New Architecture

TA: 紀伯翰

Review

Review

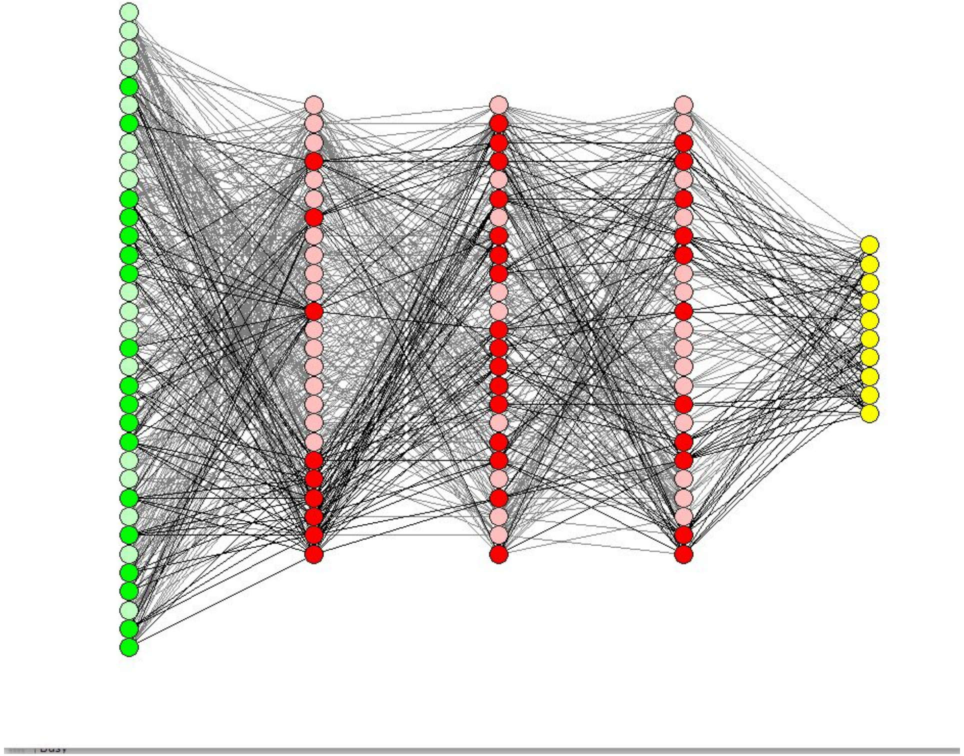
Fully Connected Network

Convolutional Neural Network

Recurrent Neural Network

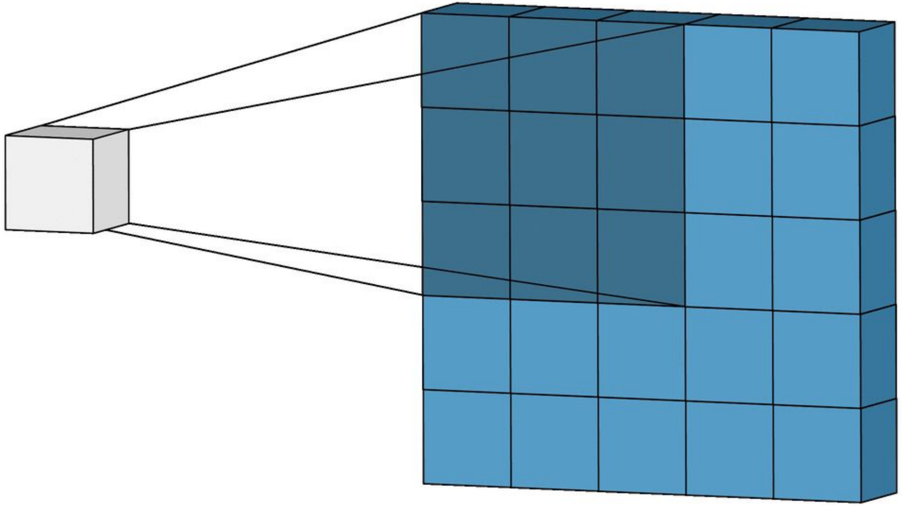
Basic Module - (1)

Fully Connected Network
Lots of Model ….



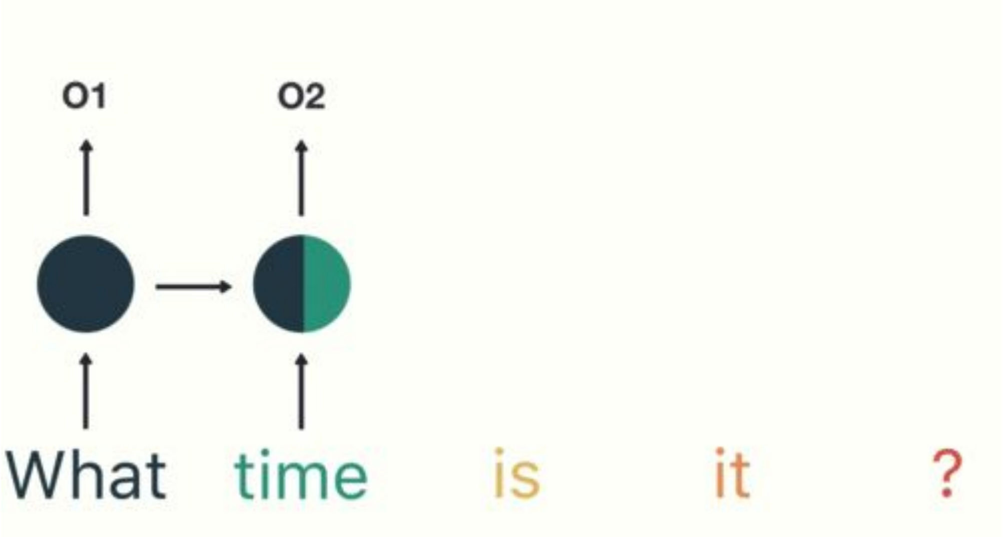
Basic Module - (2)

Convolutional Neural Network
ResNet
DenseNet
Inception Network

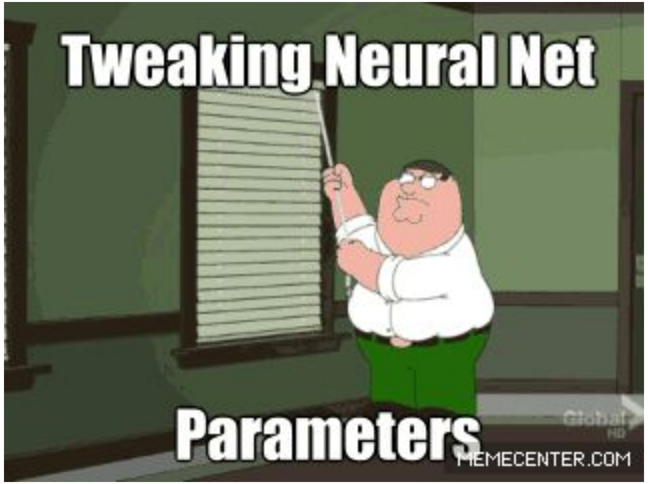


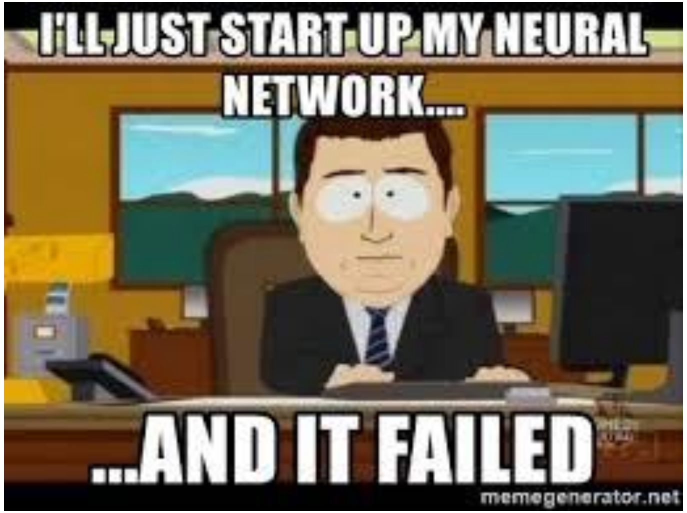
Basic Module - (3)

Recurrent Neural Network
Seq2seq
LSTM
Pointer Network ...



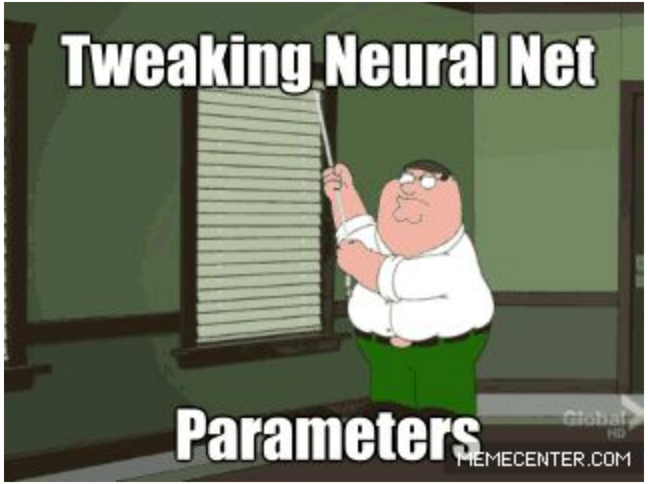
Stack them and hope the new model will be better !

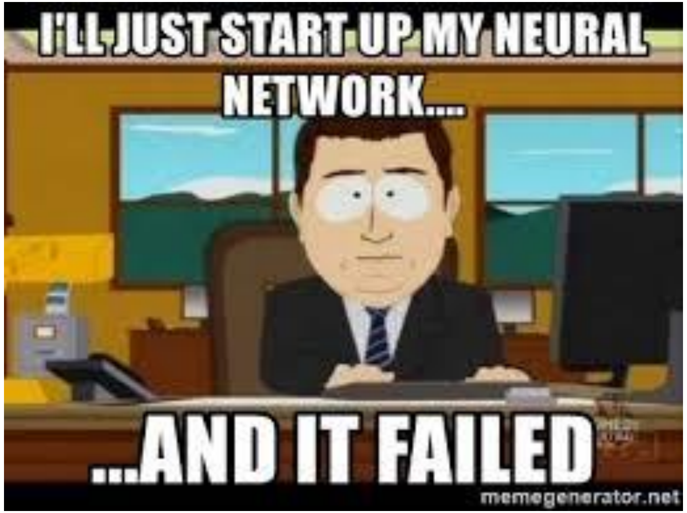

Tweaking Neural Net
Parameters
Global HD
MEMECENTER.COM


I'LL JUST START UP MY NEURAL NETWORK....
...AND IT FAILED
memegenerator.net

THAT'S NOT ENOUGH
WE HAVE TO GO DEEPER
quickmeme.com

LET'S GET STARTED


Tweaking Neural Net
Parameters
Global HD
MEMECENTER.COM


I'LL JUST START UP MY NEURAL NETWORK....
...AND IT FAILED
memegenerator.net

End

Follow up SOTA structure

Stable and explore faster

Why New Architecture ?

1. Increase Performance !
2. Extract better feature from data
3. Generalization
4. Reduce Parameters or explainable

Today’ s New Architecture

The variant structure design from the old module in 2019.

The cool application of architecture in 2019.

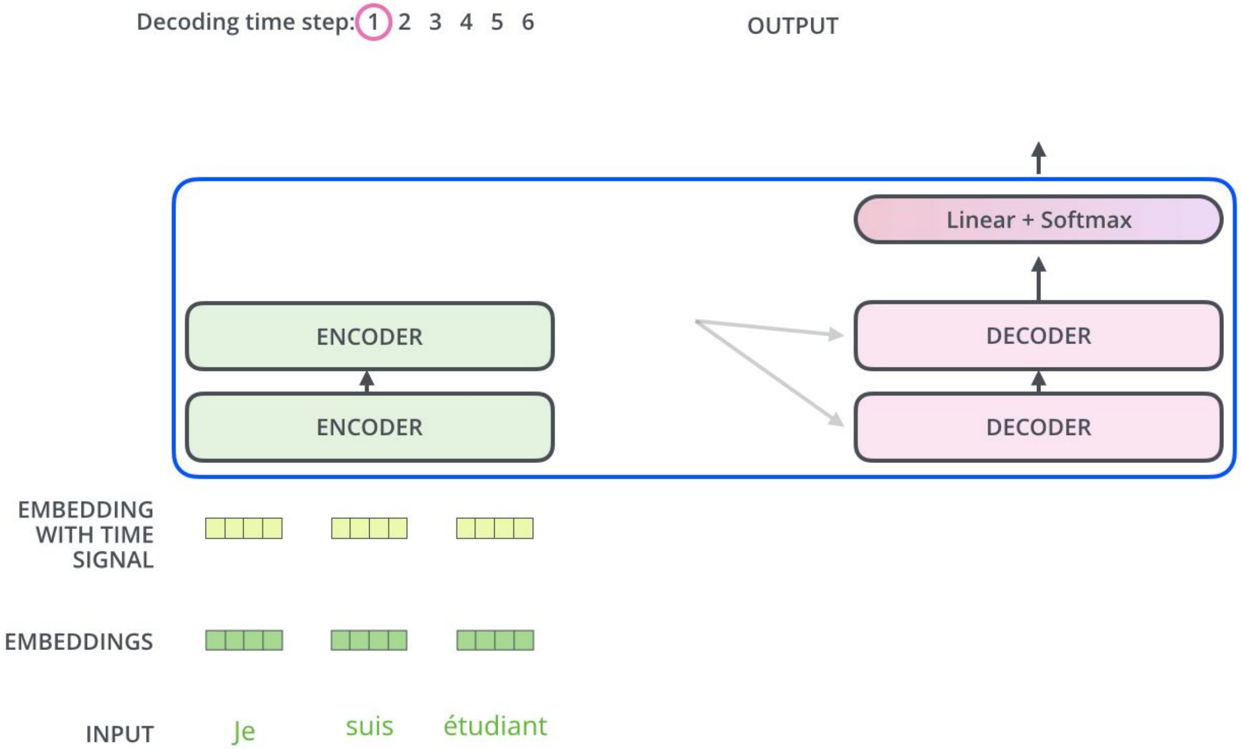
The SOTA models in the trend for 2019.

