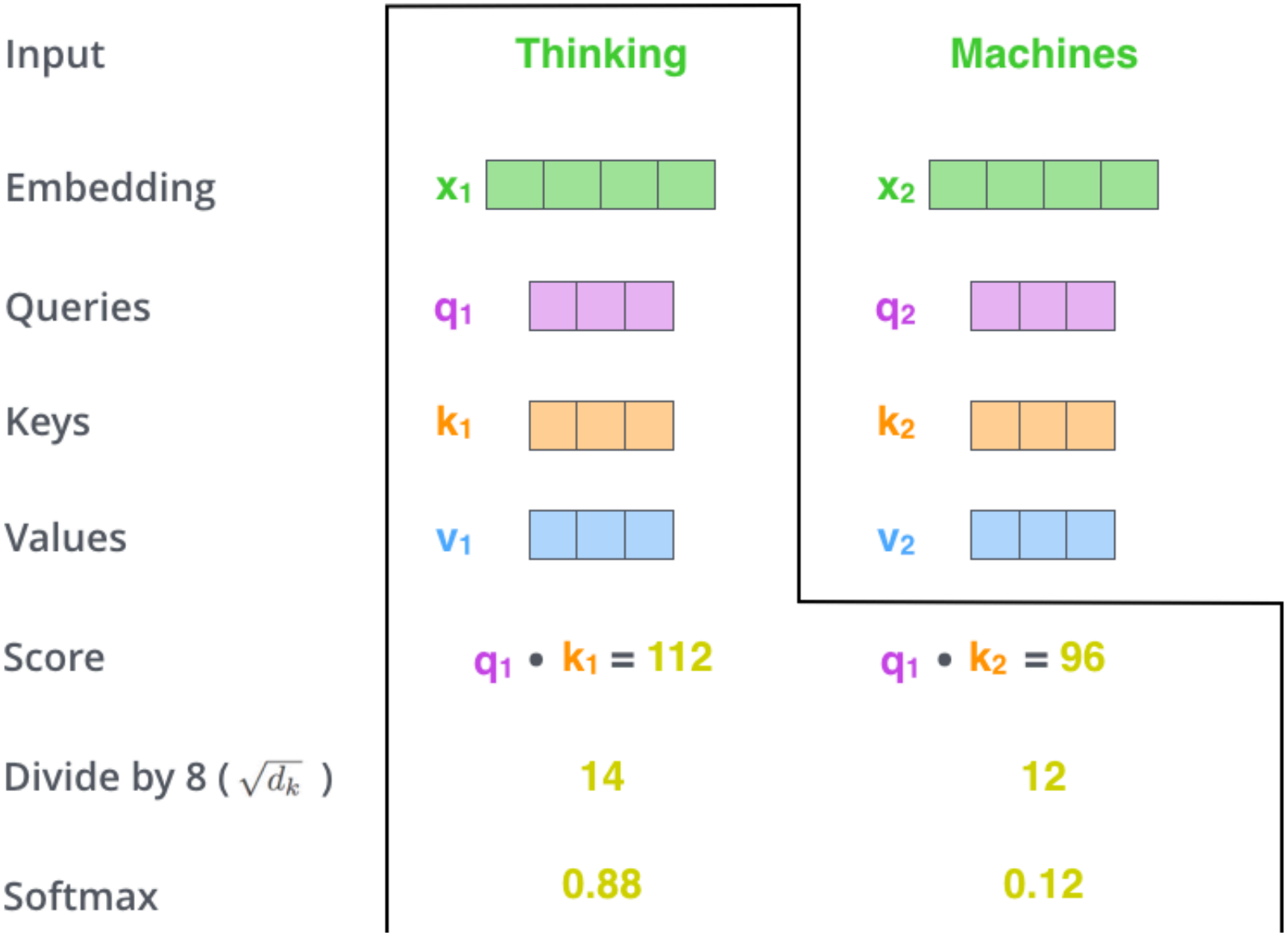


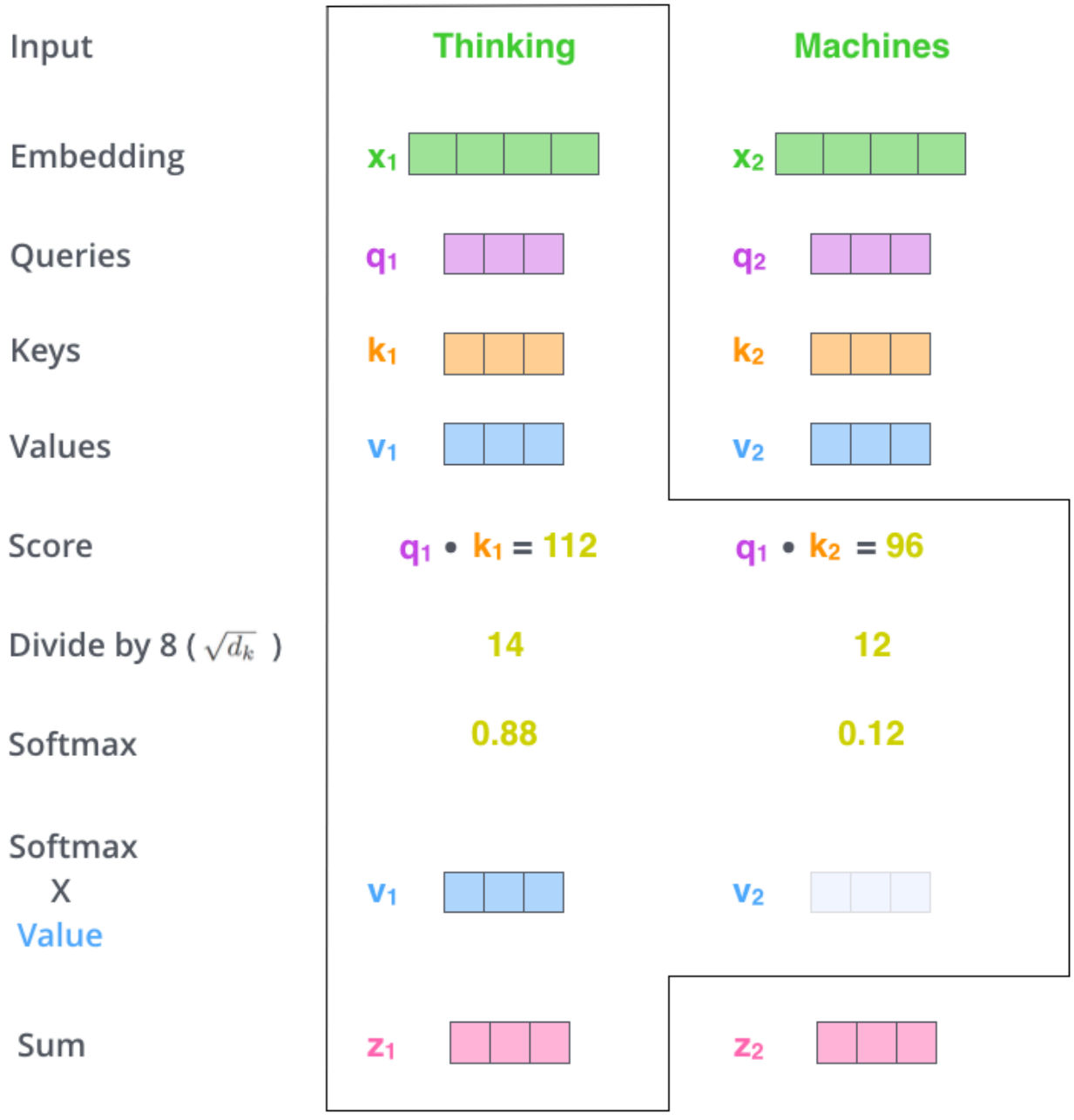
Use matrices W^Q, W^K and W^V to project input into query, key and value vectors

