On Identifiability in Transformers

Yejoon Lee

Apr 25, 2022

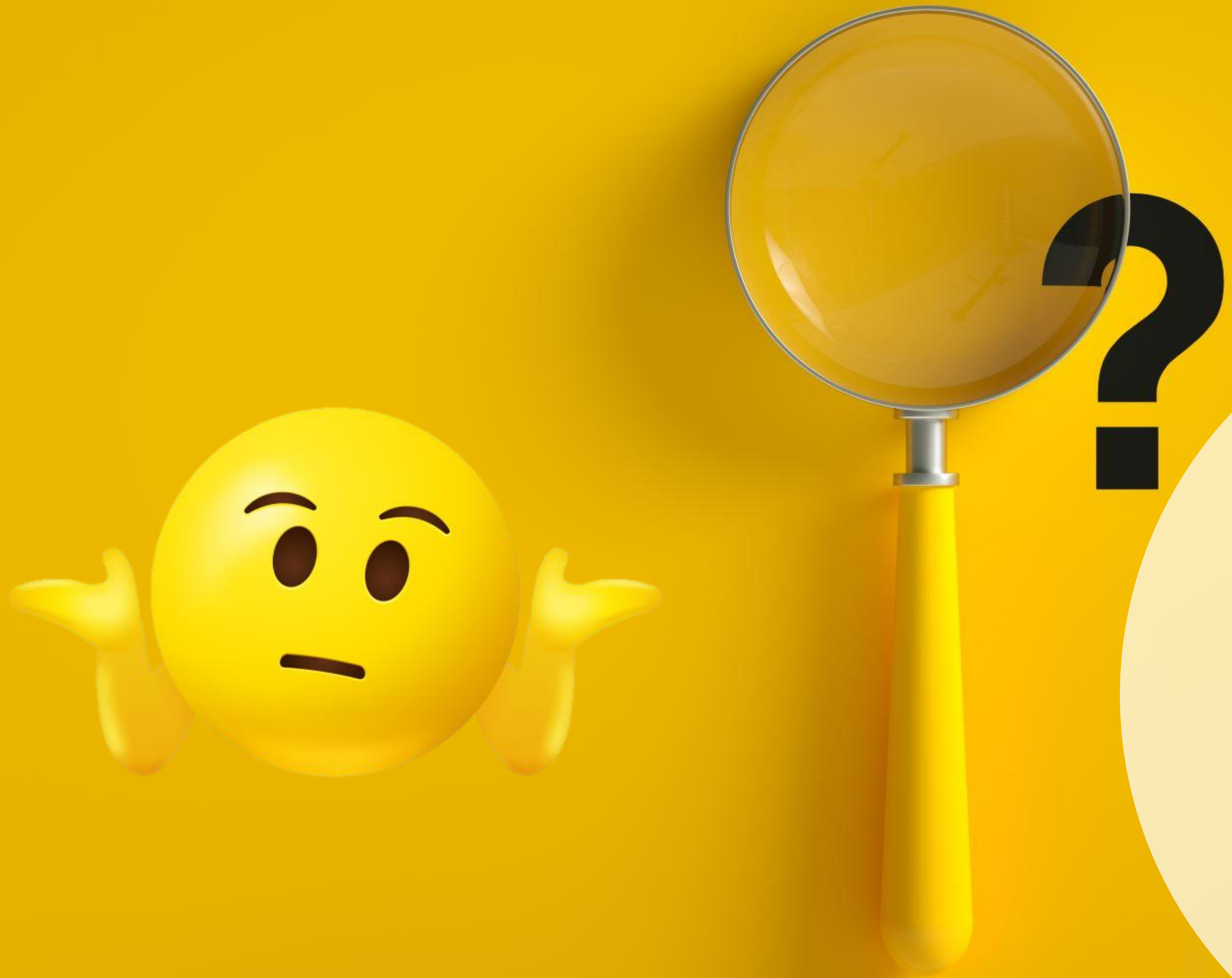
Featured Papers

- Brunner et al., 2020, On Identifiability In Transformers, ICLR

ON IDENTIFIABILITY IN TRANSFORMERS

Gino Brunner[1*], Yang Liu[2*], Damián Pascual[1*], Oliver Richter[1], Massimiliano Ciaramita[3], Roger Wattenhofer[1]
Departments of [1]Electrical Engineering and Information Technology, [2]Computer Science
ETH Zurich, Switzerland
[3]Google Research, Zurich, Switzerland

- Preliminaries:
  - Transformer
  - BERT



Do I Need to Know This?