Outline

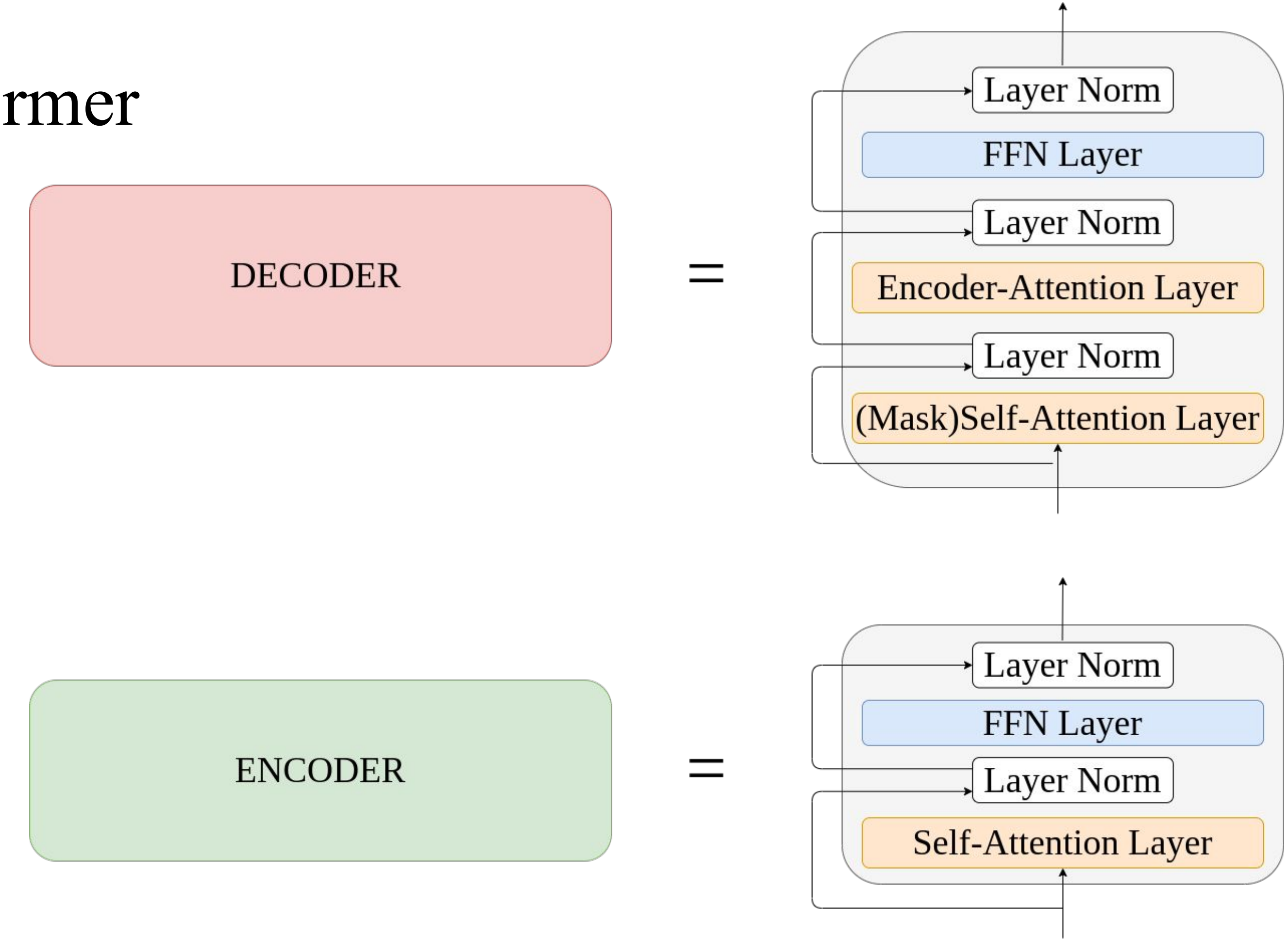
- RNN less structure
  - Transformer
    - Sandwich transformers
    - Universal Transformer
  - Residual Shuffle Exchange Network
  - BERT
    - ALBERT, Reformer
- StyleGAN

Transformer

Transformer - Encoding & Decoding



Transformer



DECODER

=

Layer Norm

FFN Layer

Layer Norm

Encoder-Attention Layer

Layer Norm

(Mask)Self-Attention Layer

ENCODER

=

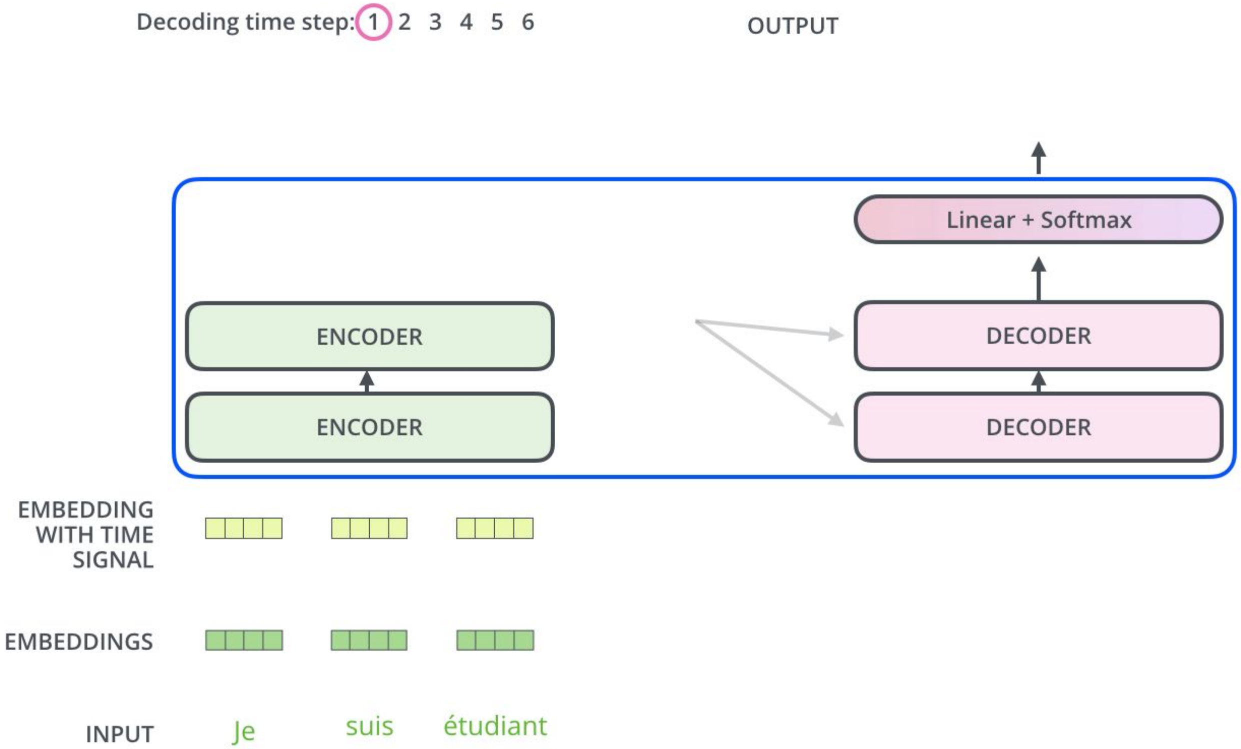
Layer Norm

FFN Layer

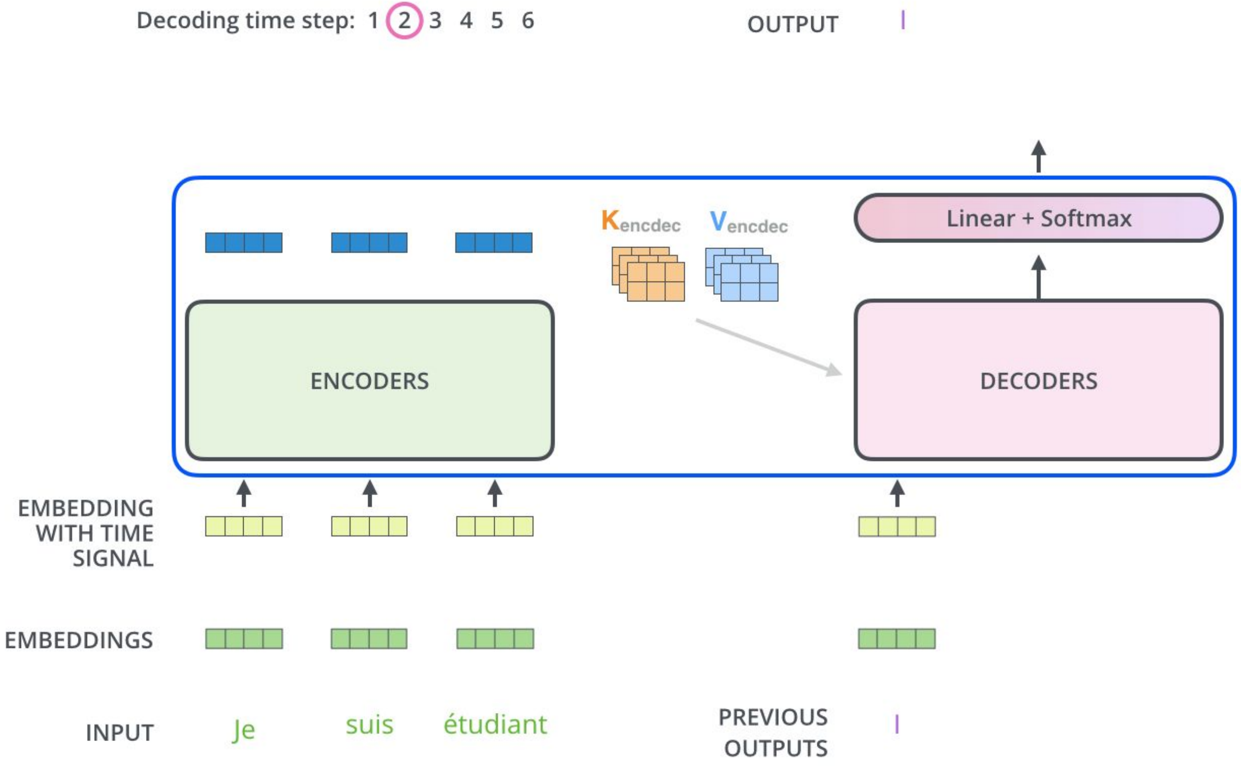
Layer Norm

Self-Attention Layer

Transformer - Encoding & Decoding



Transformer - Decoding



Sandwich Transformers

Motivation

Designing a Better Transformer

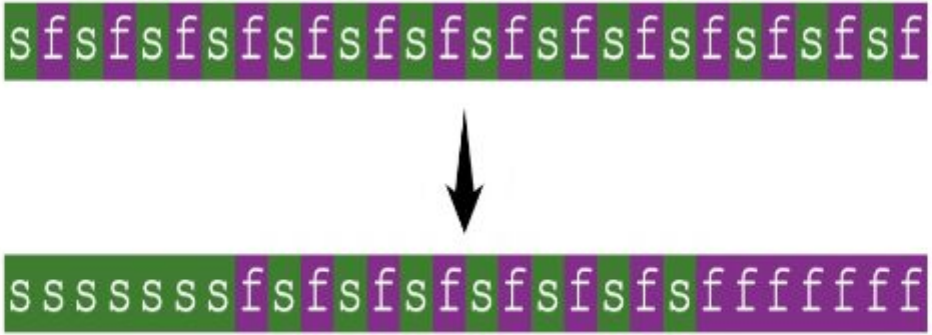
Reorder the sublayer ?

Could we increase the performance just by reorder the sublayer module ?

Highlight

1. models with more self-attention toward the bottom and more feedforward sublayers toward the top tend to perform better in general.
2. No extra parameters, memory requirement.

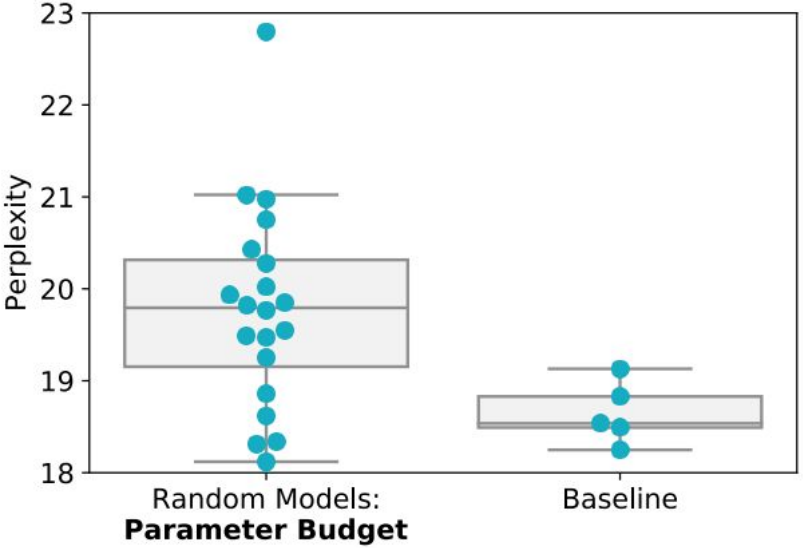
Sandwich transformer

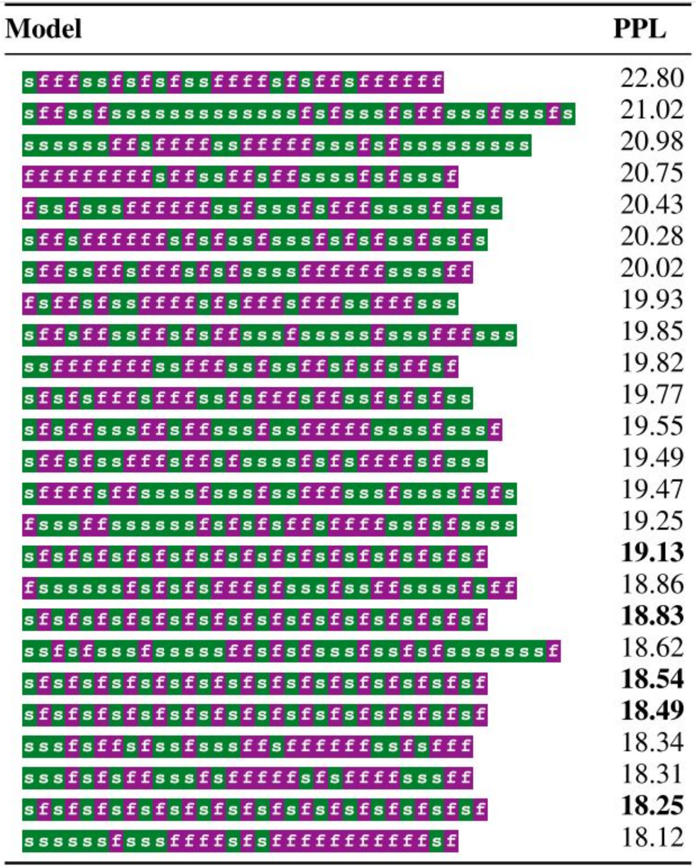


sfsfsfsfsfsfsfsfsfsfsfsfsfsf

sssssssfsfsfsfsfsfsfsfffffff

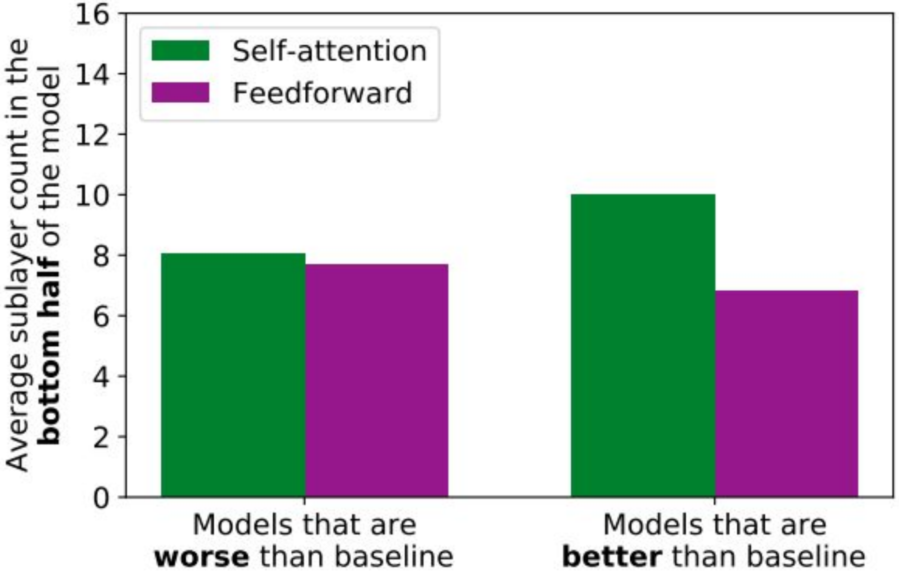
Same parameters but different order

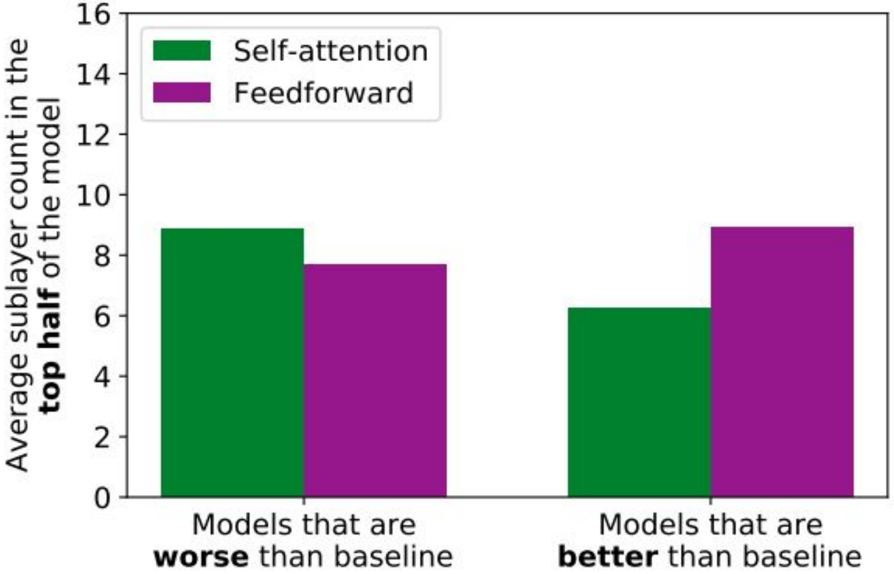




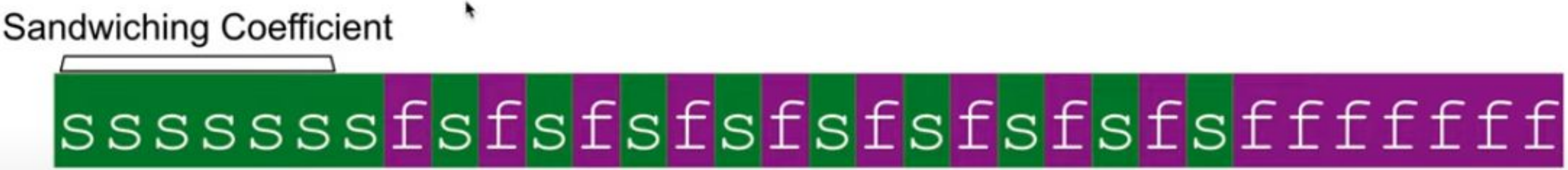
Model	PPL
sfffssfsfsfssffffsfsffsffffff	22.80
sffssfsssssssssssssfsfsssfsffsssfsssfs	21.02
ssssssffsffffssfffffsssfsfsssssssss	20.98
fffffffffsffssffsffssssfsfsssf	20.75
fssfsssffffffssfsssfsfffssssfsfss	20.43
sffsffffffsfsfssfsssfsfsfssfssfs	20.28
sffssffsfffsfsfssssffffffssssff	20.02
fsffsfssffffsfsfffsfffssfffsss	19.93
sffsffssffsfsffsssfsssssfsssfffsss	19.85
ssfffffffssfffssfssffsfsfsffsf	19.82
sfsfsfffsfffssfsfffsffssfsfsfss	19.77
sfsffsssffsffsssfssfffffssssfsssf	19.55
sffsfssfffsffsfssssfsfsffffsfsss	19.49
sffffsffssssfsssfssfffsssfssssfsfs	19.47
fsssffssssssfsfsfsffsffffssfsfssss	19.25
sfsfsfsfsfsfsfsfsfsfsfsfsfsfsfsf	**19.13**
fssssssfsfsfsfffsfsssfssffssssfsff	18.86
sfsfsfsfsfsfsfsfsfsfsfsfsfsfsfsf	**18.83**
ssfsfsssfsssssffsfsfsssfssfsfsssssssf	18.62
sfsfsfsfsfsfsfsfsfsfsfsfsfsfsfsf	**18.54**
sfsfsfsfsfsfsfsfsfsfsfsfsfsfsfsf	**18.49**
sssfsffsfssfsssffsffffffssfsfff	18.34
sssfsfsffsssfsfffffsfsffffsssff	18.31
sfsfsfsfsfsfsfsfsfsfsfsfsfsfsfsf	**18.25**
ssssssfsssffffsfsfffffffffffsf	18.12

Are Balanced Architectures Better?