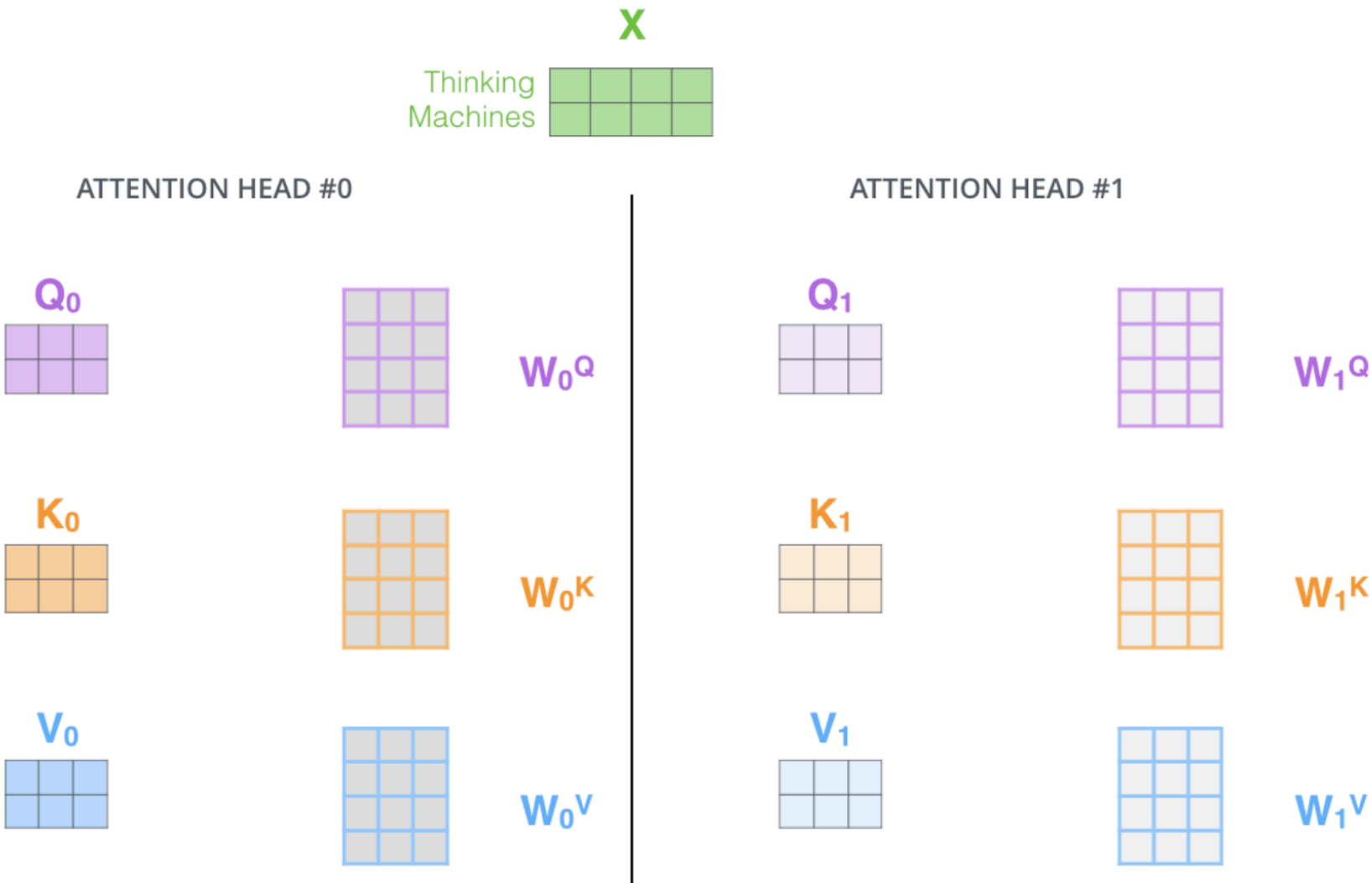
$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

d_k is the dimension of key vectors

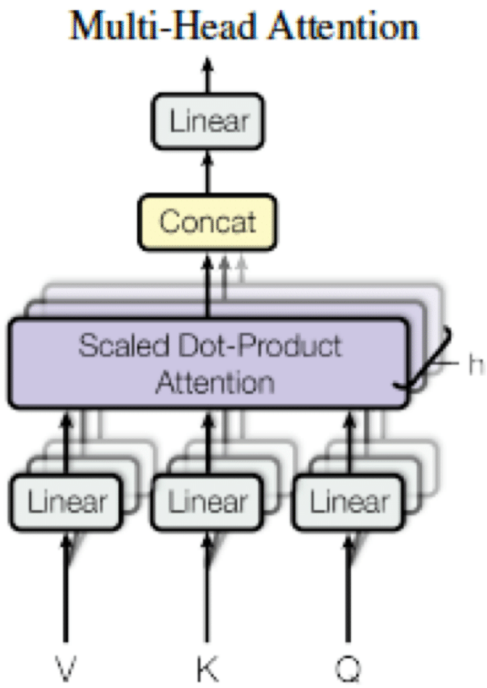

Input
Thinking
Machines
Embedding
X_1
X_2
Queries
q_1
q_2
W^Q
Keys
k_1
k_2
W^K
Values
v_1
v_2
W^V

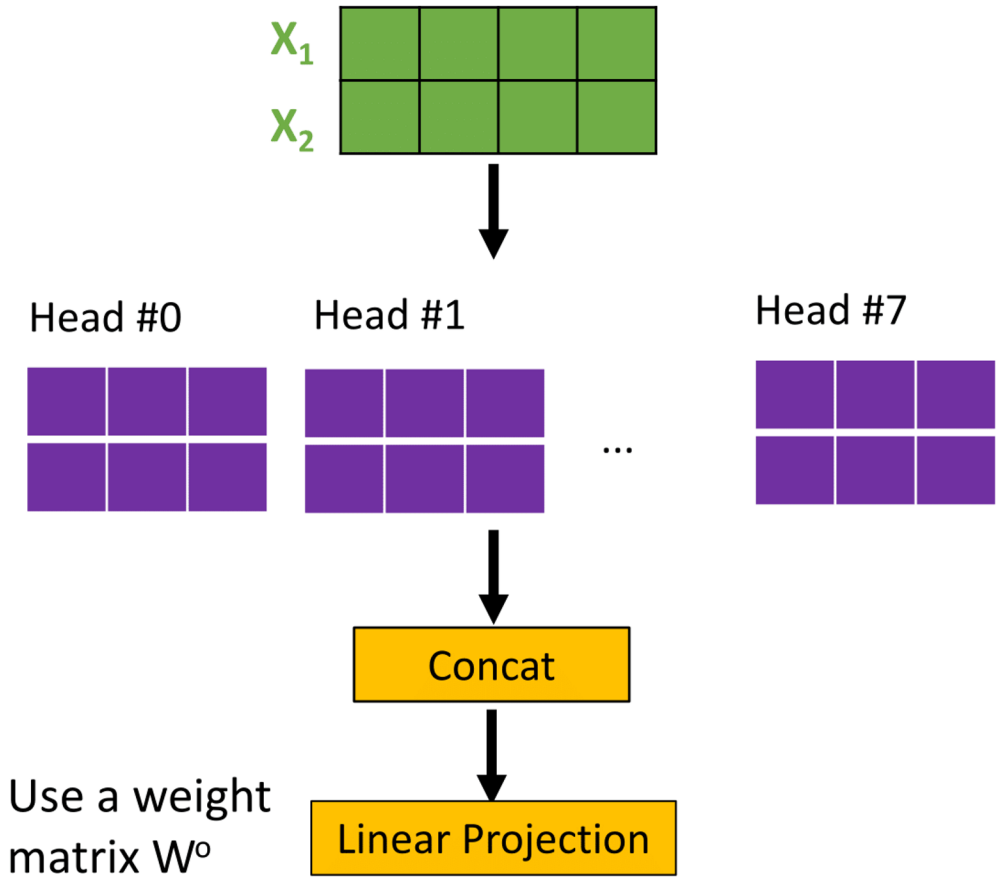

Input
Thinking
Machines
Embedding
x1
x2
Queries
q1
q2
Keys
k1
k2
Values
v1
v2
Score
q1 • k1 = 112
q1 • k2 = 96
Divide by 8 (√dk)
14
12
Softmax
0.88
0.12