Possible questions (in case you're not familiar with transformer)

Q. Isn't all of this only **limited to one specific model**, *transformer,* compared to post-hoc methods (LIME, SHAP, gradient-based, etc)?

A. Yes…

Q. Is *transformer* **that** important?

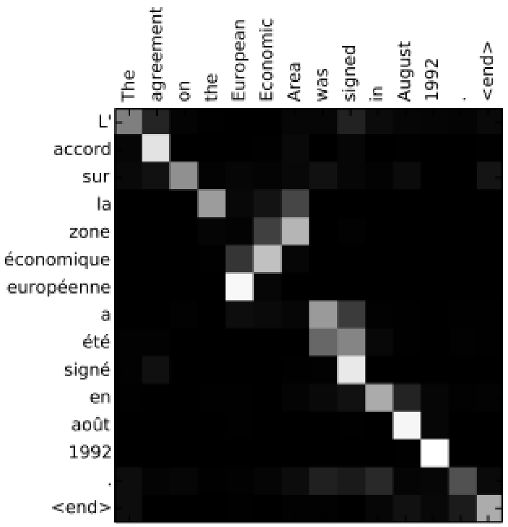
YES

GTC 2022 Keynote with NVIDIA CEO Jensen Huang

Okay. But..

Q. Why don't we just use post-hoc methods?

A. Good. However, ML practitioners are **tempted to use attention as interpretation**. Plus, post-hoc methods have downsides.



(a)

Bahdanau et al., 2014, Neural Machine Translation by Jointly Learning to Align and Translate, ICLR
Xu et al., 2015, Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, ICML

Table of Contents

- **Transformer & BERT**
- ***On Identifiability in Transformers***
- Follow-up works

Transformer & BERT

Transformer-based Language Models

- Transformer (*Vaswani et al., 2017*) is originally composed of **encoder** and **decoder.**
- However, so-called "transformer-based language models" often have different structures.
  - BERT (*Devlin et al., 2019*) : **encoder-only**
  - GPT-2 (*Radford et al, 2019*): **decoder-only**
  - T5 (*Raffel et al., 2020*): **encoder and decoder**

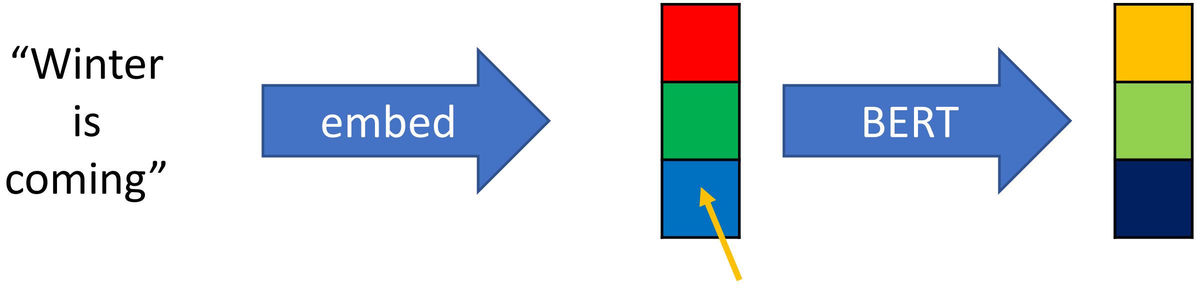
- **Self-attention** makes *transformer* a *transformer*. *On Identifiability in Transformers* uses **BERT** for experiments.