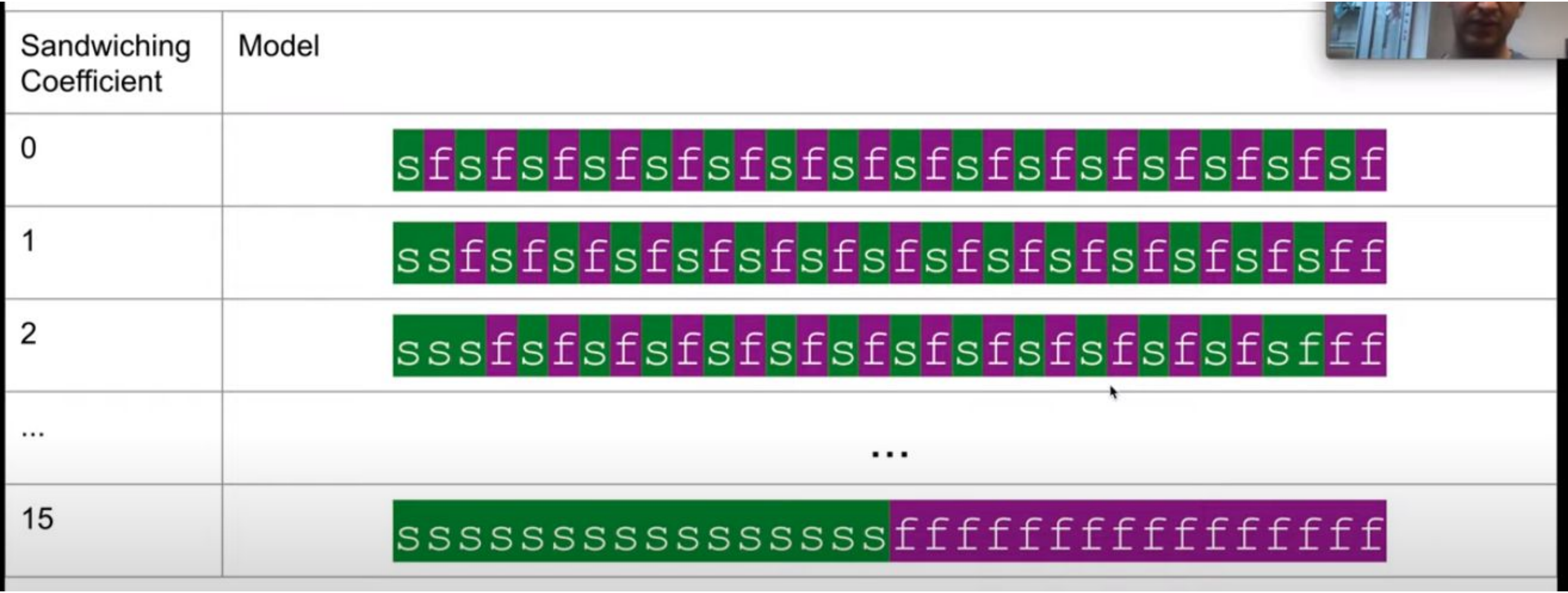


Sandwich Coefficient

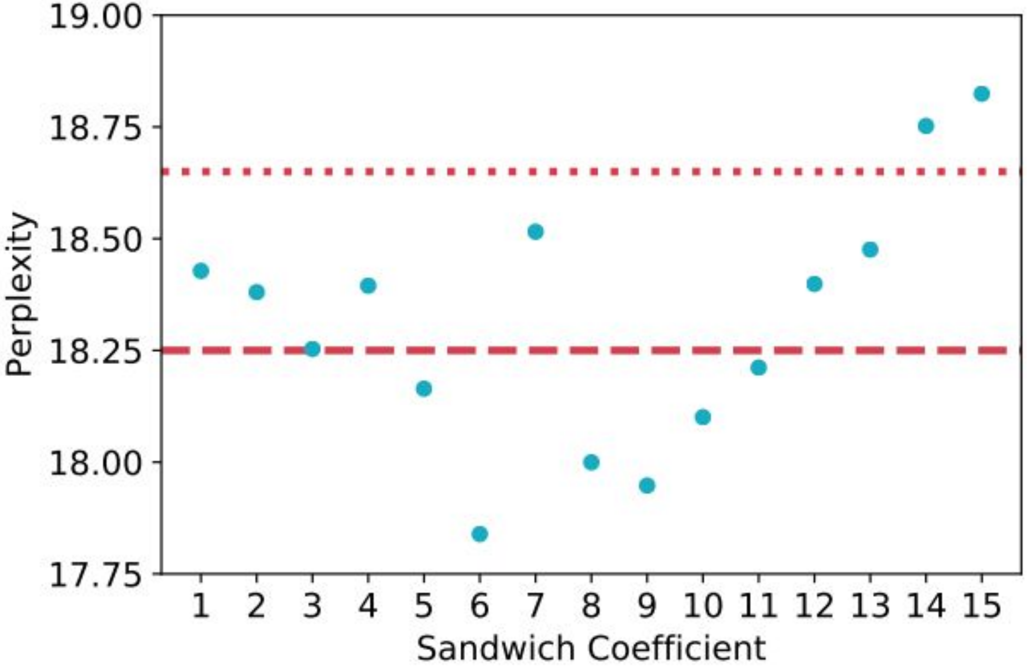


Sandwiching Coefficient

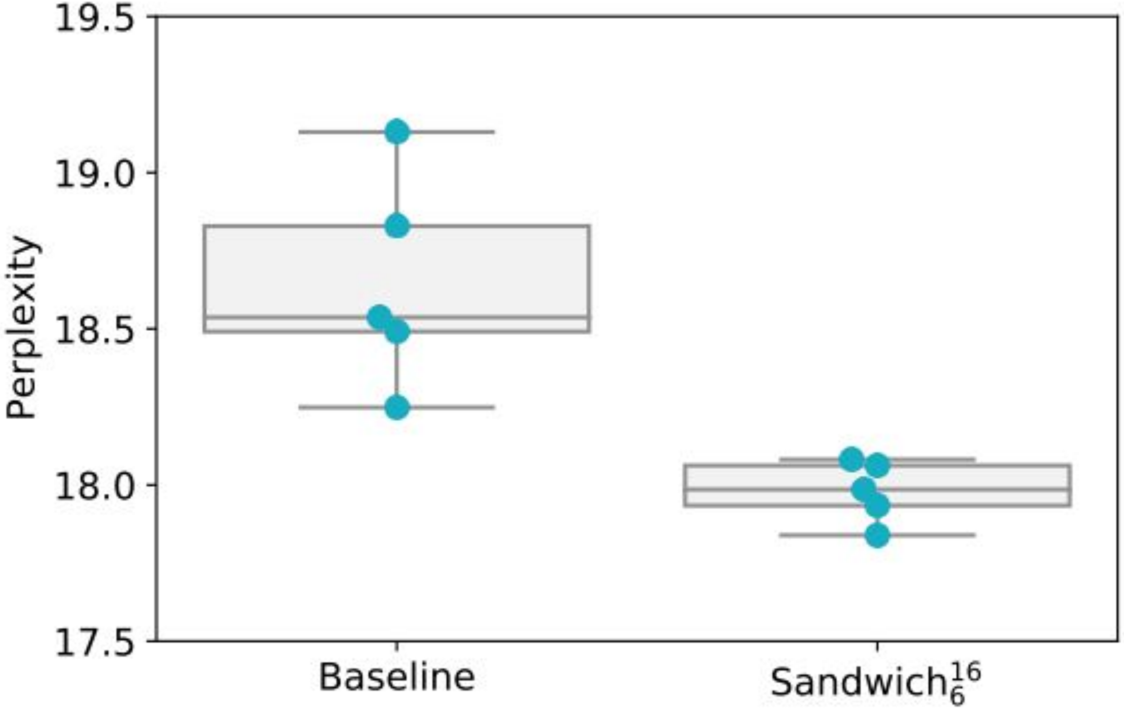
s s s s s s s f s f s f s f s f s f s f s f s f s f s f s f f f f f f f



Sandwiching Coefficient	Model
0	sfsfsfsfsfsfsfsfsfsfsfsfsfsfsfsf
1	ssfsfsfsfsfsfsfsfsfsfsfsfsfsfsff
2	sssfsfsfsfsfsfsfsfsfsfsfsfsfsfff
...	...
15	ssssssssssssssssffffffffffffffff


19.00
18.75
18.50
18.25
18.00
17.75
Perplexity
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Sandwich Coefficient

Experiment


19.5
19.0
18.5
18.0
17.5
Perplexity
Baseline
$\mathrm{Sandwich}_6^{16}$

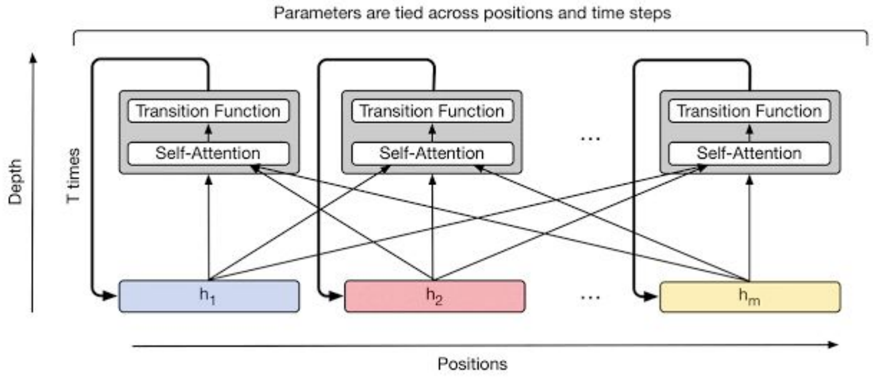
Universal Transformer

Motivation

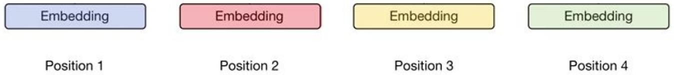
Transformer: Translation Good! / algorithmic task bad !
Neural GPU: Translation Bad! / algorithmic task Good!
Neural Turing Machine: Translation Bad! / algorithmic task Good!

Universal Transformer born !

Universal Transformer



Dynamic Halting



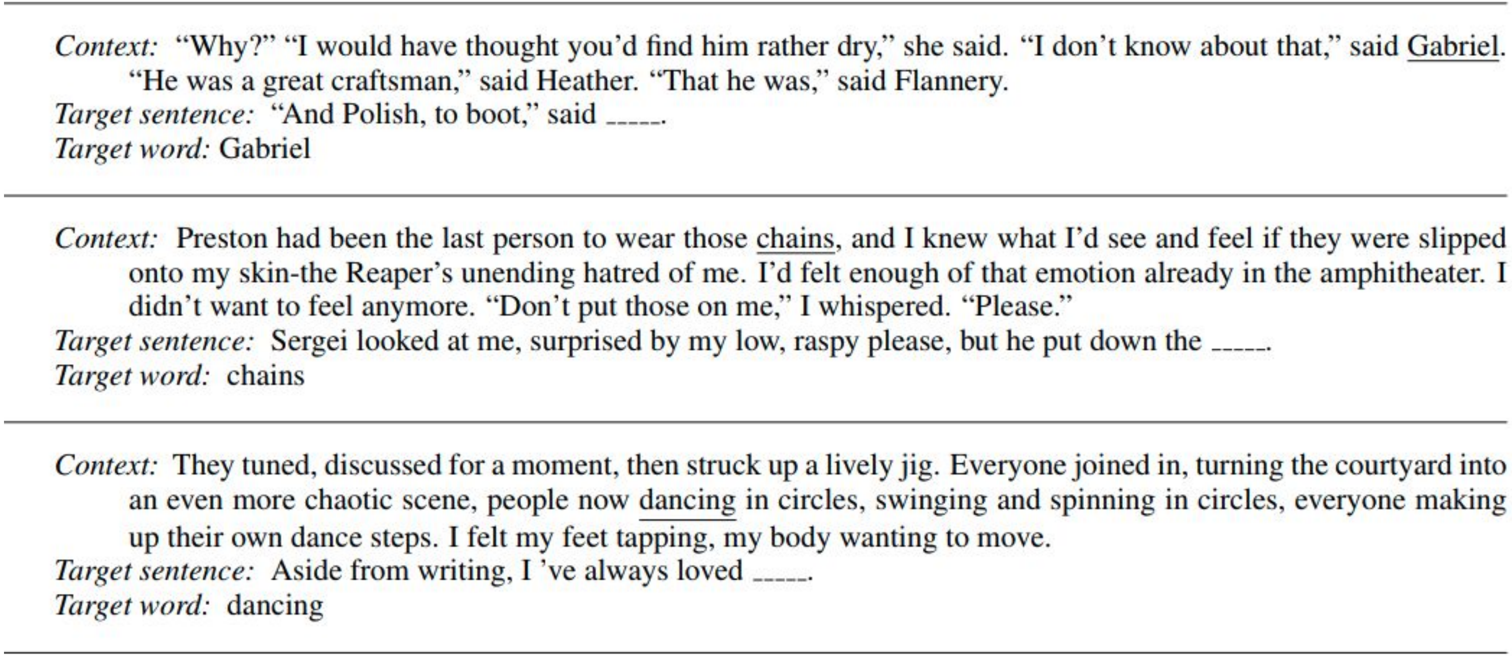
Embedding Embedding Embedding Embedding

Position 1 Position 2 Position 3 Position 4

Experiment(s)