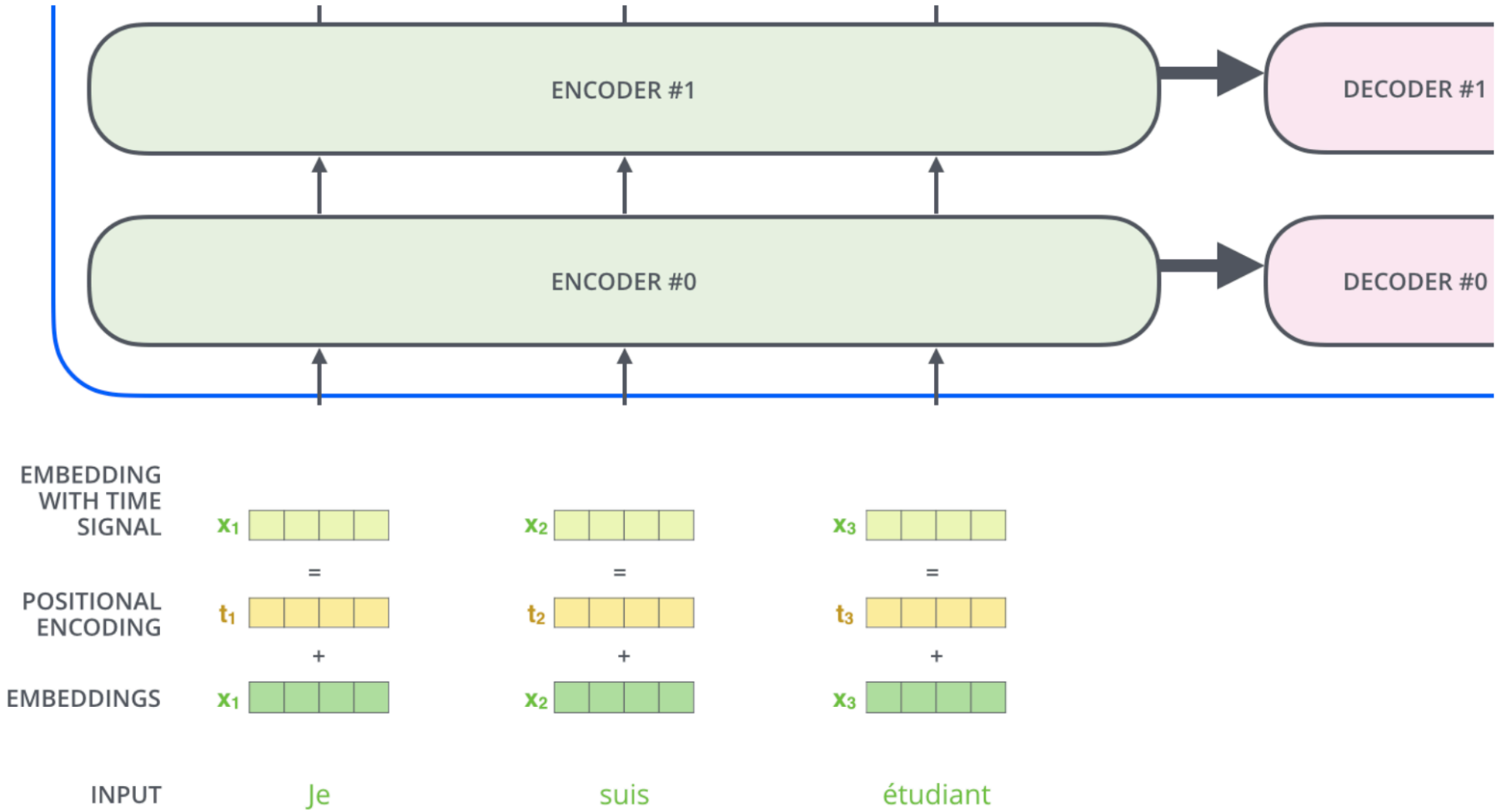

Input
Thinking
Machines
Embedding
x1
x2
Queries
q1
q2
Keys
k1
k2
Values
v1
v2
Score
q1 • k1 = 112
q1 • k2 = 96
Divide by 8 (√dk)
14
12
Softmax
0.88
0.12
Softmax X Value
v1
v2
Sum
z1
z2

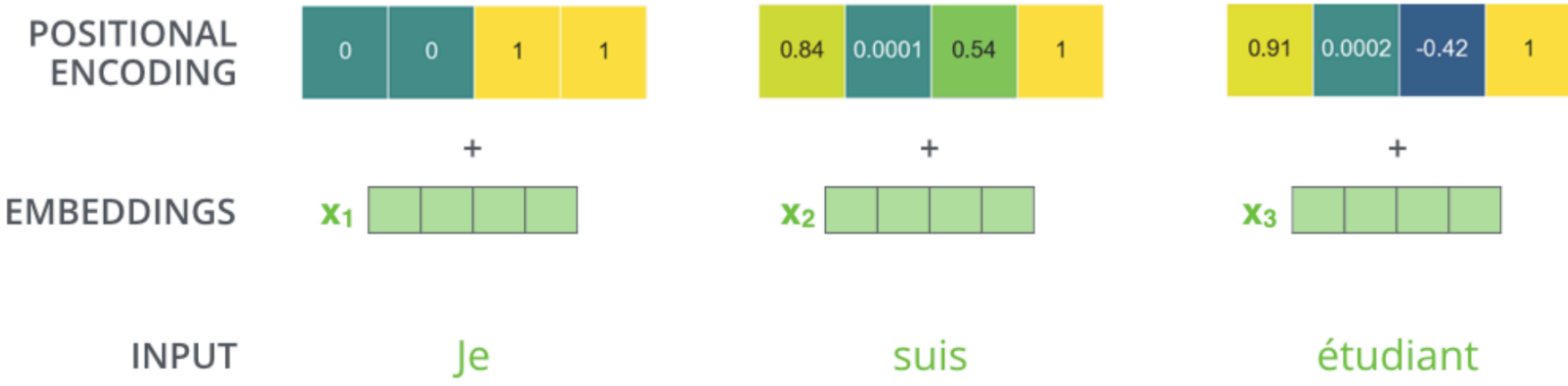

X
Thinking
Machines
ATTENTION HEAD #0
ATTENTION HEAD #1
Q0
K0
V0
W0Q
W0K
W0V
Q1
K1
V1
W1Q
W1K
W1V

Multi-Head Attention






ENCODER #1
DECODER #1
ENCODER #0
DECODER #0
EMBEDDING WITH TIME SIGNAL
POSITIONAL ENCODING
EMBEDDINGS
INPUT
x1
x2
x3
t1
t2
t3
=
+
Je
suis
étudiant


POSITIONAL ENCODING
0
0
1
1
0.84
0.0001
0.54
1
0.91
0.0002
-0.42
1
+
+
+
EMBEDDINGS
x1
x2
x3
INPUT
Je
suis
étudiant