⟹ Let's look at the **self-attention** **in** **BERT** **Disclaimer: This presentation does not fully address neither Transformer nor BERT.*

Vaswani et al., 2017, Attention is All You Need, NeurIPS
Devlin et al., 2019, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, NAACL
Radford et al., 2019, Language Models are Unsupervised Multitask Learners
Raffel et al., 2020, Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

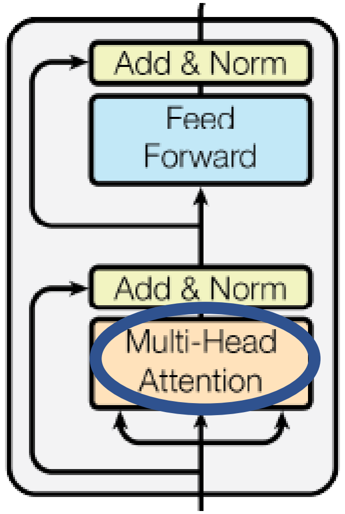
BERT produces a representation of words

- BERT is essentially a stacked *BERT layers.*
- Each layer **receives embedding as an input** and **produces new embedding as an output.** (embedding = vector representation of the word)
- BERT is short for Bidirectional Encoder **Representations** from Transformers



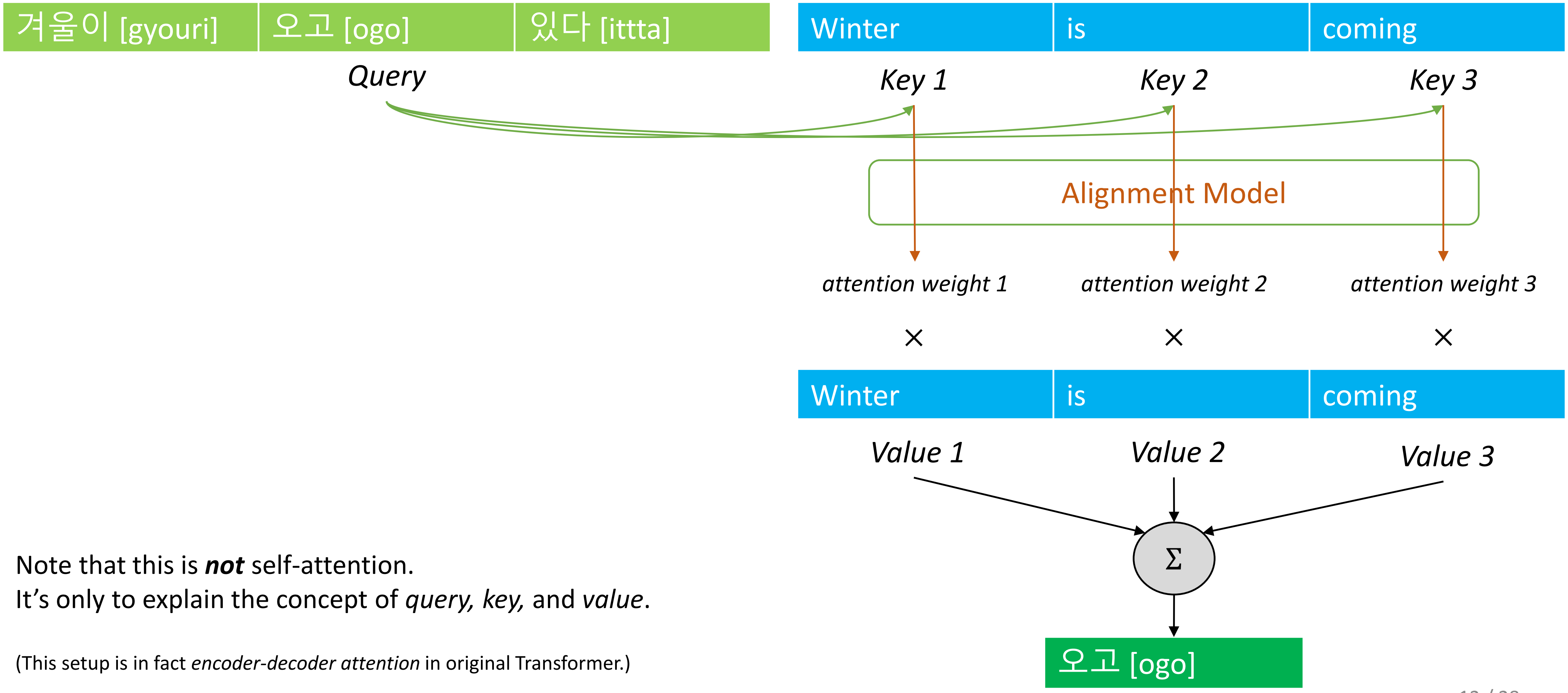
Each block represents ***d*-dimensional vector**;
d known as ***embedding dimension***