1. Lambda Question Answering Dataset
2. WMT 14 En-De translation task

Lambda Question Answering Dataset

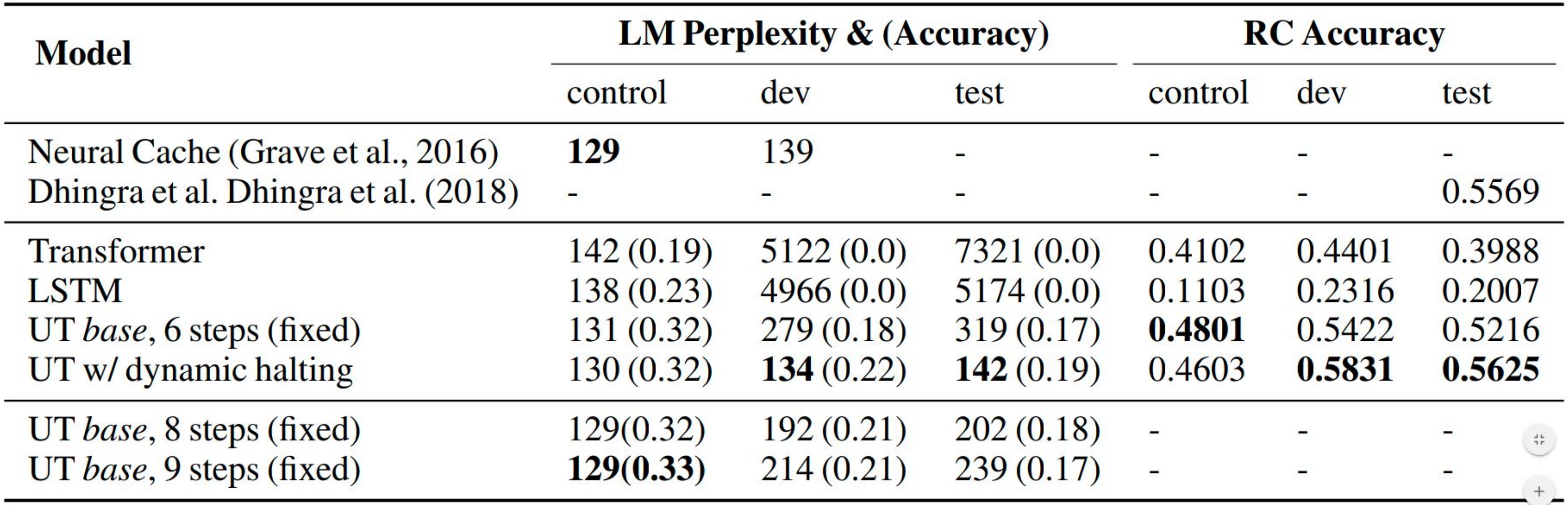


Context: “Why?” “I would have thought you’d find him rather dry,” she said. “I don’t know about that,” said Gabriel. “He was a great craftsman,” said Heather. “That he was,” said Flannery.
Target sentence: “And Polish, to boot,” said _____.
Target word: Gabriel

Context: Preston had been the last person to wear those chains, and I knew what I’d see and feel if they were slipped onto my skin-the Reaper’s unending hatred of me. I’d felt enough of that emotion already in the amphitheater. I didn’t want to feel anymore. “Don’t put those on me,” I whispered. “Please.”
Target sentence: Sergei looked at me, surprised by my low, raspy please, but he put down the _____.
Target word: chains

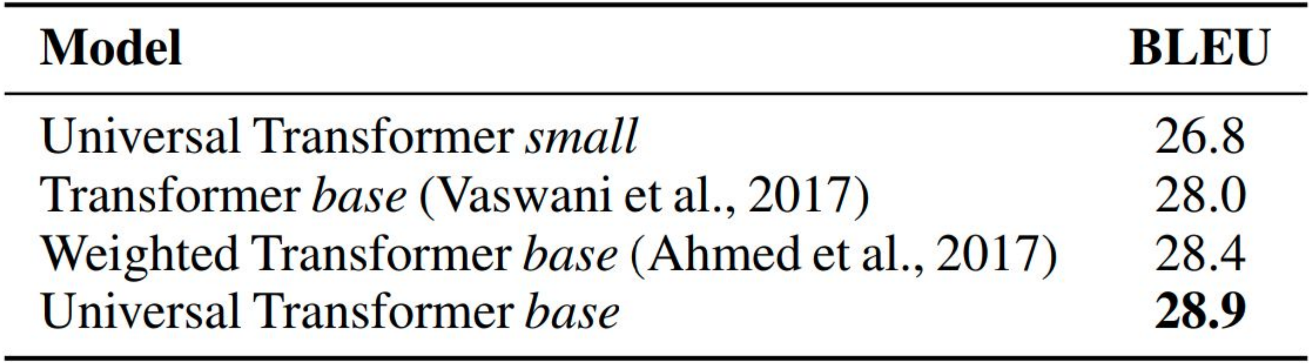
Context: They tuned, discussed for a moment, then struck up a lively jig. Everyone joined in, turning the courtyard into an even more chaotic scene, people now dancing in circles, swinging and spinning in circles, everyone making up their own dance steps. I felt my feet tapping, my body wanting to move.
Target sentence: Aside from writing, I ’ve always loved _____.
Target word: dancing

Result(s)



Model	LM Perplexity & (Accuracy)			RC Accuracy		
	control	dev	test	control	dev	test
Neural Cache (Grave et al., 2016)	**129**	139	-	-	-	-
Dhingra et al. Dhingra et al. (2018)	-	-	-	-	-	0.5569
Transformer	142 (0.19)	5122 (0.0)	7321 (0.0)	0.4102	0.4401	0.3988
LSTM	138 (0.23)	4966 (0.0)	5174 (0.0)	0.1103	0.2316	0.2007
UT *base*, 6 steps (fixed)	131 (0.32)	279 (0.18)	319 (0.17)	**0.4801**	0.5422	0.5216
UT w/ dynamic halting	130 (0.32)	**134** (0.22)	**142** (0.19)	0.4603	**0.5831**	**0.5625**
UT *base*, 8 steps (fixed)	129(0.32)	192 (0.21)	202 (0.18)	-	-	-
UT *base*, 9 steps (fixed)	**129(0.33)**	214 (0.21)	239 (0.17)	-	-	-

WMT 14 En-De translation task



Model	BLEU
Universal Transformer *small*	26.8
Transformer *base* (Vaswani et al., 2017)	28.0
Weighted Transformer *base* (Ahmed et al., 2017)	28.4
Universal Transformer *base*	**28.9**

Residual Shuffle Exchange Network

Residual Shuffle Exchange Network

HighLight:

1. Less parameters compare to other models for the same tasks.
2. Sequence processing in O(n log n) Time, specialize application on long sequence.
3. Shuffle & Exchange operators capture distant informations replace attention .

Shuffle-Exchange Network

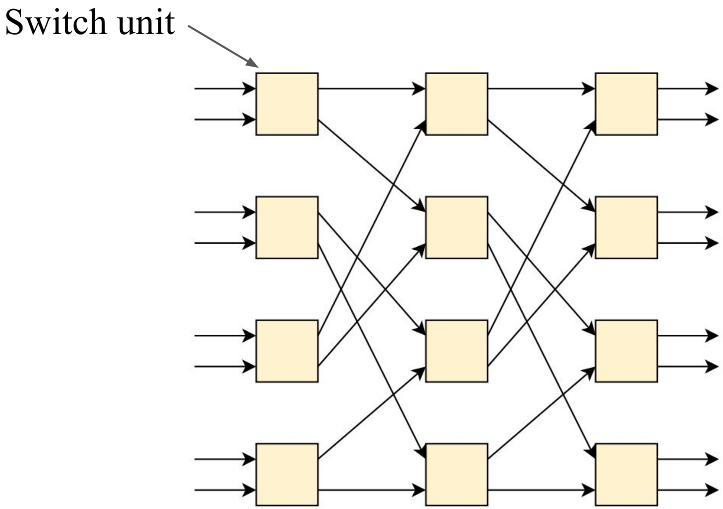


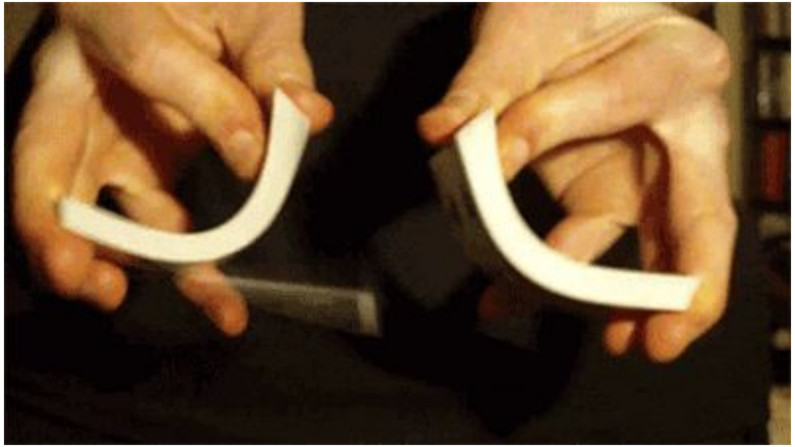
Figure 1: Shuffle-Exchange network.

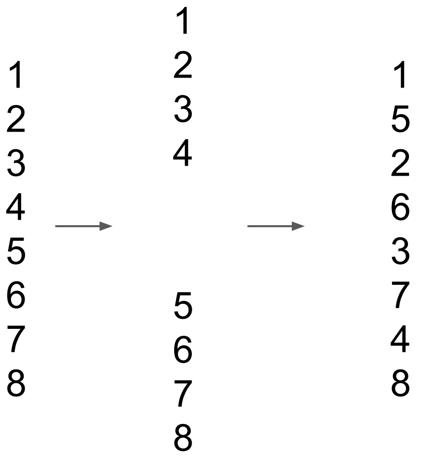
Shuffle & Exchange

Where Shuffle ? Where Exchange ?

Shuffle, Exchange

Perfect shuffle:





1
2
3
4
5
6
7
8

1
2
3
4

5
6
7
8

1
5
2
6
3
7
4
8

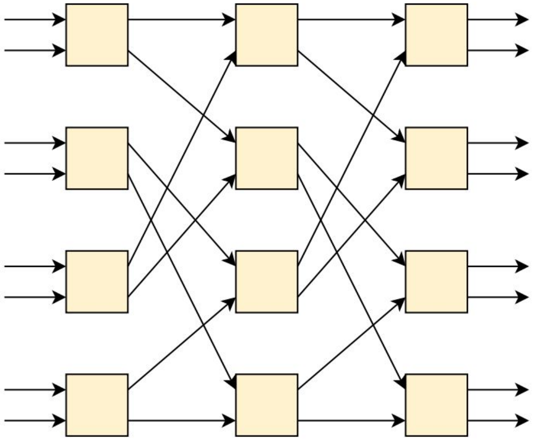
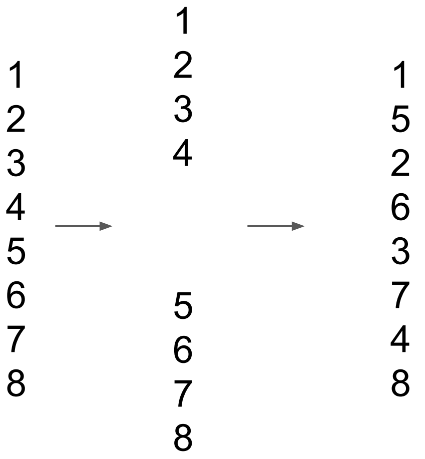


Figure 1: Shuffle-Exchange network.



1
2
3
4
5
6
7
8

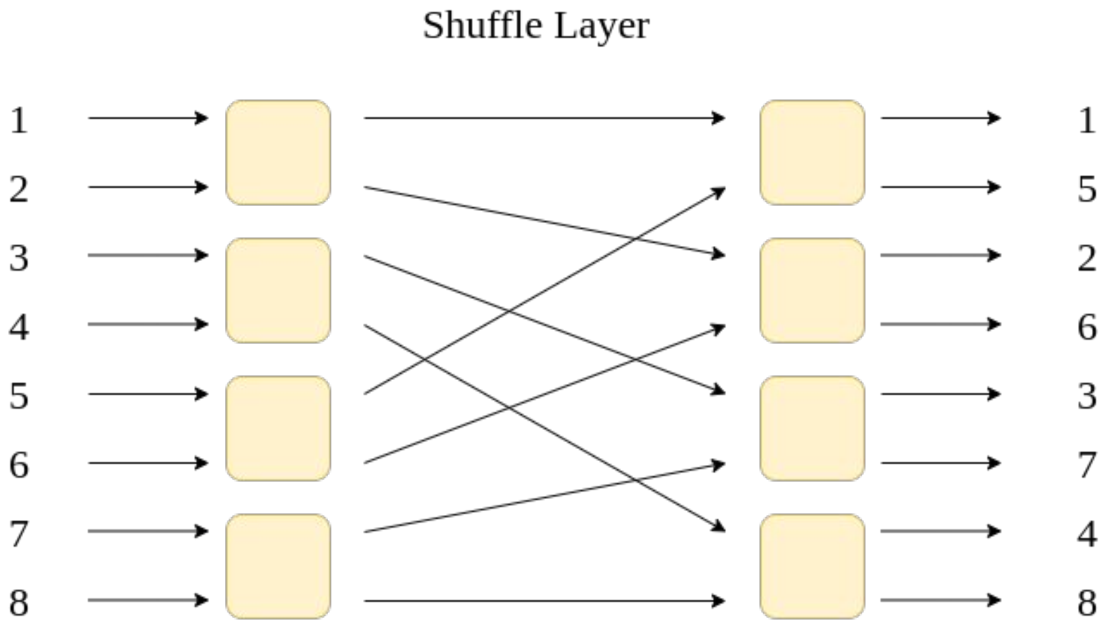
1
2
3
4

5
6
7
8

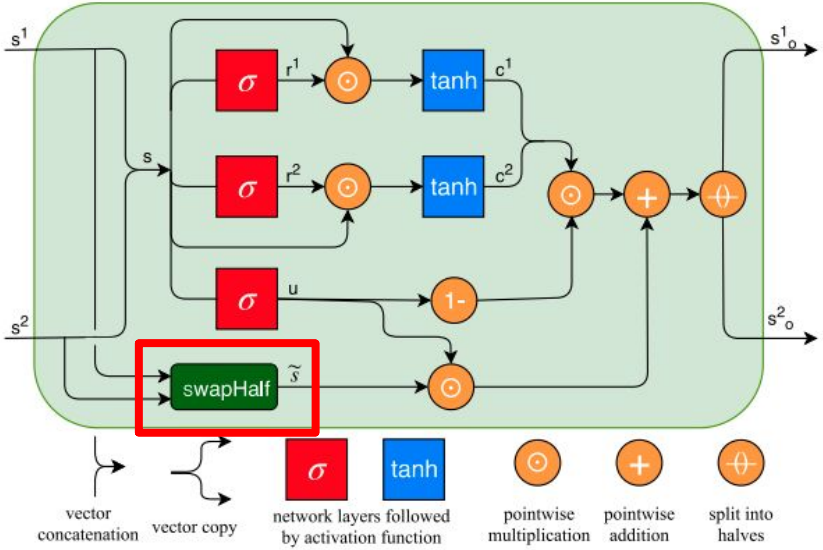
1
5
2
6
3
7
4
8

????

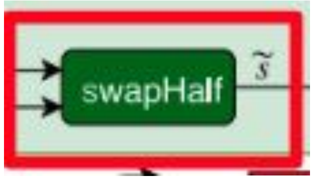
Shuffle



Neural Shuffle Exchange Network - Switch Unit



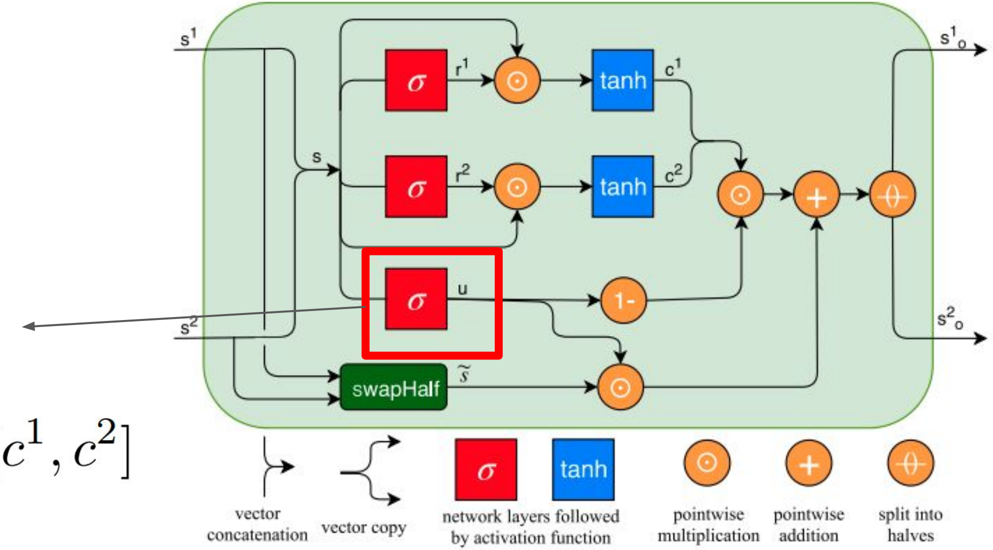
Exchange



$$\text{swapHalf}\left(\overset{\text{s1}}{\begin{bmatrix} a \\ b \end{bmatrix}}, \overset{\text{s2}}{\begin{bmatrix} c \\ d \end{bmatrix}}\right) = \left[\overset{\text{s1}}{\begin{bmatrix} a \\ d \end{bmatrix}}, \overset{\text{s2}}{\begin{bmatrix} c \\ b \end{bmatrix}}\right]$$

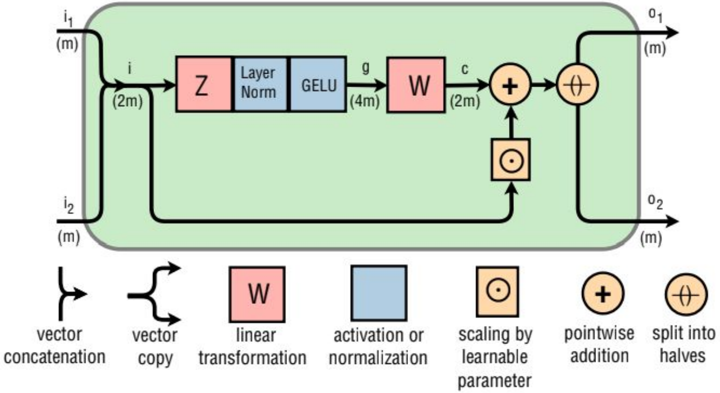
Neural Shuffle Exchange Network - Switch Unit

Determine
How Exchange



$$[s_o^1, s_o^2] = u \odot \tilde{s} + (1 - u) \odot [c^1, c^2]$$

Residual Shuffle Exchange Network



Shuffle-Exchange Network, Benes network

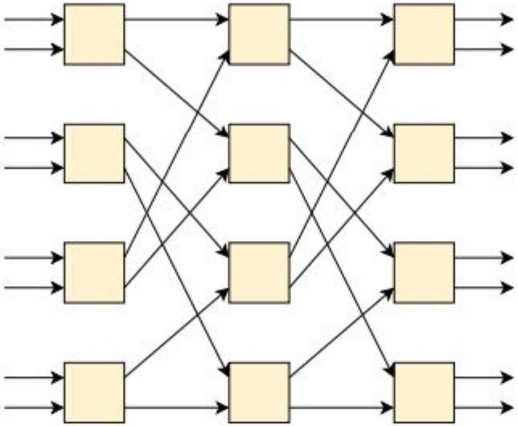


Figure 1: Shuffle-Exchange network.