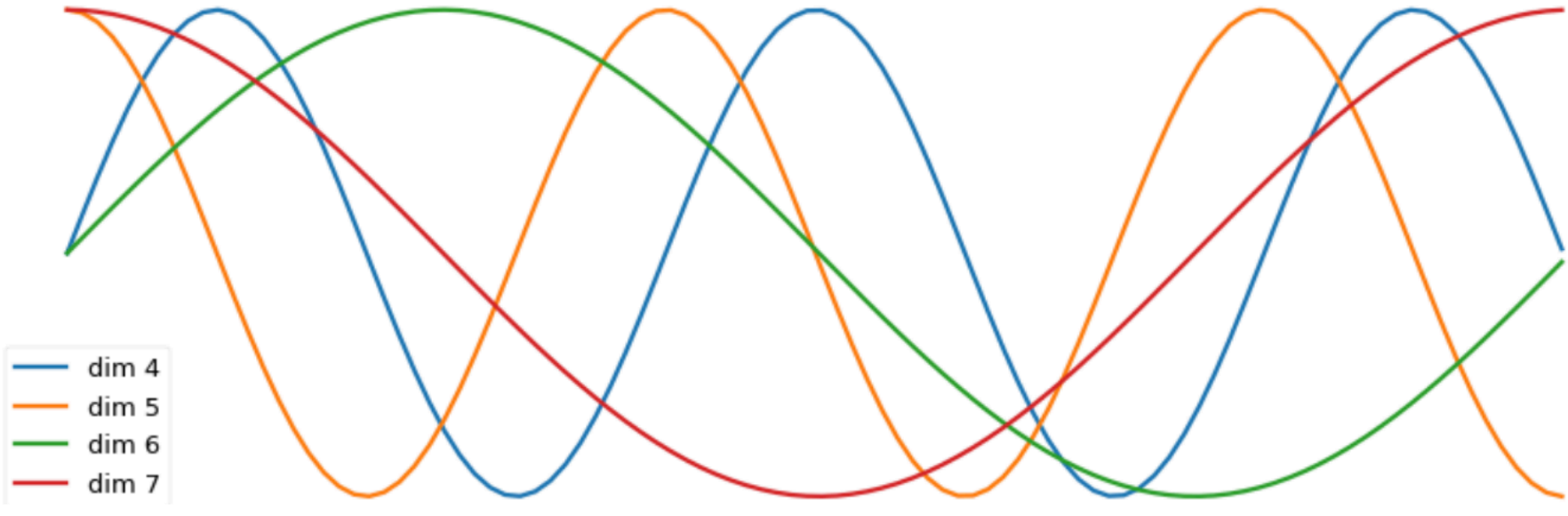
Position Encoding

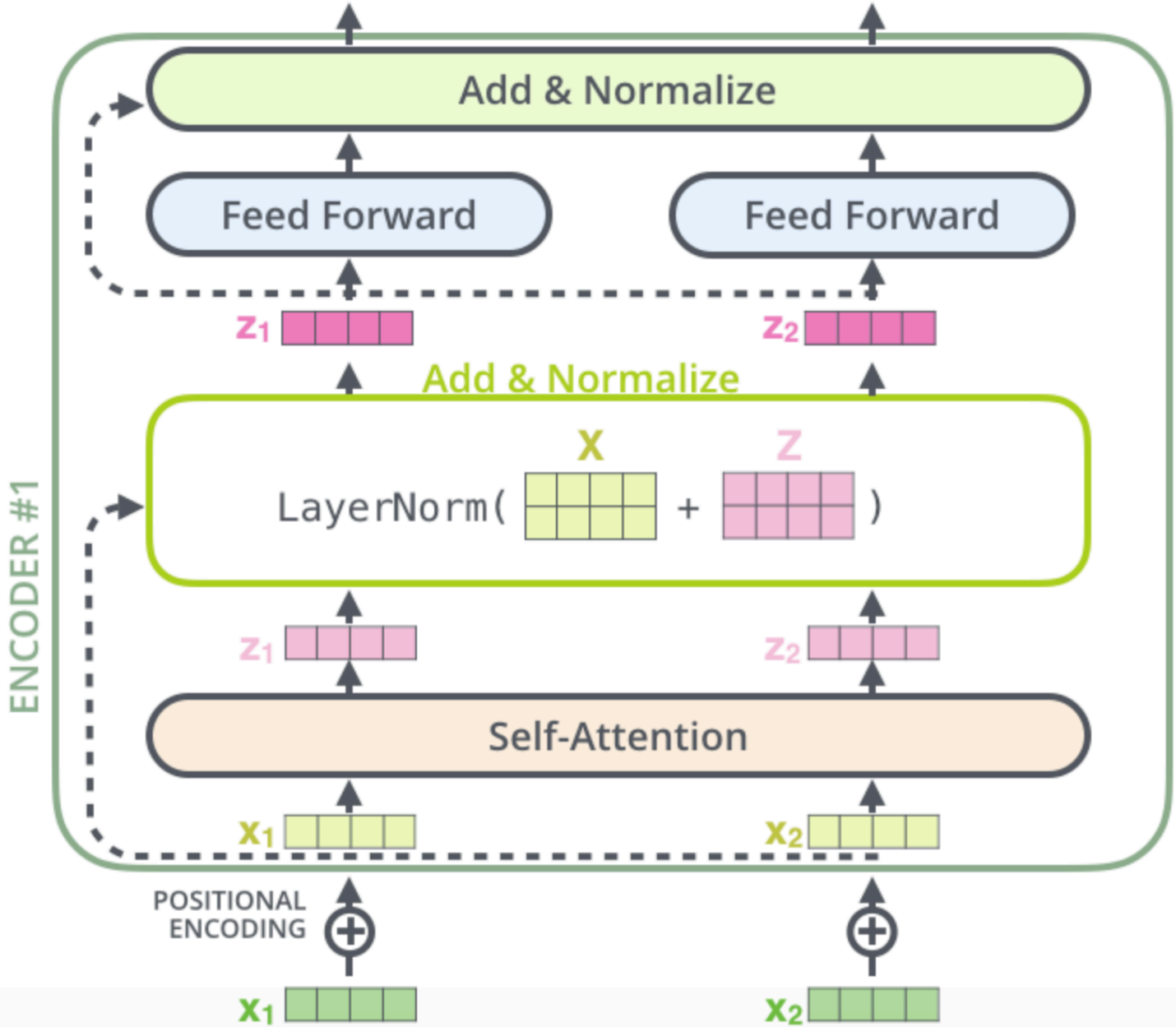
- Position Encoding is used to make use of the order of the sequence
  - Since the model contains no recurrence and no convolution
- In Vawasni et al., 2017, authors used sine and cosine functions of different frequencies

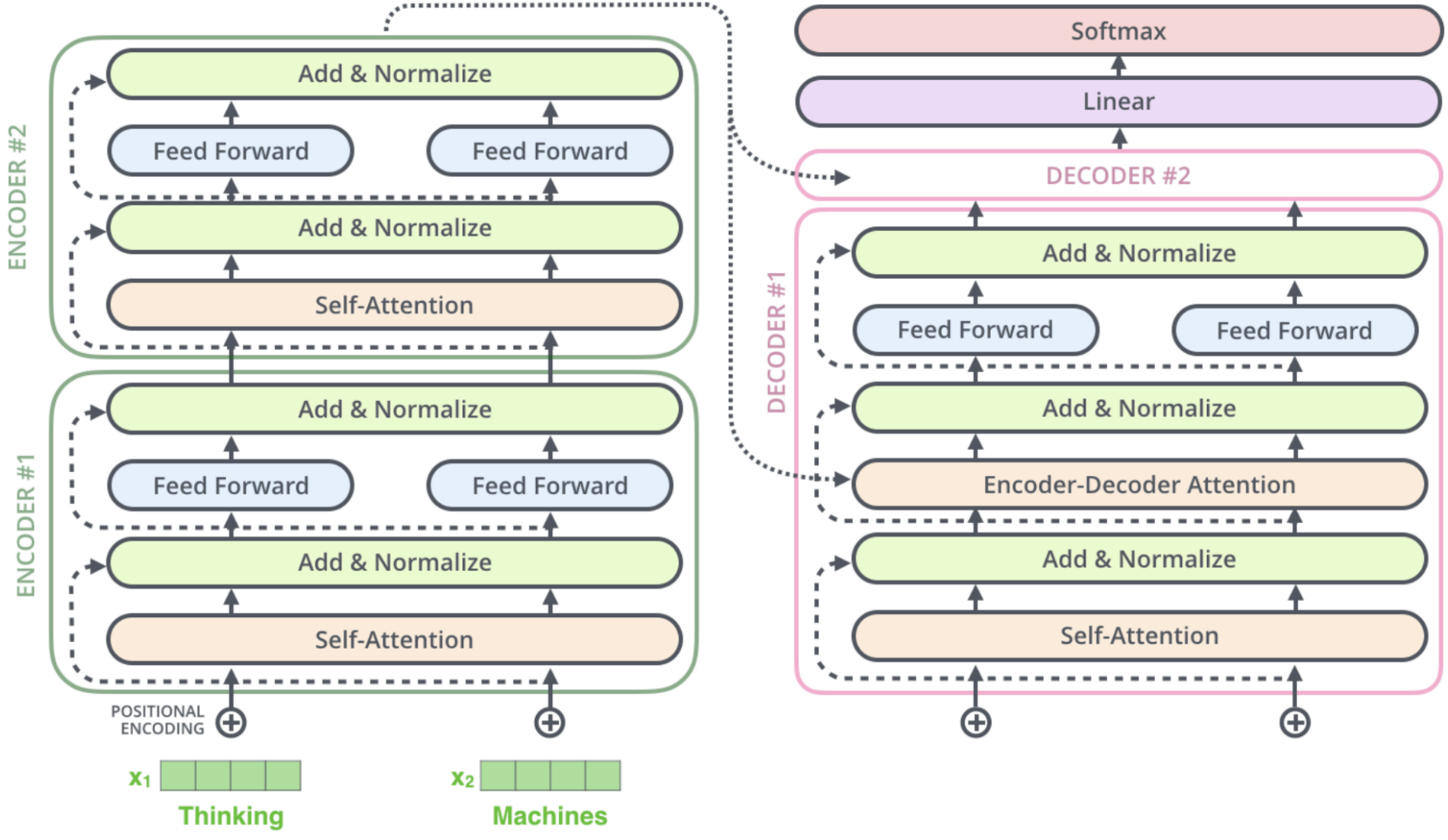
$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right)$$