Layer of BERT (= encoder of Transformer)



Essence is ***Multi-Head Attention.***

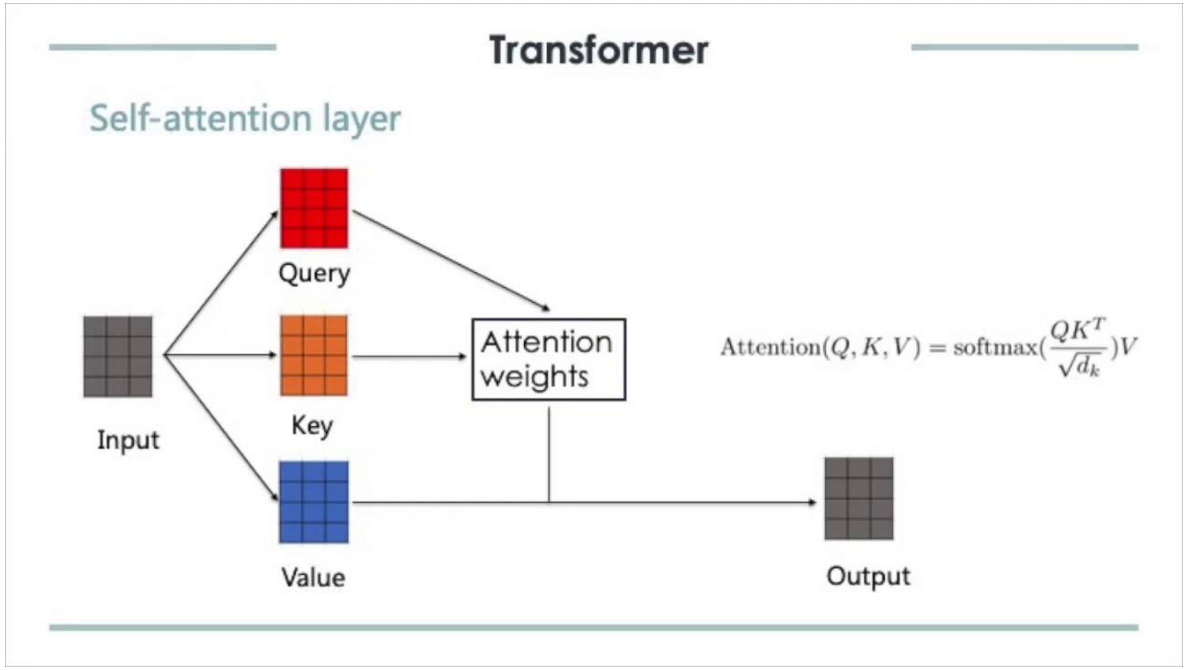
Query, Key, and Value



Note that this is ***not*** self-attention.
It's only to explain the concept of *query, key,* and *value*.

(This setup is in fact *encoder-decoder attention* in original Transformer.)