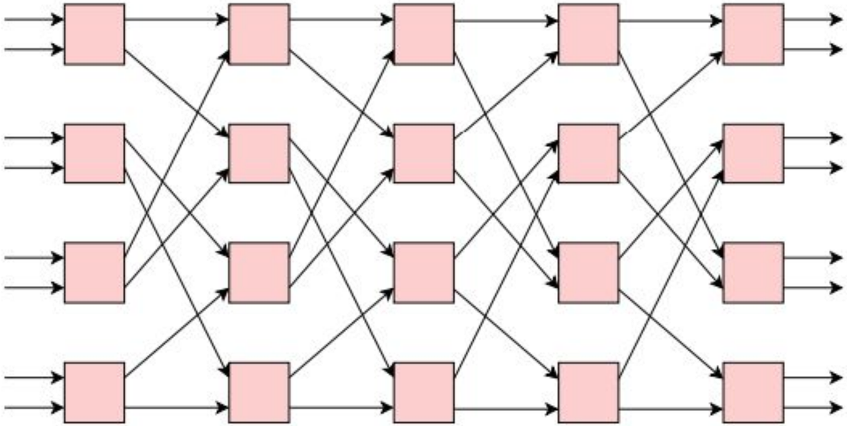
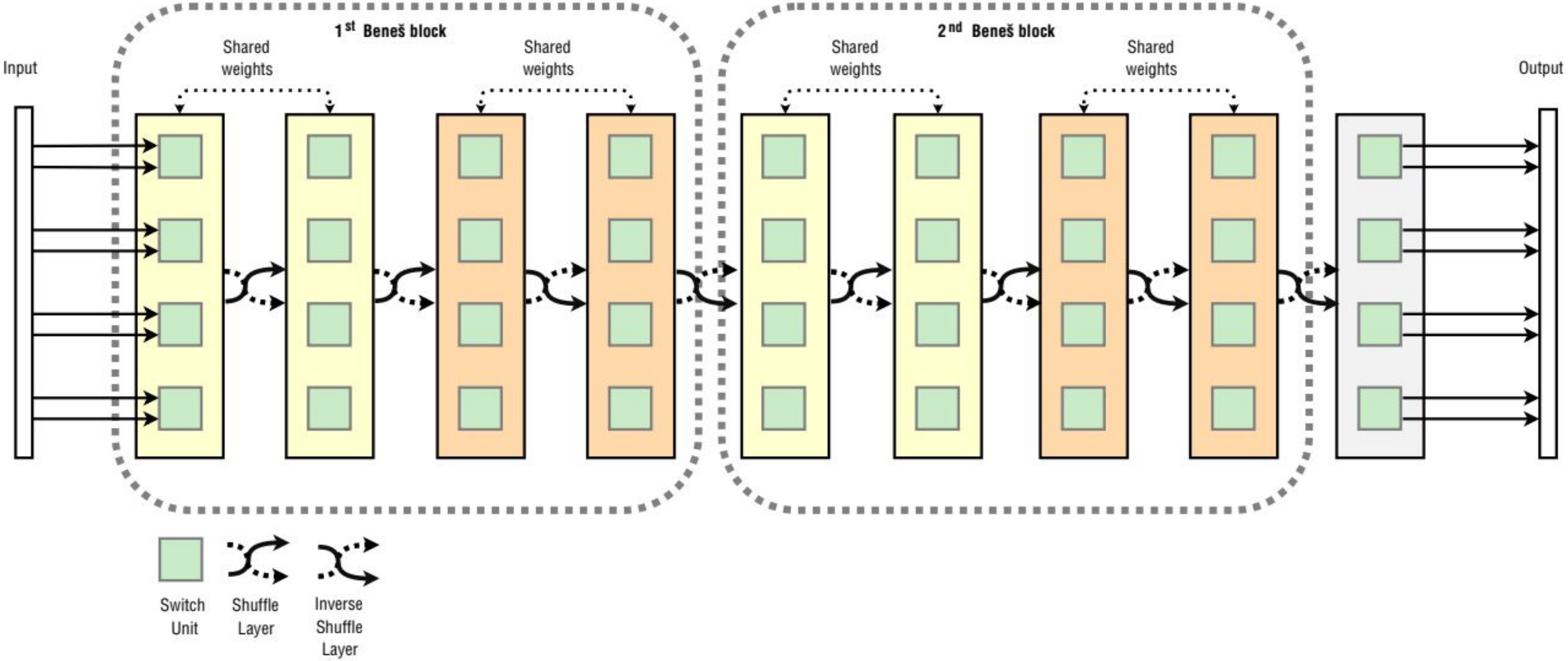


Figure 2: Beneš network.

Residual Shuffle Exchange Network



Experiment(s)

5 Experiment Environments (mention 1)

1. Lambda Question Answering Dataset ✓
2. MusicNet Dataset
3. Multiplication Task
4. Sort Task
5. Adding Task

Experiment(s)

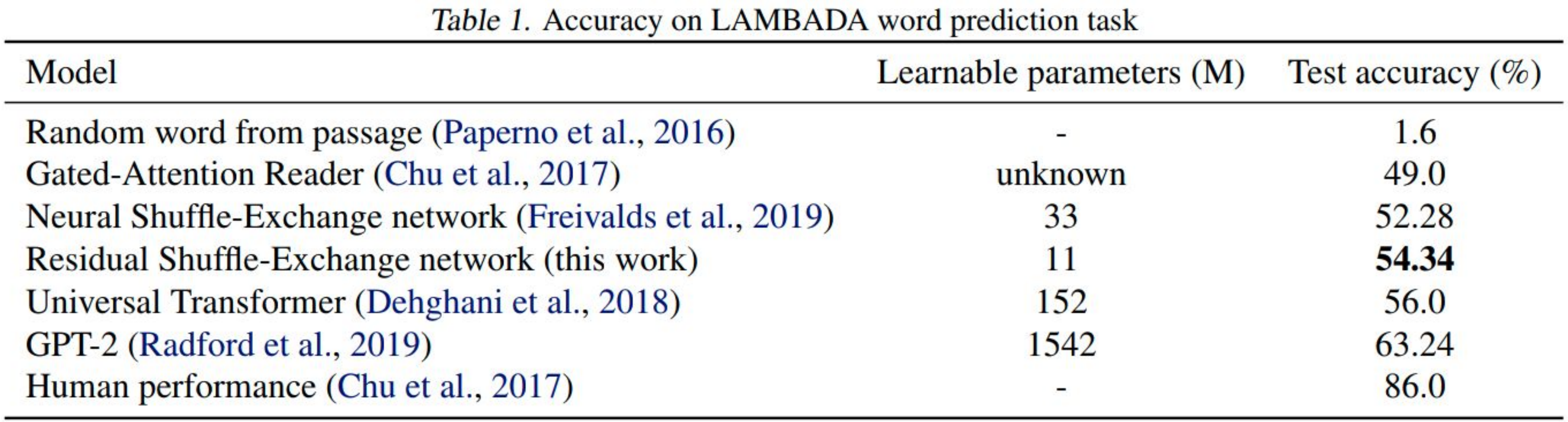
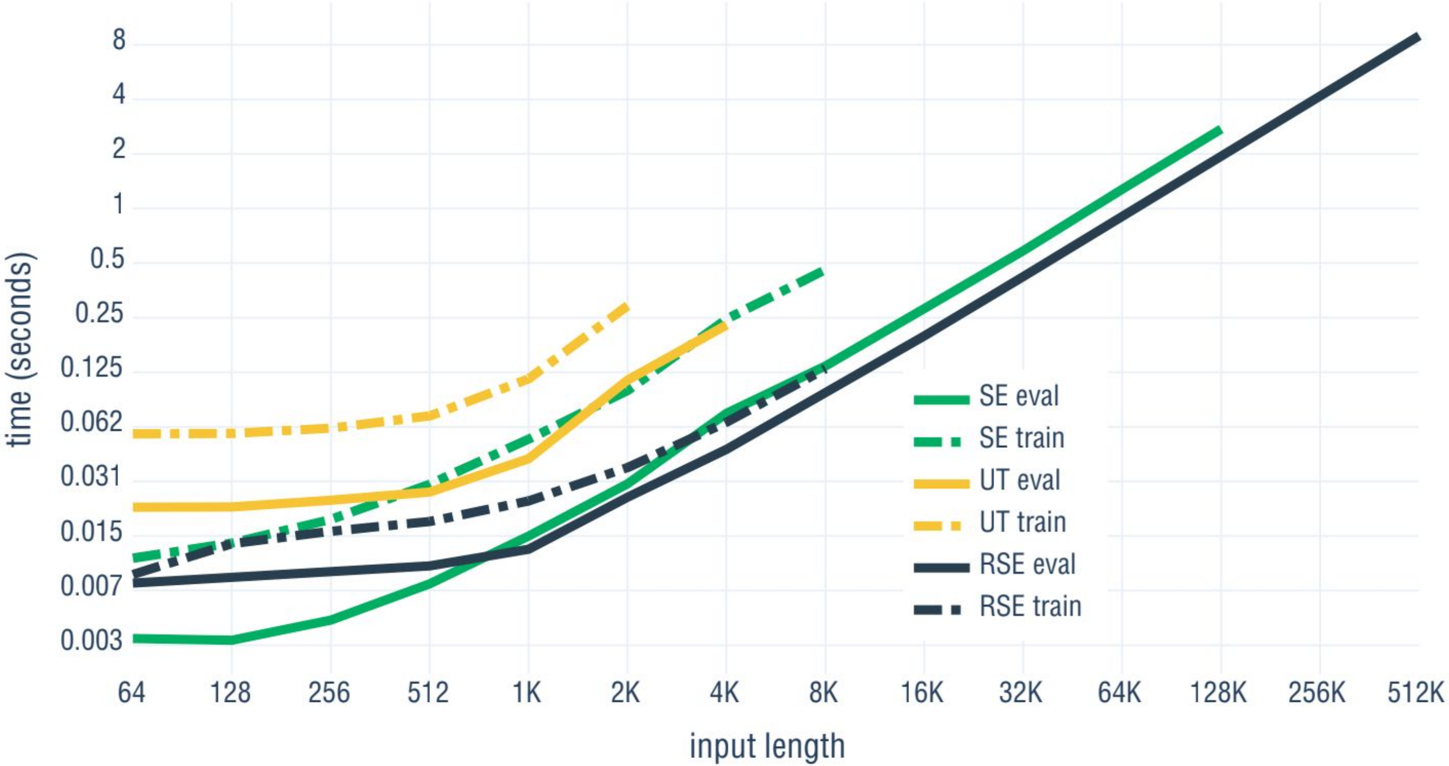


Table 1. Accuracy on LAMBADA word prediction task

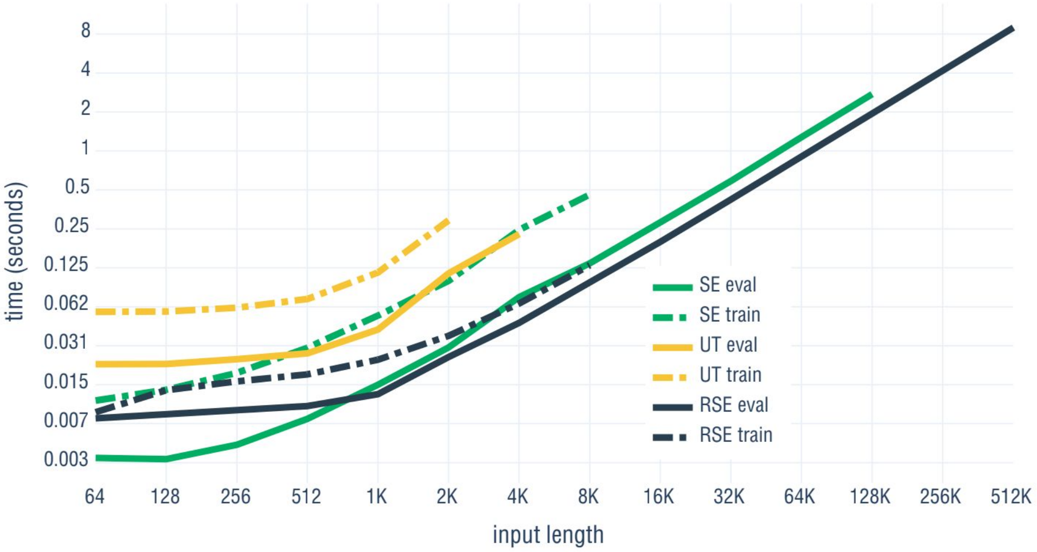
Model	Learnable parameters (M)	Test accuracy (%)
Random word from passage (Paperno et al., 2016)	-	1.6
Gated-Attention Reader (Chu et al., 2017)	unknown	49.0
Neural Shuffle-Exchange network (Freivalds et al., 2019)	33	52.28
Residual Shuffle-Exchange network (this work)	11	**54.34**
Universal Transformer (Dehghani et al., 2018)	152	56.0
GPT-2 (Radford et al., 2019)	1542	63.24
Human performance (Chu et al., 2017)	-	86.0

Experiment(s)



Memory Requirement

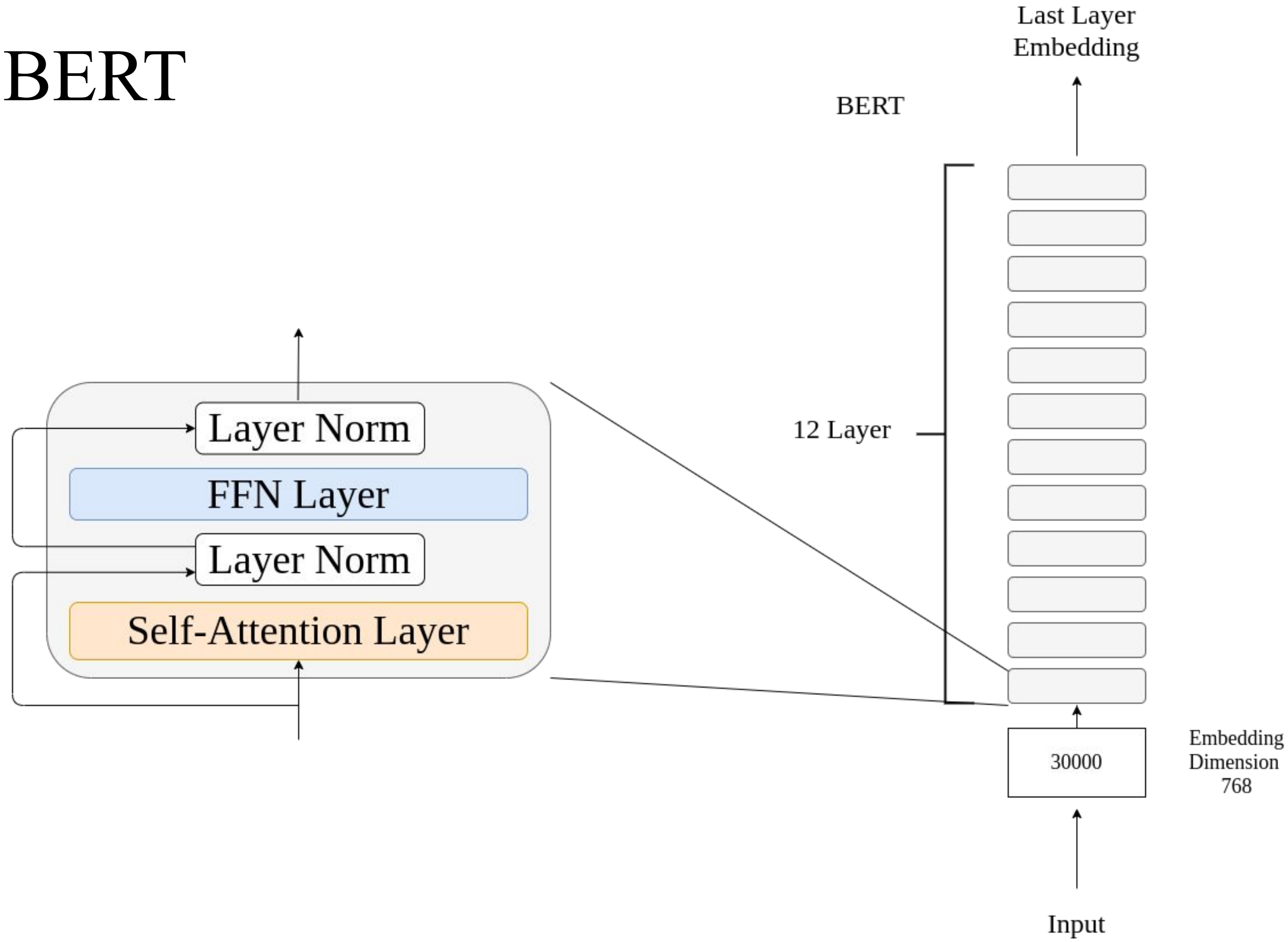
1. Evaluation: 11G GPU memory:
   4x longer sequence than Neural Shuffle-Exchange Network.
   128x longer sequence than Universal Transformer.