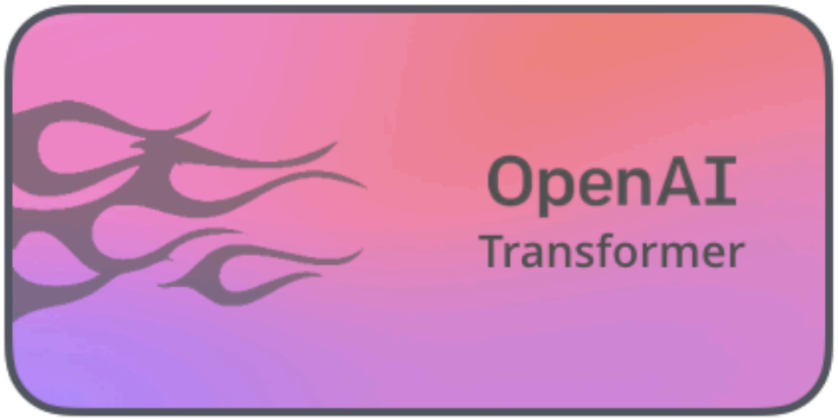
$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right)$$

  - *pos* is the position and i is the dimension

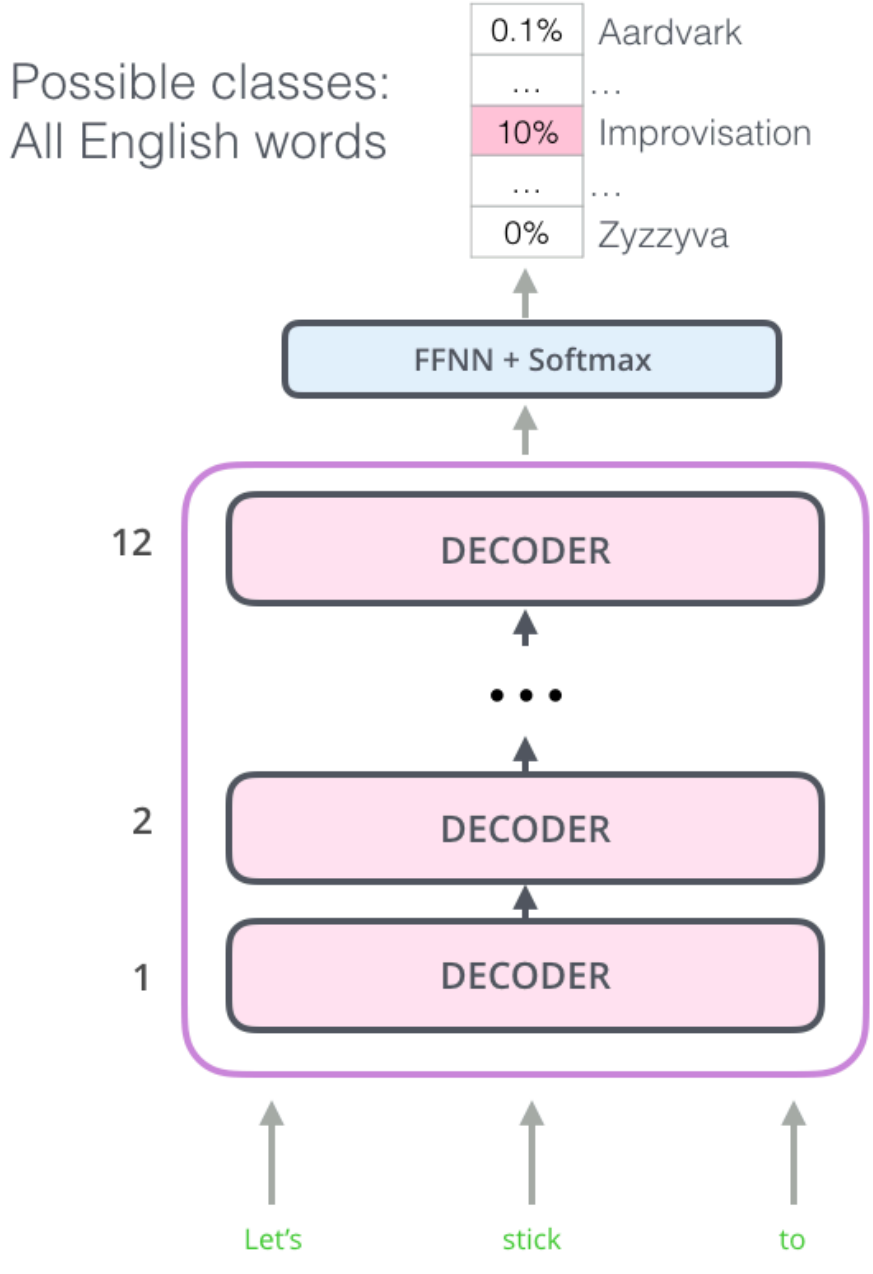

dim 4
dim 5
dim 6
dim 7


Add & Normalize
Feed Forward
Feed Forward
z1
z2
Add & Normalize
X
Z
LayerNorm(+)
z1
z2
Self-Attention
x1
x2
ENCODER #1
POSITIONAL ENCODING
x1
x2

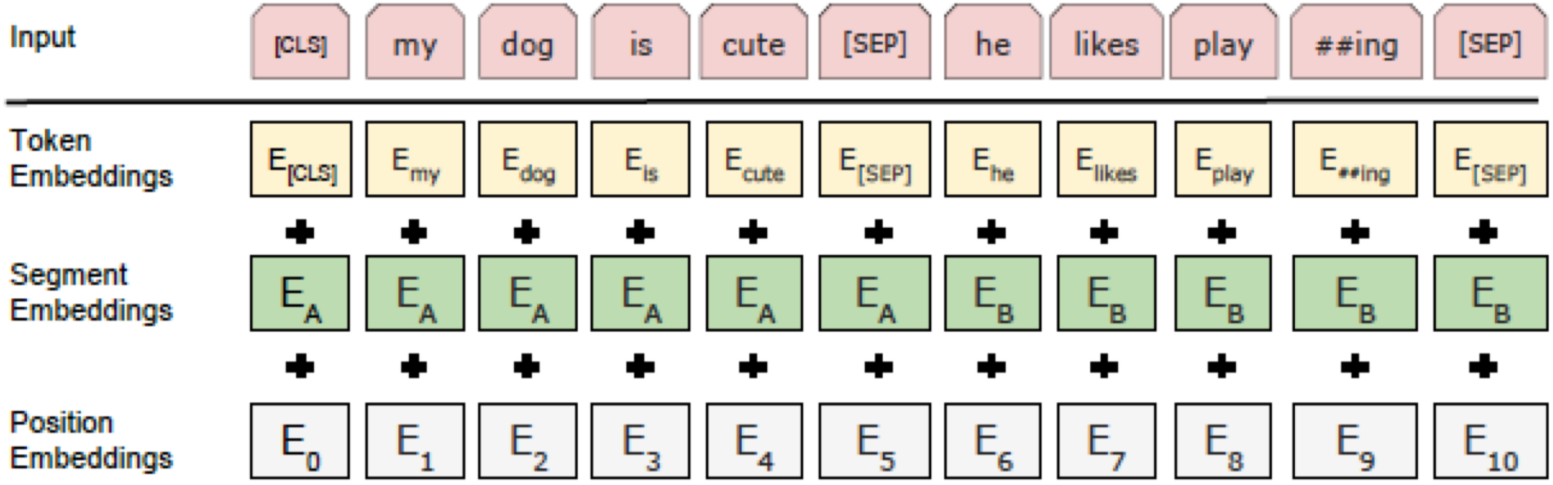

Softmax
Linear
DECODER #2
DECODER #1
Add & Normalize
Feed Forward
Feed Forward
Add & Normalize
Encoder-Decoder Attention
Add & Normalize
Self-Attention
ENCODER #2
ENCODER #1
Add & Normalize
Feed Forward
Feed Forward
Add & Normalize
Self-Attention
POSITIONAL ENCODING
x1
x2
Thinking
Machines