Single-Head Self-Attention



https://www.youtube.com/watch?v=5T38-2J5CcY

Why ***self***-attention?
: query and key/value comes from the same embedding

$$Q = EW^Q$$
$$K = EW^K$$
$$V = EW^V$$

$$A = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)$$
$$E_{new} = AV$$

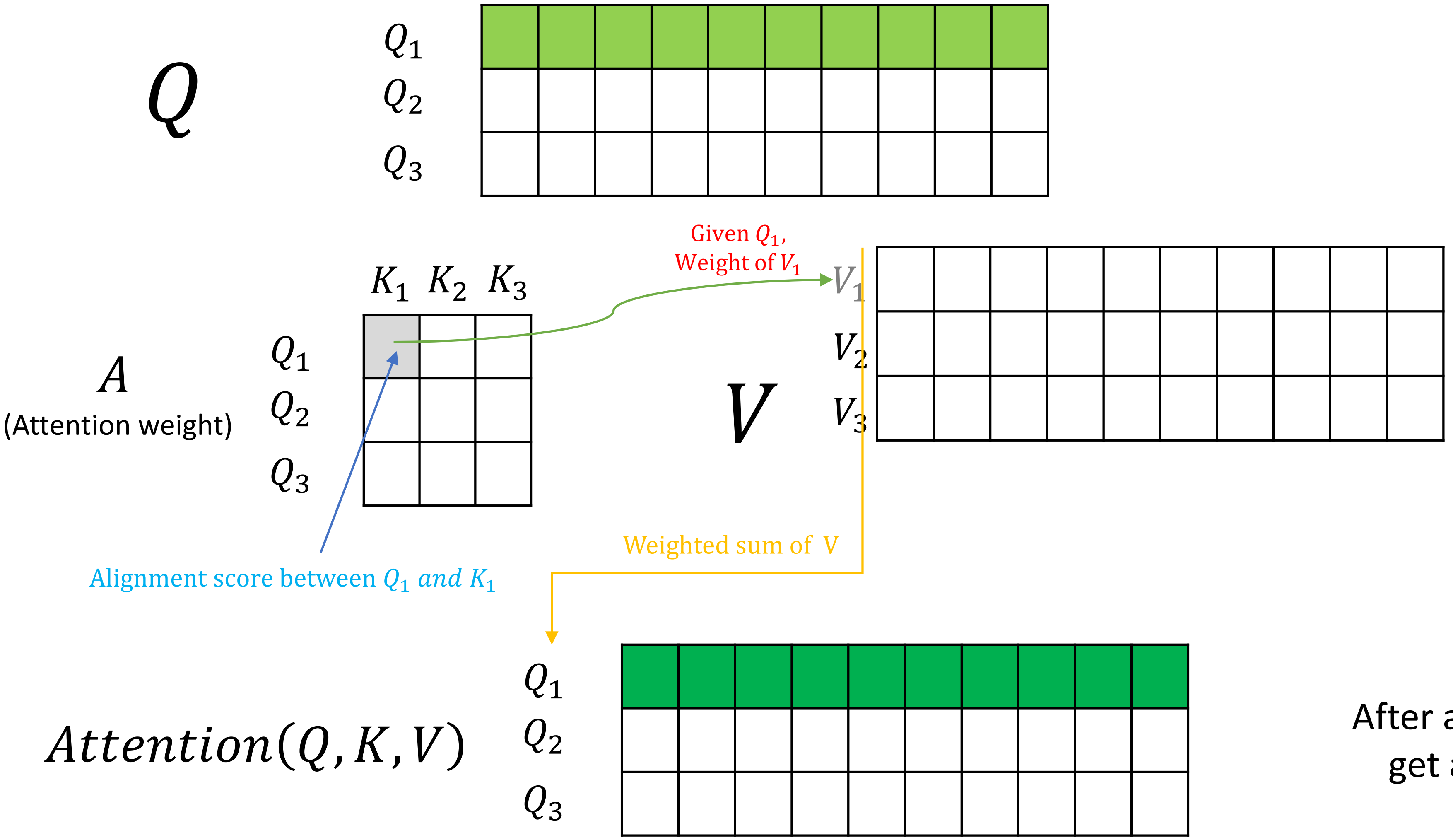
This formula itself is the alignment model *(scaled dot-product attention)*