BERT

BERT

BERT


Last Layer Embedding
BERT
12 Layer
Layer Norm
FFN Layer
Layer Norm
Self-Attention Layer
30000
Embedding Dimension 768
Input

Bert: Pre-training Task

Mask Language Model:

Input: [CLS] 劉碩他要簽 [MASK] → [CLS] 劉碩他要簽博

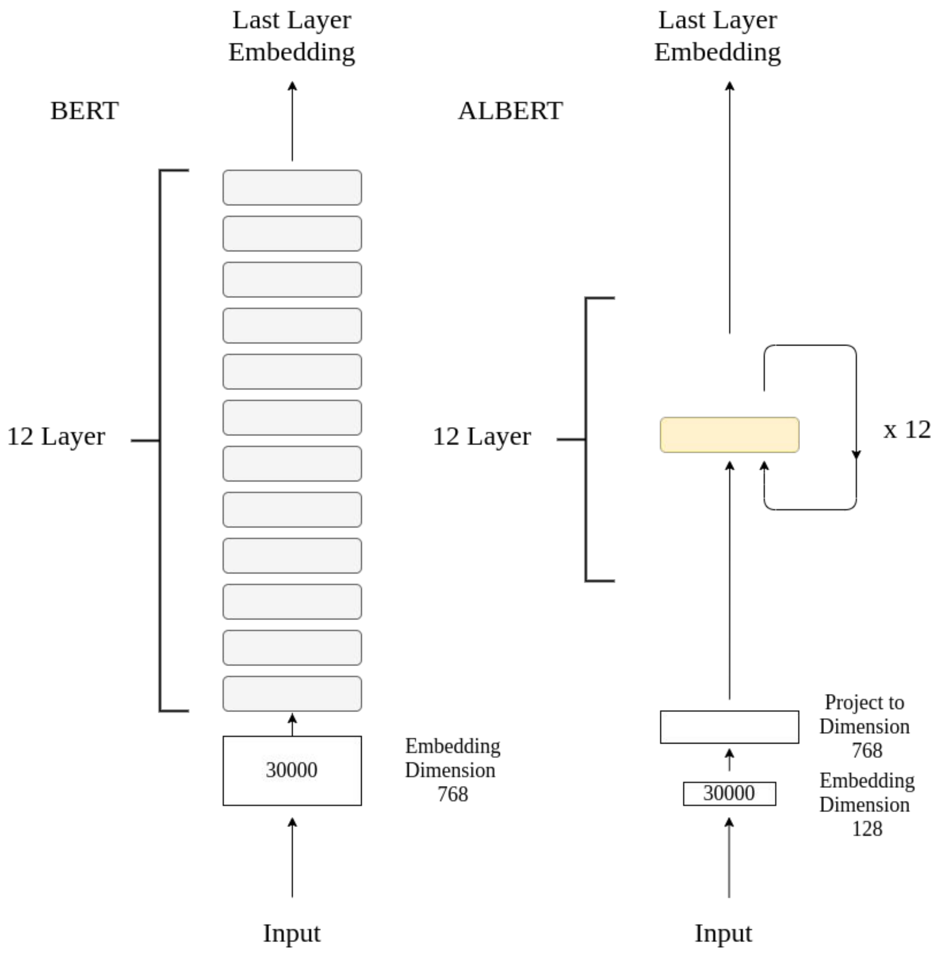
Next Sentence Prediction:

Input: ❌ [CLS] 我要當老師 [SEP] 人民的法槌

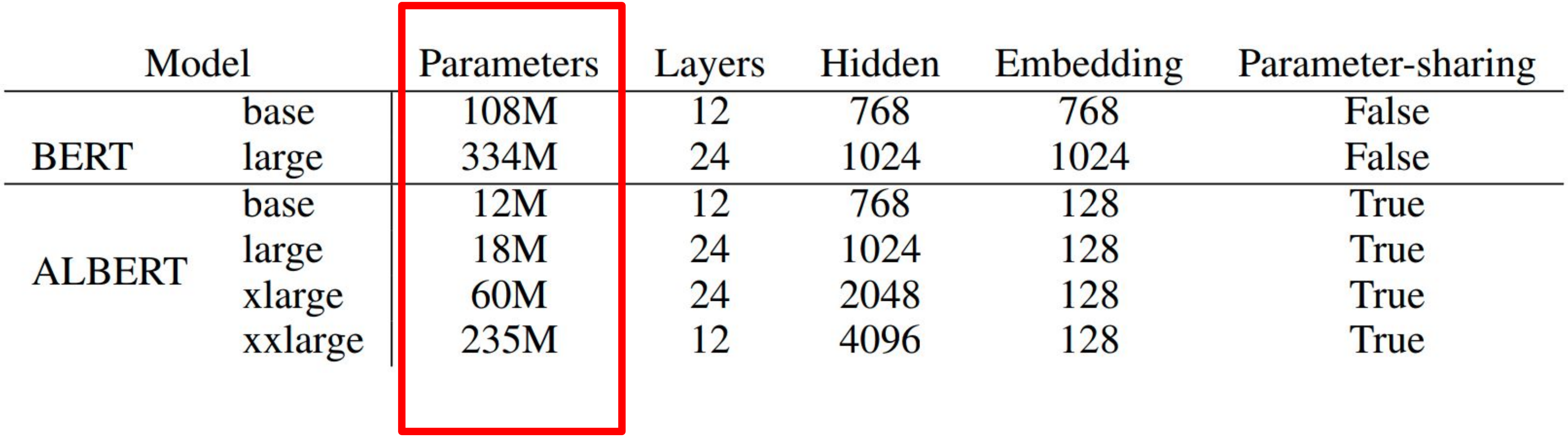
Input: ✔ [CLS] 這隻手是人民的意志 [SEP] 人民的法槌

ALBERT

ALBERT


Last Layer Embedding
BERT
12 Layer
30000
Embedding Dimension 768
Input
Last Layer Embedding
ALBERT
12 Layer
x 12
Project to Dimension 768
30000
Embedding Dimension 128
Input

ALBERT



Model		Parameters	Layers	Hidden	Embedding	Parameter-sharing
BERT	base	108M	12	768	768	False
	large	334M	24	1024	1024	False
ALBERT	base	12M	12	768	128	True
	large	18M	24	1024	128	True
	xlarge	60M	24	2048	128	True
	xxlarge	235M	12	4096	128	True

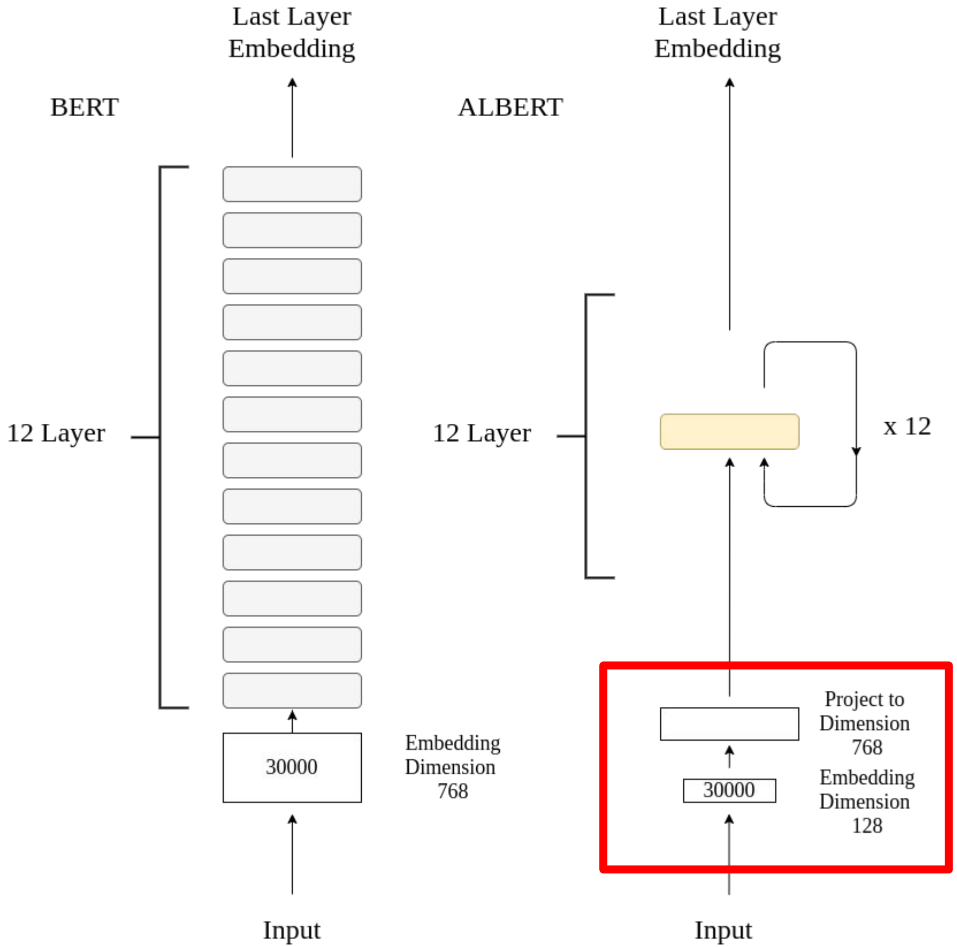
ALBERT

1.Factorize Embedding Matrix

Original BERT:
30000 x 768 = 23.04M

ALBERT:
30000 x128 = 3.8M
128 x 768 = 0.098M
Total: 3.898M

Reduce Parameters !

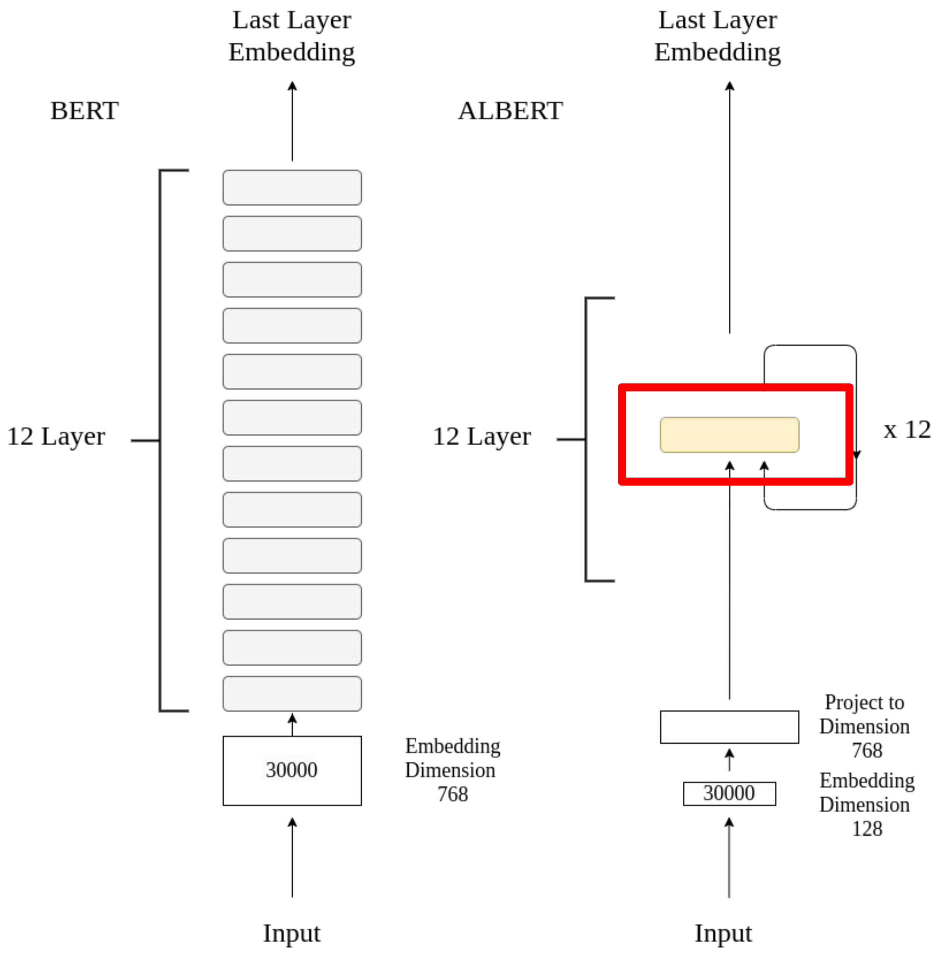


ALBERT

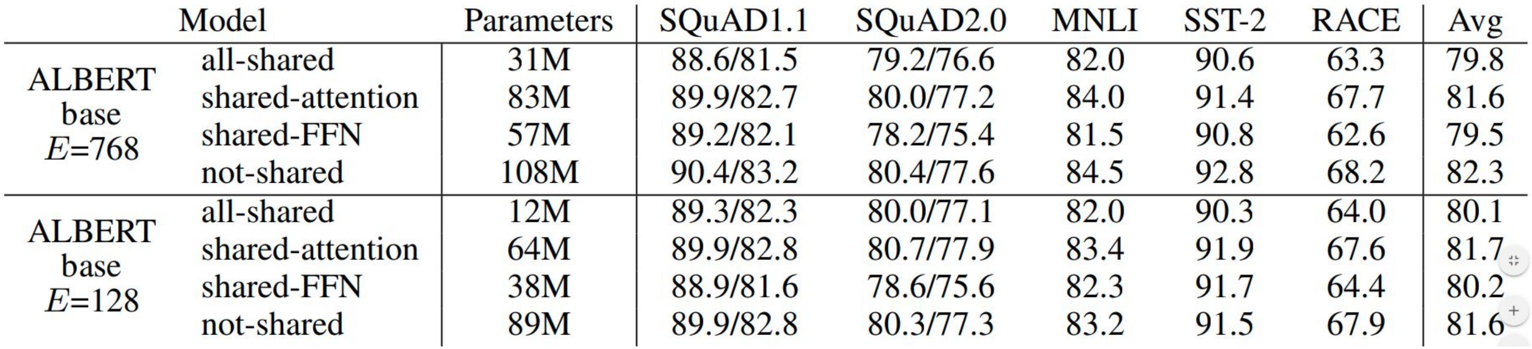
2.Shared Same Parameters across Layer

1/ 12 BERT Parameters

Reduce Parameters !!!