OpenAI
Transformer


DECODERS


Possible classes:
All English words
0.1%
Aardvark
...
...
10%
Improvisation
...
...
0%
Zyzzyva
FFNN + Softmax
12
DECODER
2
DECODER
1
DECODER
Let's
stick
to

Input Representation

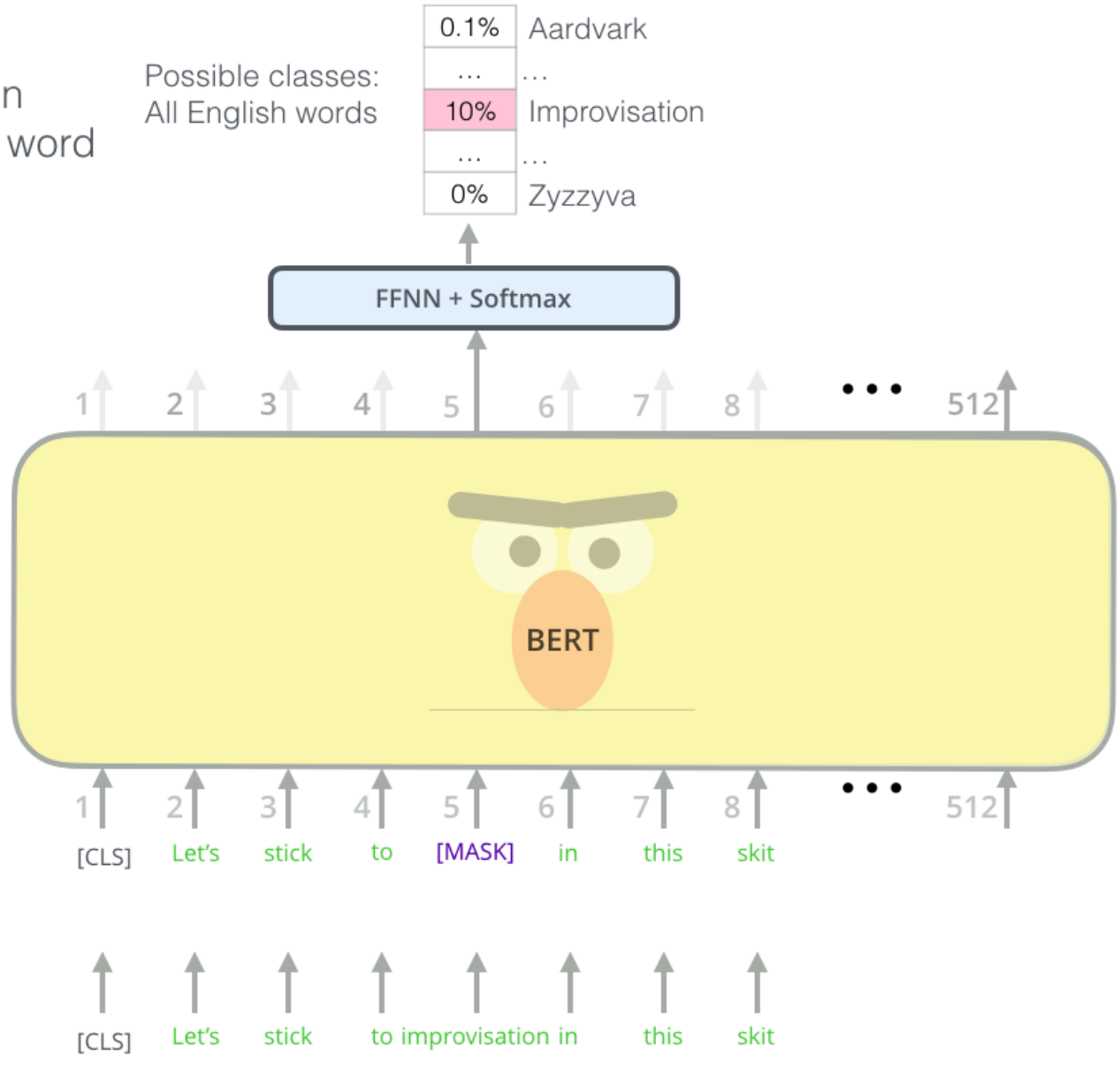


- Token Embeddings: Use pretrained WordPiece embeddings
- Position Embeddings: Use learned Position Embeddings
  - Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. Convolutionalsequence to sequence learning. arXiv preprint arXiv:1705.03122v2, 2017.
- Added sentence embedding to every tokens of each sentence
- Use [CLS] for the classification tasks
- Separate sentences by using a special token [SEP]

Task#1: Masked LM

- 15% of the words are masked at random
  - and the task is to predict the masked words based on its left and right context
- Not all tokens were masked in the same way (example sentence “My dog is hairy”)
  - 80% were replaced by the <MASK> token: “My dog is <MASK>”
  - 10% were replaced by a random token: “My dog is apple”
  - 10% were left intact: “My dog is hairy”