$E: embedding;\ (d_s, d)\ where\ d_s\ denotes\ the\ length\ of\ input\ and\ d\ denotes\ the\ embedding\ dimension$

$Q: Query;\ (d_s, d)$
$K: Key;\ (d_s, d)$
$V: Value;\ (d_s, d)$
$A: Attention\ weight;\ (d_s, d_s)$

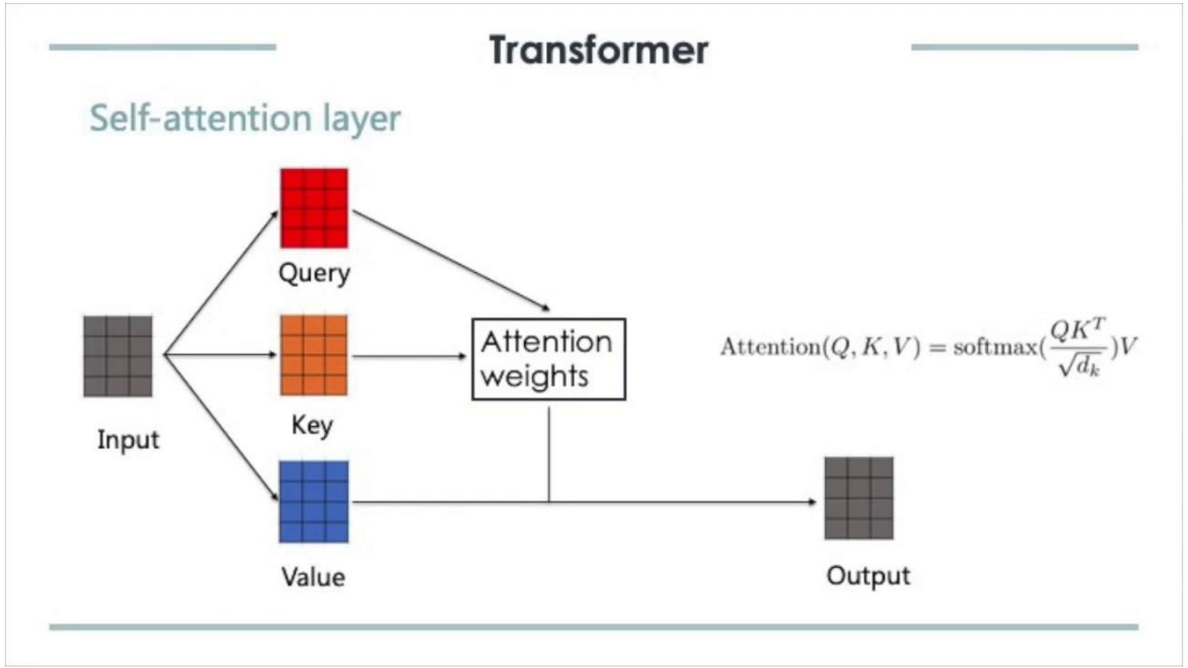
$W^Q, W^K, W^V\ are\ the\ trainable\ weights.$

Single-Head Self-Attention



After attention, we hopefully get a ***better embedding***.