Share Parameters Experiment(s)



Model		Parameters	SQuAD1.1	SQuAD2.0	MNLI	SST-2	RACE	Avg
ALBERT base E=768	all-shared	31M	88.6/81.5	79.2/76.6	82.0	90.6	63.3	79.8
	shared-attention	83M	89.9/82.7	80.0/77.2	84.0	91.4	67.7	81.6
	shared-FFN	57M	89.2/82.1	78.2/75.4	81.5	90.8	62.6	79.5
	not-shared	108M	90.4/83.2	80.4/77.6	84.5	92.8	68.2	82.3
ALBERT base E=128	all-shared	12M	89.3/82.3	80.0/77.1	82.0	90.3	64.0	80.1
	shared-attention	64M	89.9/82.8	80.7/77.9	83.4	91.9	67.6	81.7
	shared-FFN	38M	88.9/81.6	78.6/75.6	82.3	91.7	64.4	80.2
	not-shared	89M	89.9/82.8	80.3/77.3	83.2	91.5	67.9	81.6

ALBERT: Pre-training Task

Mask Language Model:

Input: [CLS] 劉碩他要簽 [MASK] ⟶ [CLS] 劉碩他要簽博

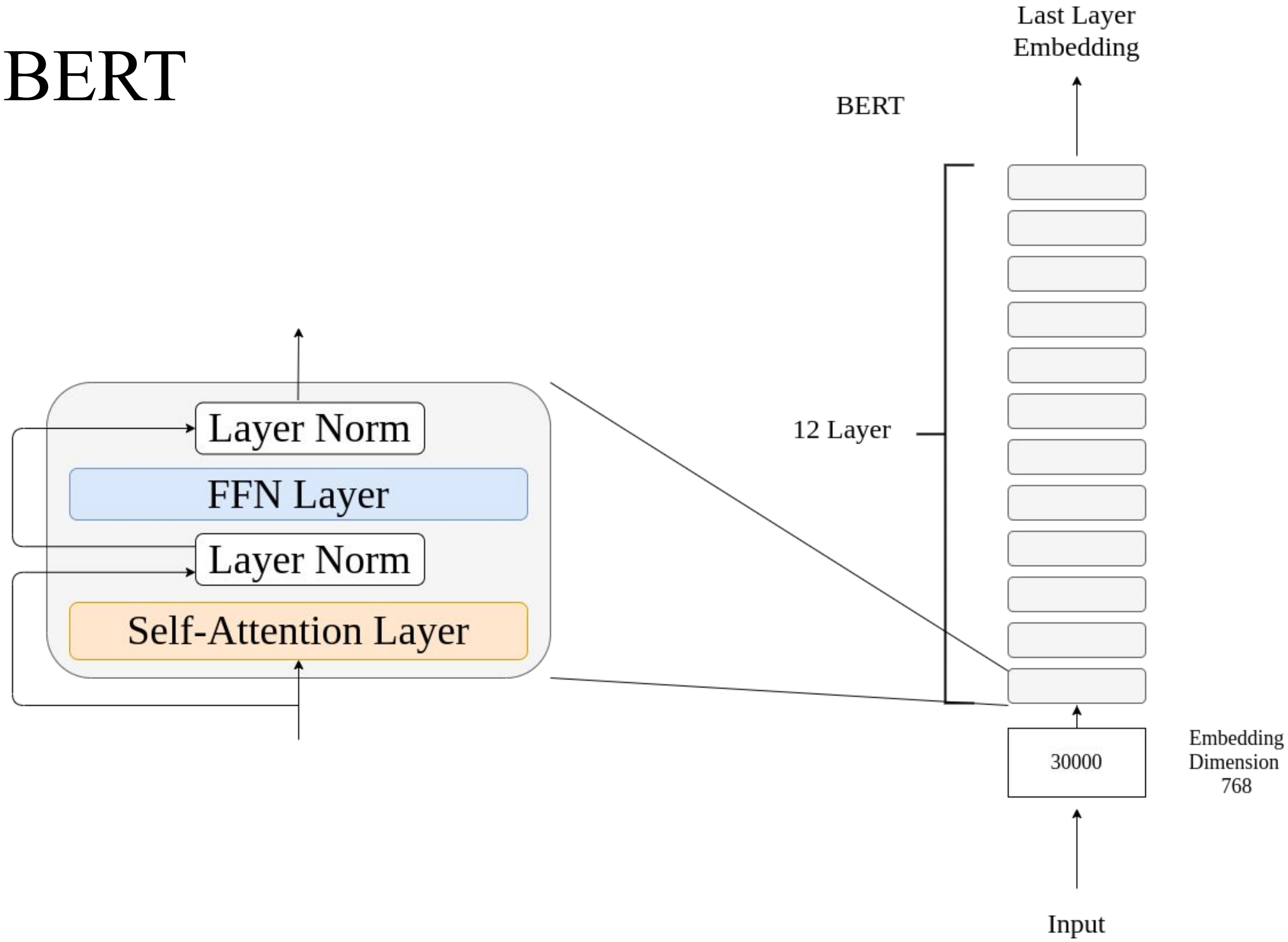
Sentence Order Prediction:

Input: ✗ [CLS] 人民的法槌 [SEP] 這隻手是人民的意志

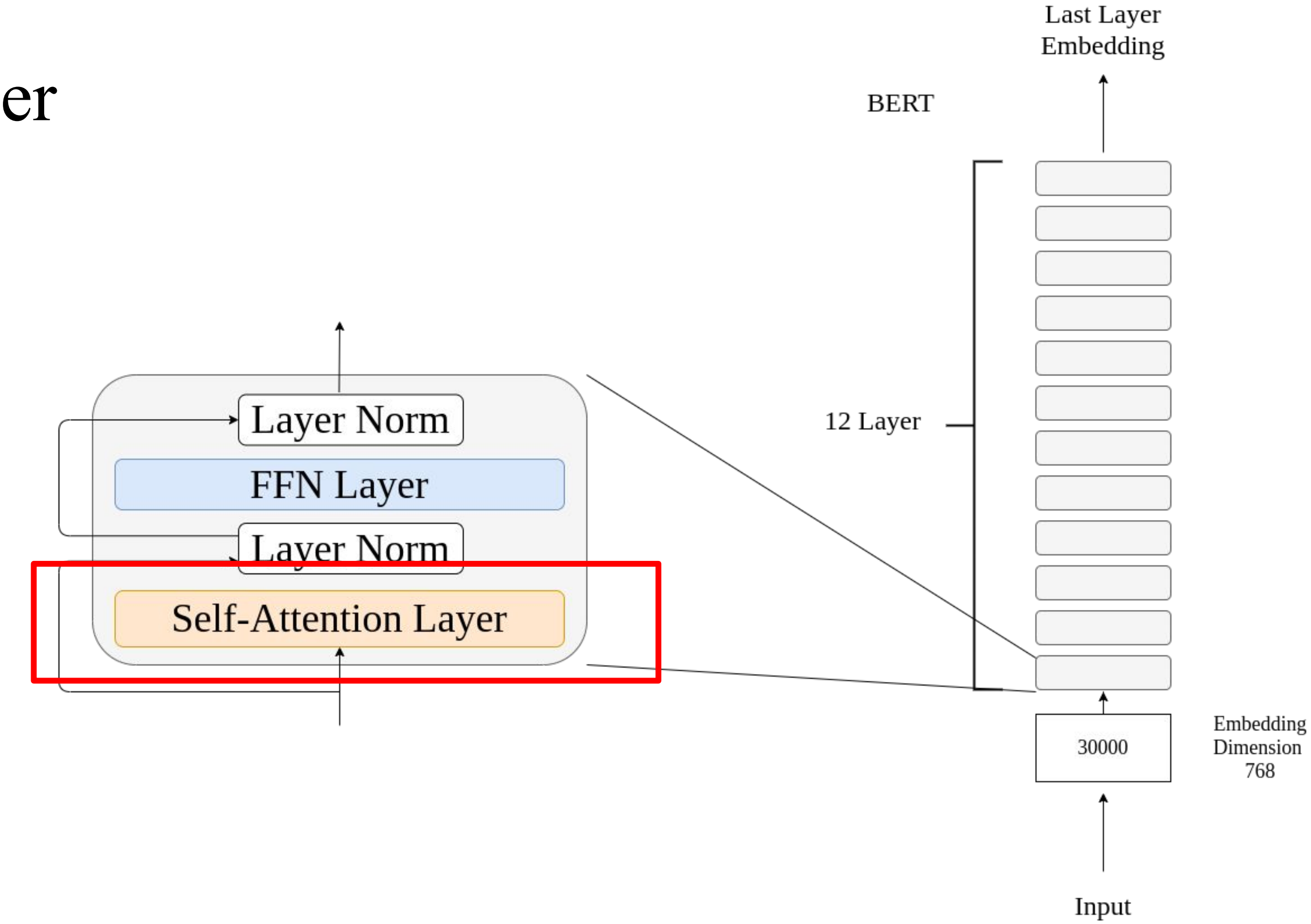
Input: ✓ [CLS] 這隻手是人民的意志 [SEP] 人民的法槌

Reformer

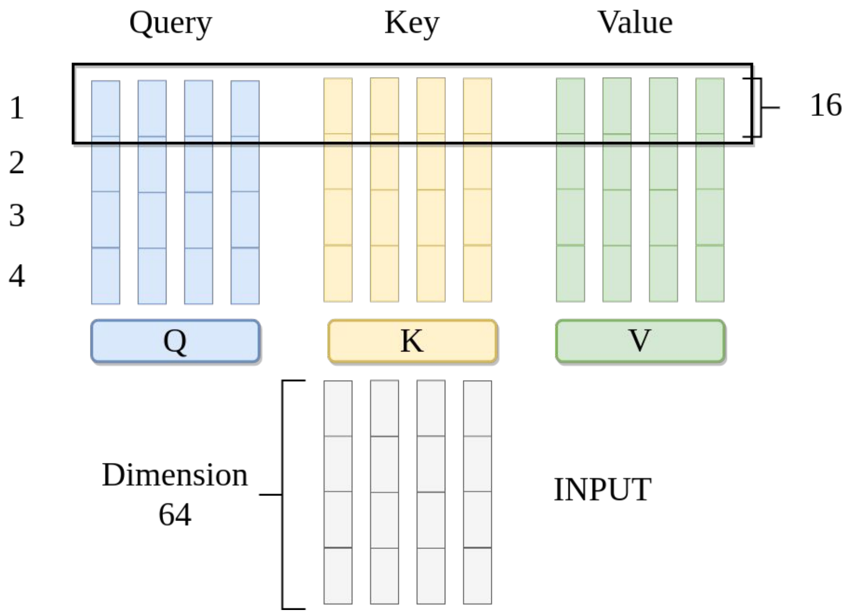
BERT


Last Layer Embedding
BERT
12 Layer
Layer Norm
FFN Layer
Layer Norm
Self-Attention Layer
30000
Embedding Dimension 768
Input