Use the output of the masked word's position to predict the masked word

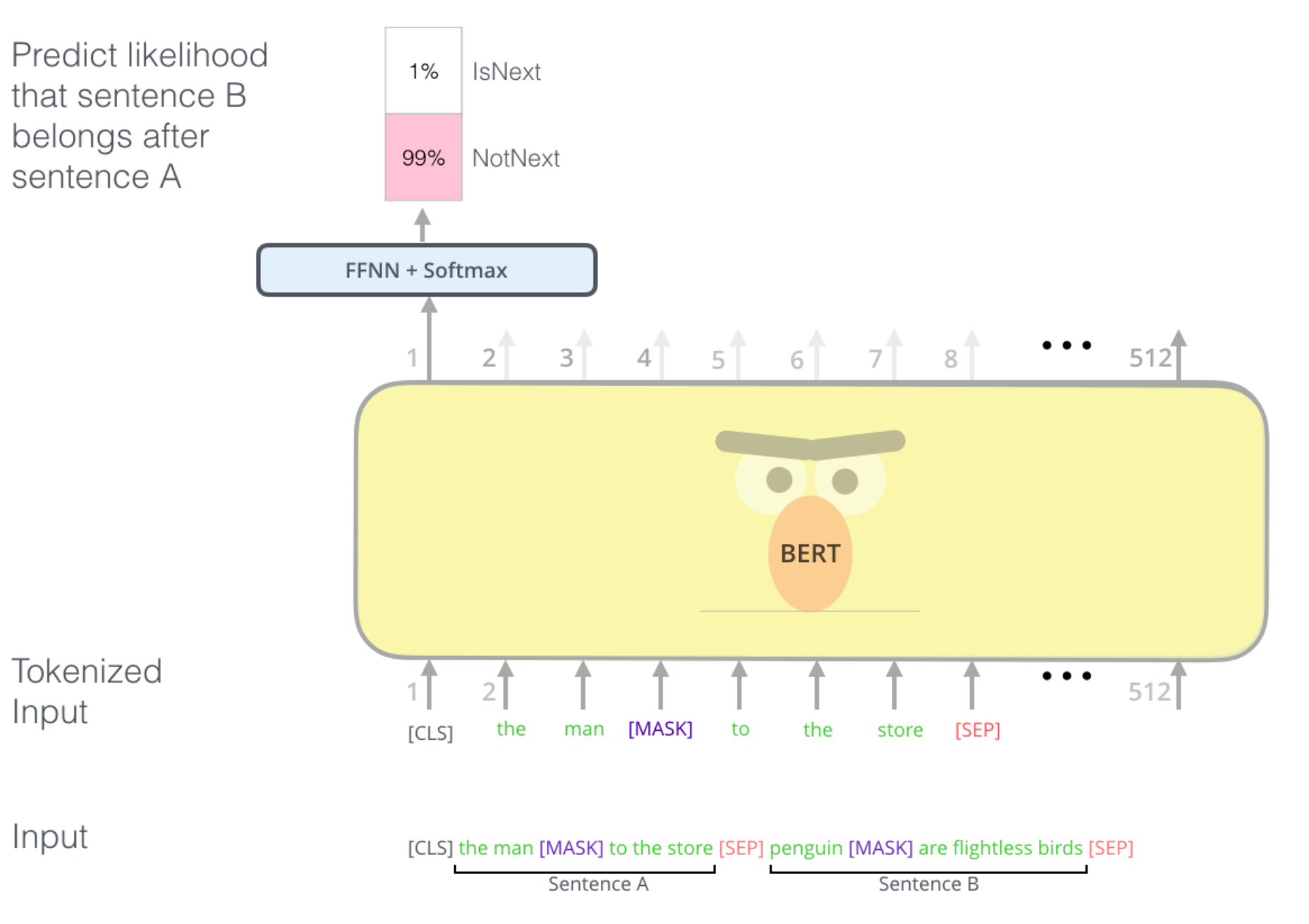


Task#2: Next Sentence Prediction

- Motivation
  - Many downstream tasks are based on understanding the relationship between two text sentences
    - Question Answering (QA) and Natural Language Inference (NLI)
  - Language modeling does not directly capture that relationship
- The task is pre-training binarized next sentence prediction task

Input = [CLS] the man went to [MASK] store [SEP] he bought a gallon [MASK] milk [SEP]
Label = isNext

Input = [CLS] the man [MASK] to the store [SEP]penguin [MASK] are flight ##less birds [SEP]
Label = NotNext


Predict likelihood that sentence B belongs after sentence A
1%
IsNext
99%
NotNext
FFNN + Softmax
1
2
3
4
5
6
7
8
512
BERT
Tokenized Input
[CLS]
the
man
[MASK]
to
the
store
[SEP]
Input
[CLS] the man [MASK] to the store [SEP] penguin [MASK] are flightless birds [SEP]
Sentence A
Sentence B

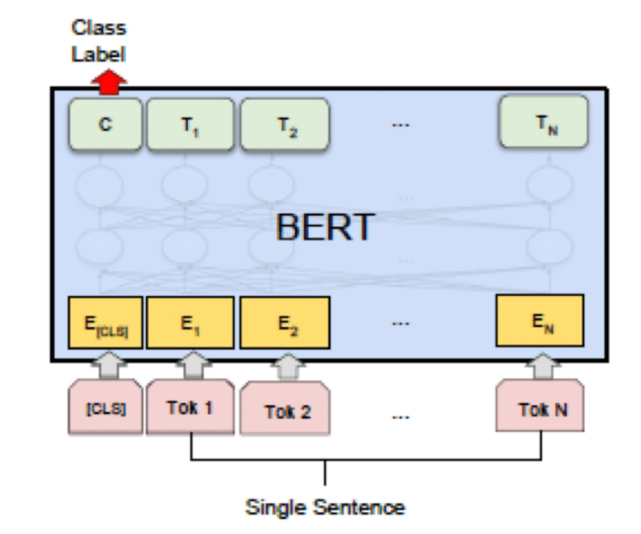
Pre-training procedure

- Training data: BooksCorpus (800M words) + English Wikipedia (2,500M words)
- To generate each training input sequences: sample two spans of text (A and B) from the corpus
  - The combined length is ≤ 500 tokens
  - 50% B is the actual next sentence that follows A and 50% of the time it is a random sentence from the corpus
- The training loss is the sum of the mean masked LM likelihood and the mean next sentence prediction likelihood

Fine-tuning procedure

- For sequence-level classification task
  - Obtain the representation of the input sequence by using the final hidden state (hidden state at the position of the special token [CLS]) $C \in R^H$
  - Just add a classification layer and use softmax to calculate label probabilities. Parameters W $\in R^{K \times H}$

$$P = \text{softmax}(CW^T)$$



Fine-tuning procedure

- For sequence-level classification task
  - All of the parameters of BERT and *W* are fine-tuned jointly
- Most model hyperparameters are the same as in pre-training
  - except the batch size, learning rate, and number of training epochs

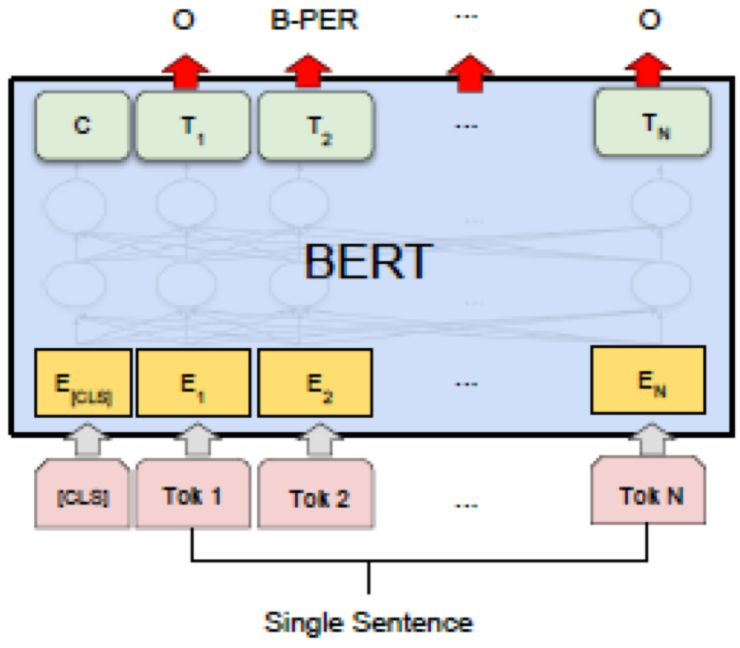
Fine-tuning procedure

- Token tagging task (e.g., Named Entity Recognition)
  - Feed the final hidden representation $T_i \in R^H$ for each token i into a classification layer for the tagset (NER label set)
  - To make the task compatible with WordPiece tokenization

```
Jim   Hen   ##son was a puppet ##eer
I-PER I-PER X     O   O O      X
```

    - Predict the tag for the first sub-token of a word
    - No prediction is made for X

Fine-tuning procedure



Single Sentence Tagging Tasks: CoNLL-2003 NER

Fine-tuning procedure

- Span-level task: SQuAD v1.1
  - Input Question:
    `Where do water droplets collide with ice crystals to form precipitation?`
  - Input Paragraph:
    `.... Precipitation forms as smaller dropletscoalesce via collision with other rain dropsor ice crystals within a cloud. ...`
  - Output Answer:
    `within a cloud`

Fine-tuning procedure

- Span-level task: SQuAD v1.1
  - Represent the input question and paragraph as a single packed sequence
    - The question uses the A embedding and the paragraph uses the B embedding
  - New parameters to be learned in fine-tuning are start vector $S \in \mathbb{R}^H$ and end vector $E \in \mathbb{R}^H$
  - Calculate the probability of word i being the start of the answer span

$$P_i = \frac{e^{S \cdot T_i}}{\sum_j e^{S \cdot T_j}}$$

  - The training objective is the log-likelihood the correct and end positions

Comparison of BERT and OpenAI GPT

OpenAI GPT	BERT
Trained on BooksCorpus (800M)	Trained on BooksCorpus (800M) + Wikipedia (2,500M)
Use sentence separater ([SEP]) and classifier token ([CLS]) only at fine-tuning time	BERT learns [SEP], [CLS] and sentence A/B embeddings during pre-training
Trained for 1M steps with a batch-size of 32,000 words	Trained for 1M steps with a batch-size of 128,000 words
Use the same learning rate of 5e-5 for all fine-tuning experiments	BERT choose a task-specific learning rate which performs the best on the development set