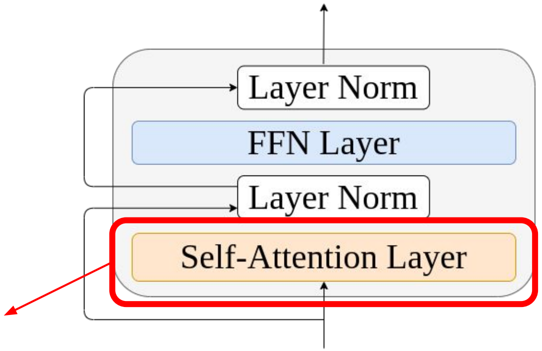
Attention Layer



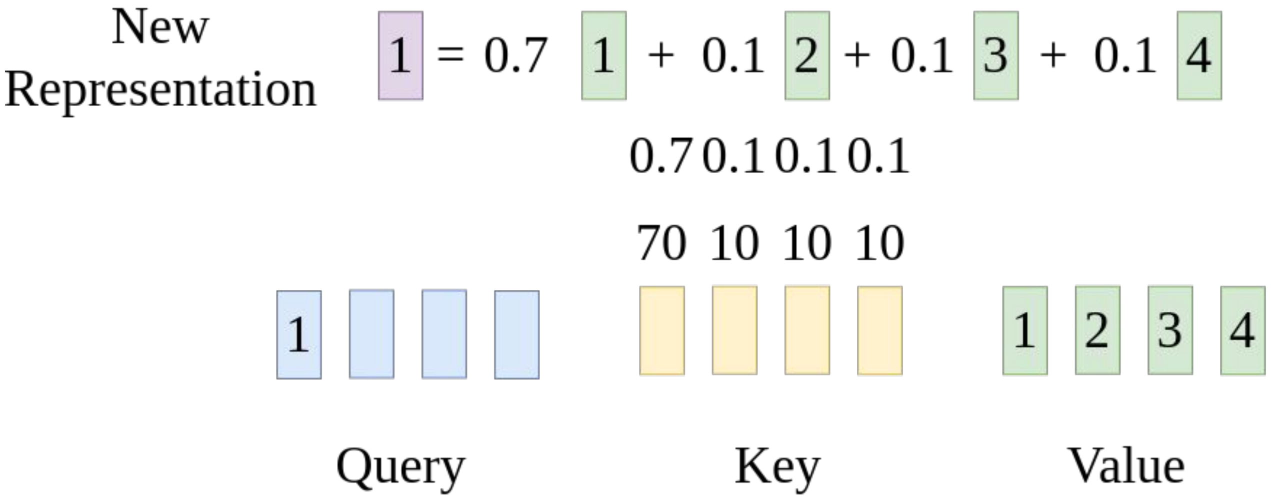
Attention Layer - Prepare Q, K, V



4 Head Attention



Attention Layer - Attention Mechanism



New Representation

1 = 0.7 1 + 0.1 2 + 0.1 3 + 0.1 4

0.7 0.1 0.1 0.1

70 10 10 10

1

1 2 3 4

Query

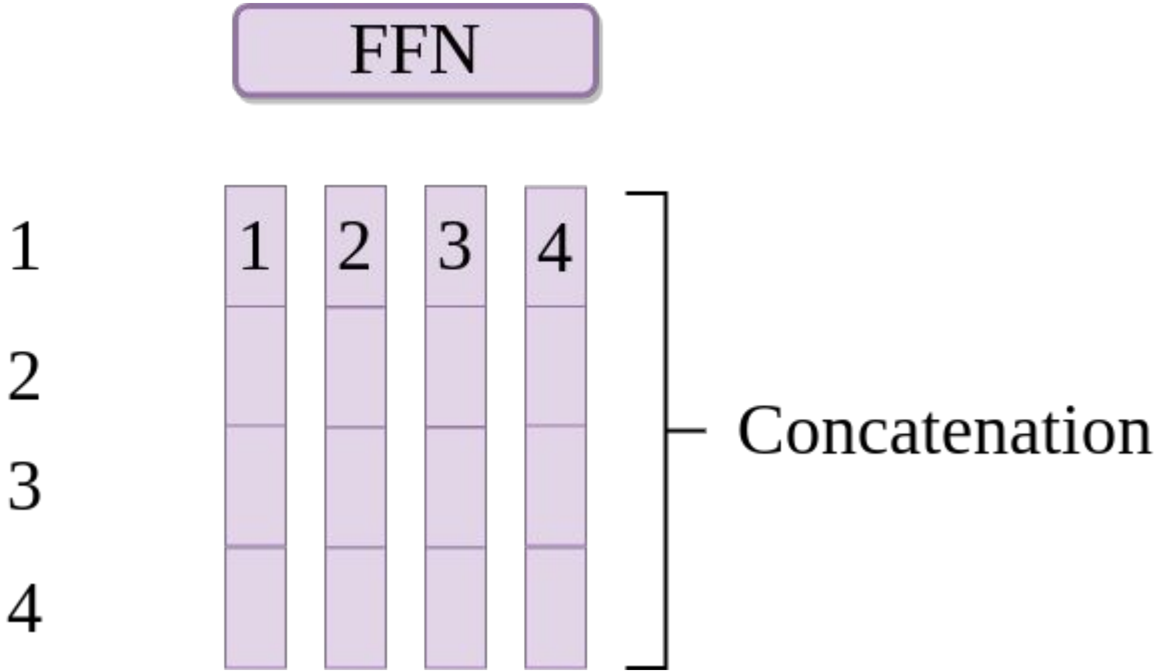
Key

Value

N = 4

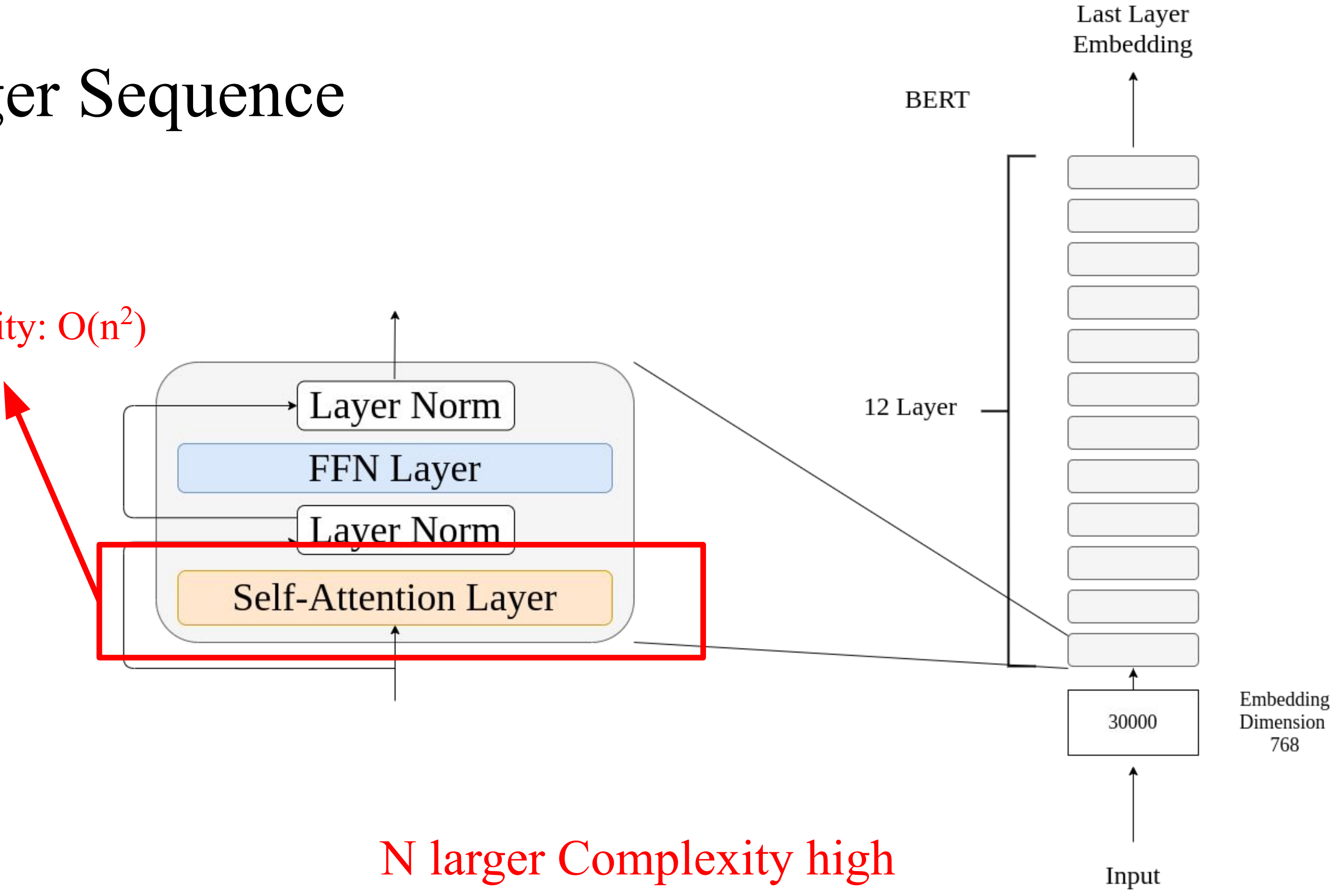
Complexity: $O(n^2)$

Concatenation - pass a FFN



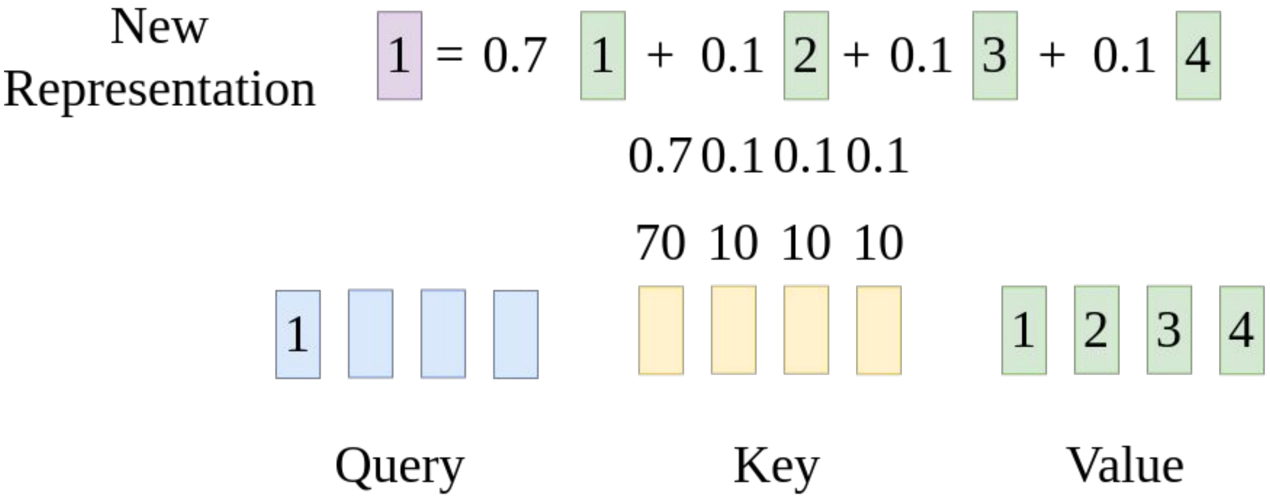
Issue: Longer Sequence

Quadratic Perplexity: $O(n^2)$



N larger Complexity high

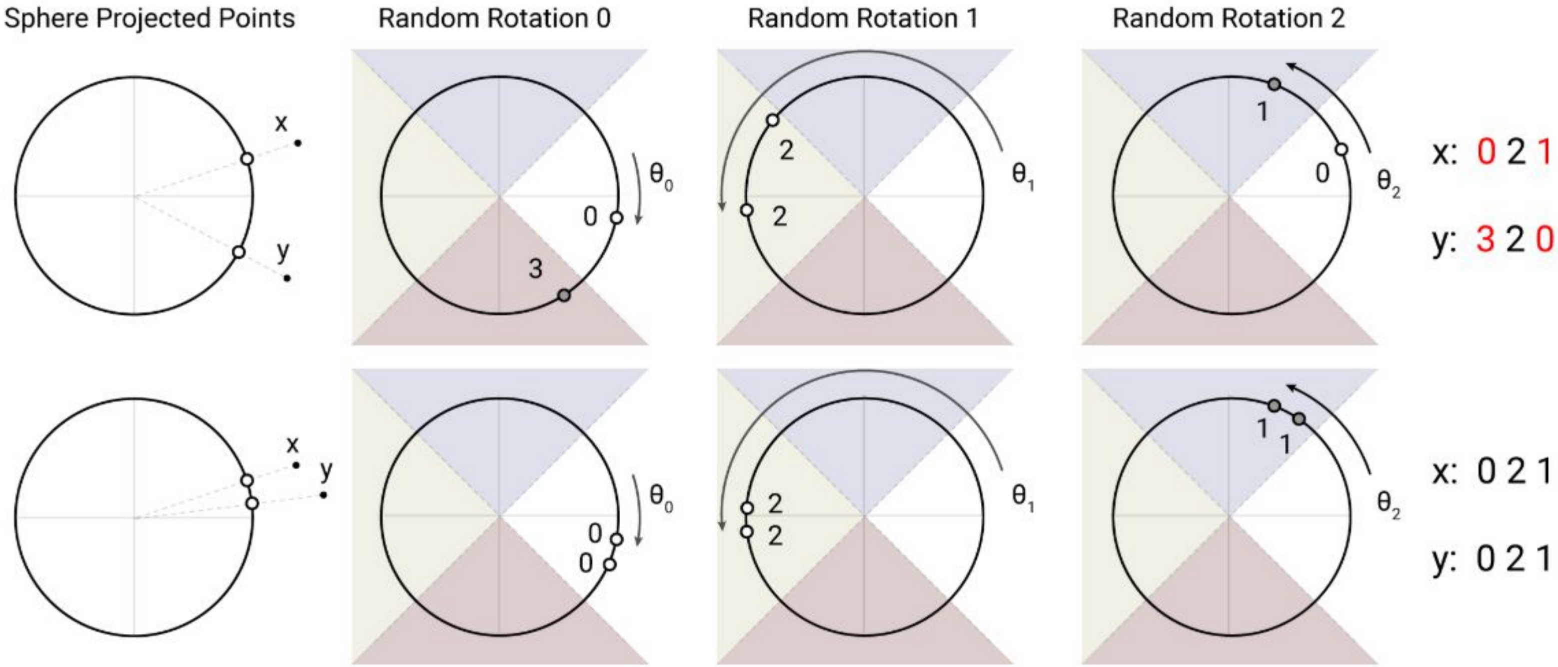
Attention Layer - Attention Mechanism



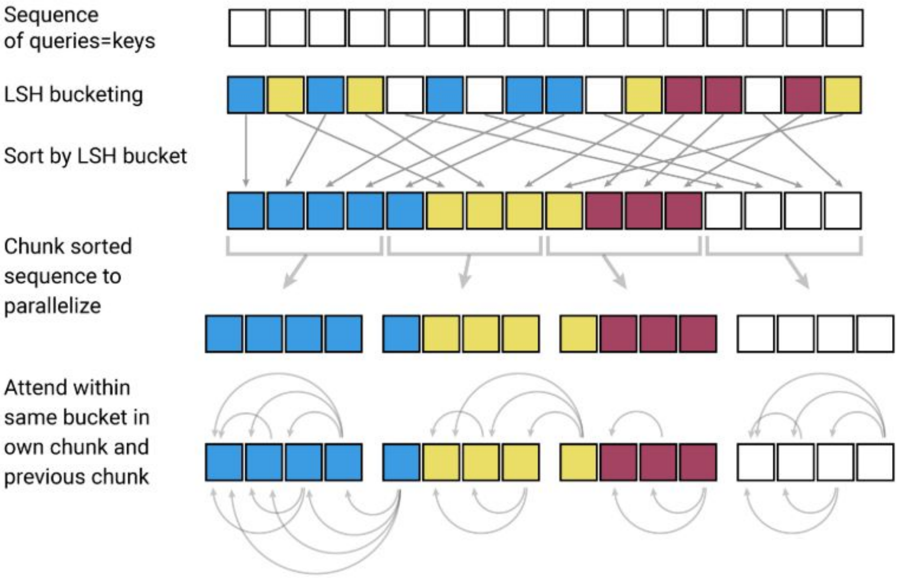
N = 8k ?

Complexity: $O(n^2)$

Hash Function



Reformer: Find a small set of candidates by hash function



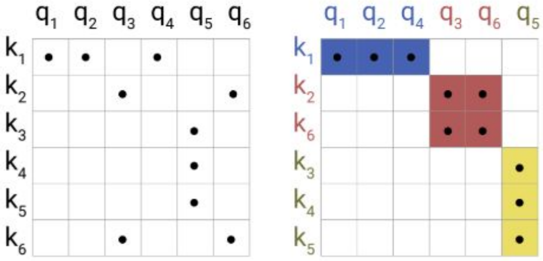
Sequence of queries=keys

LSH bucketing

Sort by LSH bucket

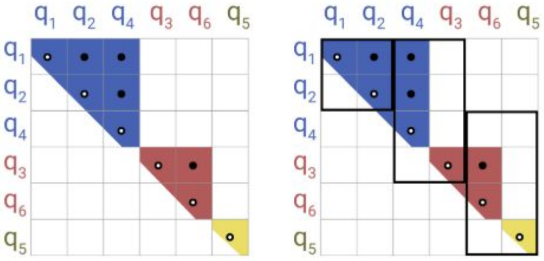
Chunk sorted sequence to parallelize

Attend within same bucket in own chunk and previous chunk



q_1 q_2 q_3 q_4 q_5 q_6
k_1 k_2 k_3 k_4 k_5 k_6
q_1 q_2 q_4 q_3 q_6 q_5
k_1 k_2 k_6 k_3 k_4 k_5

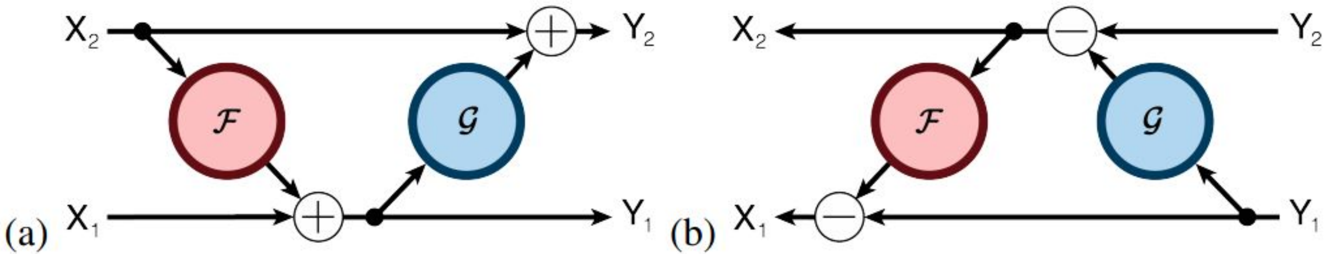
(a) Normal (b) Bucketed



q_1 q_2 q_4 q_3 q_6 q_5
q_1 q_2 q_4 q_3 q_6 q_5
q_1 q_2 q_4 q_3 q_6 q_5
q_1 q_2 q_4 q_3 q_6 q_5

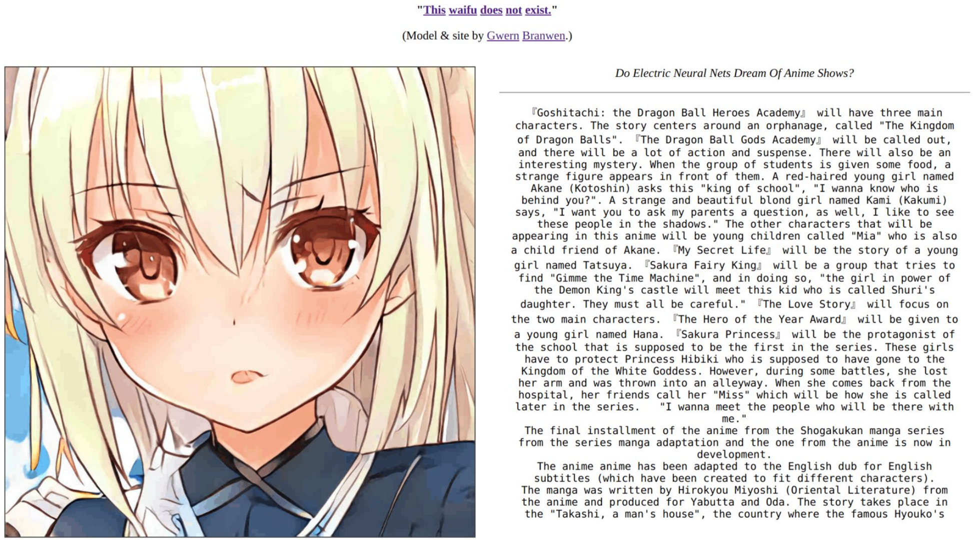
(c) Q = K (d) Chunked

Reversible Layer



Style GAN

This Waifu Does Not Exist



"This waifu does not exist."

(Model & site by Gwern Branwen.)

Do Electric Neural Nets Dream Of Anime Shows?

『Goshitachi: the Dragon Ball Heroes Academy』 will have three main characters. The story centers around an orphanage, called "The Kingdom of Dragon Balls". 『The Dragon Ball Gods Academy』 will be called out, and there will be a lot of action and suspense. There will also be an interesting mystery. When the group of students is given some food, a strange figure appears in front of them. A red-haired young girl named Akane (Kotoshin) asks this "king of school", "I wanna know who is behind you?". A strange and beautiful blond girl named Kami (Kakumi) says, "I want you to ask my parents a question, as well, I like to see these people in the shadows." The other characters that will be appearing in this anime will be young children called "Mia" who is also a child friend of Akane. 『My Secret Life』 will be the story of a young girl named Tatsuya. 『Sakura Fairy King』 will be a group that tries to find "Gimme the Time Machine", and in doing so, "the girl in power of the Demon King's castle will meet this kid who is called Shuri's daughter. They must all be careful." 『The Love Story』 will focus on the two main characters. 『The Hero of the Year Award』 will be given to a young girl named Hana. 『Sakura Princess』 will be the protagonist of the school that is supposed to be the first in the series. These girls have to protect Princess Hibiki who is supposed to have gone to the Kingdom of the White Goddess. However, during some battles, she lost her arm and was thrown into an alleyway. When she comes back from the hospital, her friends call her "Miss" which will be how she is called later in the series. "I wanna meet the people who will be there with me."

The final installment of the anime from the Shogakukan manga series from the series manga adaptation and the one from the anime is now in development.

The anime anime has been adapted to the English dub for English subtitles (which have been created to fit different characters). The manga was written by Hirokyou Miyoshi (Oriental Literature) from the anime and produced for Yabutta and Oda. The story takes place in the "Takashi, a man's house", the country where the famous Hyouko's