Outline

- Research context
- Main ideas
- BERT
- **Experiments**
- Conclusions

Experiments

- GLUE (General Language Understanding Evaluation) benchmark
  - Distribute canonical Train, Dev and Test splits
  - Labels for Test set are not provided
- Datasets in GLUE:
  - MNLI: Multi-Genre Natural Language Inference
  - QQP: Quora Question Pairs
  - QNLI: Question Natural Language Inference
  - SST-2: Stanford Sentiment Treebank
  - CoLA: The corpus of Linguistic Acceptability
  - STS-B: The Semantic Textual Similarity Benchmark
  - MRPC: Microsoft Research Paraphrase Corpus
  - RTE: Recognizing Textual Entailment
  - WNLI: Winograd NLI

GLUE Results

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	79.9	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	88.1	91.3	45.4	80.0	82.3	56.0	75.2
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.1	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**91.1**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**81.9**

Table 1: GLUE Test results, scored by the GLUE evaluation server. The number below each task denotes the number of training examples. The "Average" column is slightly different than the official GLUE score, since we exclude the problematic WNLI set. OpenAI GPT = (L=12, H=768, A=12); $\text{BERT}_{\text{BASE}}$ = (L=12, H=768, A=12); $\text{BERT}_{\text{LARGE}}$ = (L=24, H=1024, A=16). BERT and OpenAI GPT are single-model, single task. All results obtained from https://gluebenchmark.com/leaderboard and https://blog.openai.com/language-unsupervised/.

SQuAD v1.1

System	Dev		Test	
	EM	F1	EM	F1
Leaderboard (Oct 8th, 2018)				
Human	-	-	82.3	91.2
#1 Ensemble - nlnet	-	-	86.0	91.7
#2 Ensemble - QANet	-	-	84.5	90.5
#1 Single - nlnet	-	-	83.5	90.1
#2 Single - QANet	-	-	82.5	89.3
Published				
BiDAF+ELMo (Single)	-	85.8	-	-
R.M. Reader (Single)	78.9	86.3	79.5	86.6
R.M. Reader (Ensemble)	81.2	87.9	82.3	88.5
Ours				
$\text{BERT}_{\text{BASE}}$ (Single)	80.8	88.5	-	-
$\text{BERT}_{\text{LARGE}}$ (Single)	84.1	90.9	-	-
$\text{BERT}_{\text{LARGE}}$ (Ensemble)	85.8	91.8	-	-
$\text{BERT}_{\text{LARGE}}$ (Sgl.+TriviaQA)	**84.2**	**91.1**	**85.1**	**91.8**
$\text{BERT}_{\text{LARGE}}$ (Ens.+TriviaQA)	**86.2**	**92.2**	**87.4**	**93.2**

Table 2: SQuAD results. The BERT ensemble is 7x systems which use different pre-training checkpoints and fine-tuning seeds.

Reference: https://rajpurkar.github.io/SQuAD-explorer

Named Entity Recognition

System	Dev F1	Test F1
ELMo+BiLSTM+CRF	95.7	92.2
CVT+Multi (Clark et al., 2018)	-	92.6
$\text{BERT}_{\text{BASE}}$	96.4	92.4
$\text{BERT}_{\text{LARGE}}$	**96.6**	**92.8**

Table 3: CoNLL-2003 Named Entity Recognition results. The hyperparameters were selected using the Dev set, and the reported Dev and Test scores are averaged over 5 random restarts using those hyperparameters.

SWAG

- The Situations with Adversarial Generations (SWAG)

A girl is going across a set of monkey bars. She

(i) jumps up across the monkey bars.

(ii) struggles onto the bars to grab her head.

(iii) gets to the end and stands on a wooden plank.

(iv) jumps up and does a back flip.

- The only task-specific parameters is a vector $V \in \mathbb{R}^H$
- The probability distribution is the softmax over the four choices

$$P_i = \frac{e^{V \cdot C_i}}{\sum_{j=1}^{4} e^{V \cdot C_i}}$$

SWAG Result

System	Dev	Test
ESIM+GloVe	51.9	52.7
ESIM+ELMo	59.1	59.2
$\text{BERT}_{\text{BASE}}$	81.6	-
$\text{BERT}_{\text{LARGE}}$	**86.6**	**86.3**
Human (expert)†	-	85.0
Human (5 annotations)†	-	88.0

Table 4: SWAG Dev and Test accuracies. Test results were scored against the hidden labels by the SWAG authors. †Human performance is measure with 100 samples, as reported in the SWAG paper.

Ablation Studies

- To understand
  - Effect of Pre-training Tasks
  - Effect of model sizes
  - Effect of number of training steps
  - Feature-based approach with BERT

Ablation Studies

- Main findings:
  - "Next sentence prediction" (NSP) pre-training is important for sentence-pair classification task
  - Bidirectionality (using MLM pre-training task) contributes significantly to the performance improvement

Tasks	Dev Set MNLI-m (Acc)	QNLI (Acc)	MRPC (Acc)	SST-2 (Acc)	SQuAD (F1)
$\text{BERT}_{\text{BASE}}$	84.4	88.4	86.7	92.7	88.5
No NSP	83.9	84.9	86.5	92.6	87.9
LTR & No NSP	82.1	84.3	77.5	92.1	77.8
+ BiLSTM	82.1	84.1	75.7	91.6	84.9

Table 5: Ablation over the pre-training tasks using the $\text{BERT}_{\text{BASE}}$ architecture. "No NSP" is trained without the next sentence prediction task. "LTR & No NSP" is trained as a left-to-right LM without the next sentence prediction, like OpenAI GPT. "+ BiLSTM" adds a randomly initialized BiLSTM on top of the "LTR + No NSP" model during fine-tuning.

Ablation Studies

- Main findings
  - Bigger model sizes are better even for small-scale tasks

Hyperparams				Dev Set Accuracy		
#L	#H	#A	LM (ppl)	MNLI-m	MRPC	SST-2
3	768	12	5.84	77.9	79.8	88.4
6	768	3	5.24	80.6	82.2	90.7
6	768	12	4.68	81.9	84.8	91.3
12	768	12	3.99	84.4	86.7	92.9
12	1024	16	3.54	85.7	86.9	93.3
24	1024	16	3.23	86.6	87.8	93.7

Table 6: Ablation over BERT model size. #L = the number of layers; #H = hidden size; #A = number of attention heads. "LM (ppl)" is the masked LM perplexity of held-out training data.

Conclusions

- Unsupervised pre-training (pre-training language model) is increasingly adopted in many NLP tasks
- Major contribution of the paper is to propose a deep *bidirectional* architecture from Transformer
  - Advance state-of-the-art for many important NLP tasks

Links

- TensorFlow code and pre-trained models for BERT: https://github.com/google-research/bert
- PyTorch Pretrained Bert: https://github.com/huggingface/pytorch-pretrained-BERT
- BERT-pytorch: https://github.com/codertimo/BERT-pytorch
- BERT-keras: https://github.com/Separius/BERT-keras

Remark: Applying BERT for non-English languages

- Pre-trained BERT models are provided for more than 100 languages (including Vietnamese)
  - https://github.com/google-research/bert/blob/master/multilingual.md
- Be careful with tokenization!!
- For Japanese (and Chinese): “spaces were added around every character in the CJI Unicode rage before applying WordPiece” => Not a good way to do
  - Use SentencePiece: https://github.com/google/sentencepiece
  - We may need to pre-train BERT model

References

1. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.
2. Vaswani et al. (2017). Attention Is All You Need. arXiv preprint arXiv:1706.03762. https://arxiv.org/abs/1706.03762
3. The Annotated Transformer: http://nlp.seas.harvard.edu/2018/04/03/attention.html, by harvardnlp.
4. The Illustrated Transformer: http://jalammar.github.io/illustrated-transformer/
5. ELMo explained: https://www.mihaileric.com/posts/deep-contextualized-word-representations-elmo/
6. ULMFit explained: https://yashuseth.blog/2018/06/17/understanding-universal-language-model-fine-tuning-ulmfit/
7. Dissecting BERT: https://medium.com/dissecting-bert
8. The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning): http://jalammar.github.io/illustrated-bert/