Multi-Head Self-Attention



https://www.youtube.com/watch?v=5T38-2J5CcY

1. $for\ i^{th}$ $head\ of$ $\boldsymbol{total\ h\ heads}$,

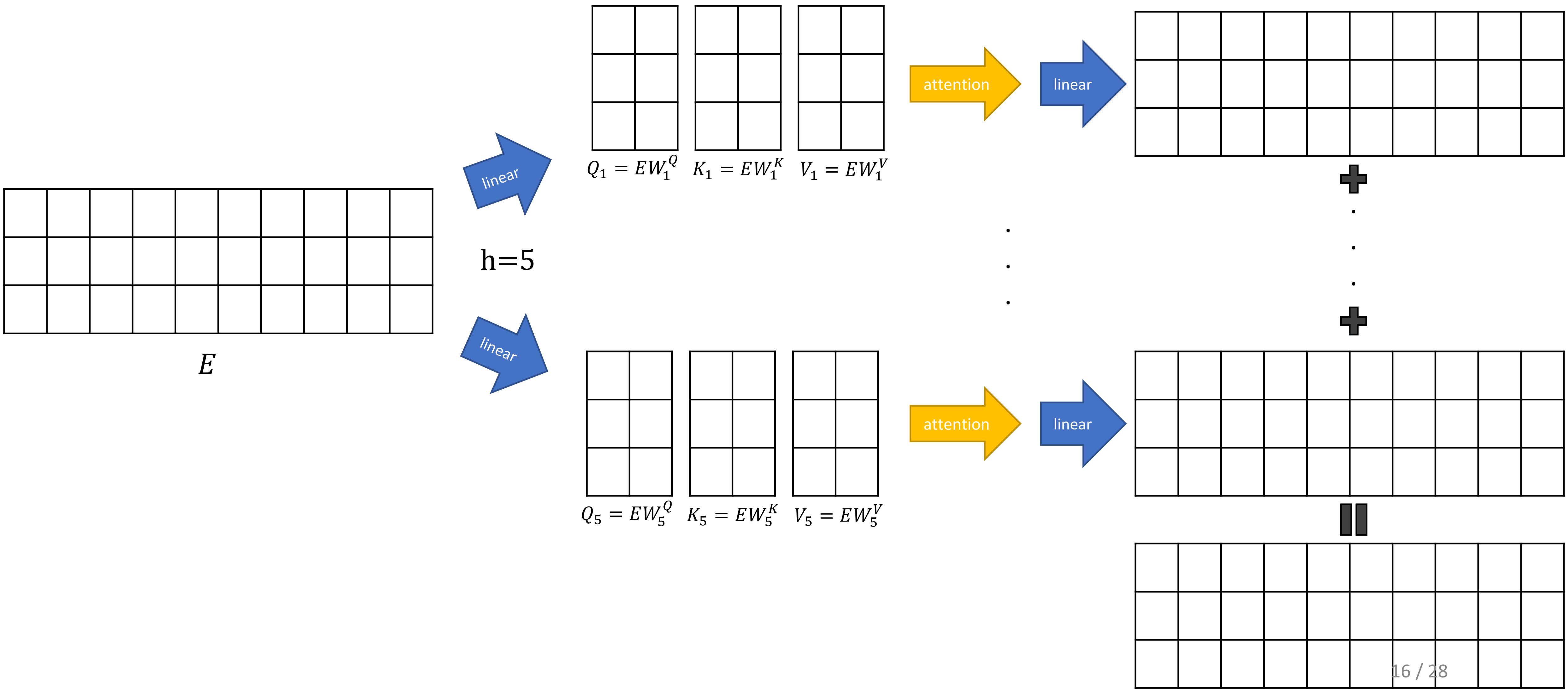
$$Q_i = EW_i^Q$$
$$K_i = EW_i^K$$
$$V_i = EW_i^V$$
$$A_i = \text{softmax}\left(\frac{Q_i K_i^T}{\sqrt{d_q}}\right)$$
$$O_i = A_i V_i$$
$$E_i = O_i W_i^H$$

$$2.\ E_{new} = \sum_{i=1}^{h} E_i$$

$E: embedding;\ (d_s, d)$
$Q: Query;\ (d_s, d_q), \boldsymbol{d_q = d/h}$
$K: Key;\ (d_s, d_q)$
$V: Value;\ (d_s, d_v), \boldsymbol{d_v = d/h}$
$A: Attention\ weight;\ (d_s, d_s)$
$W_i^H: trainable\ weight;\ (d_v, d)$

$W_i^Q, W_i^K, W_i^V, W_i^H$ $are\ the\ trainable\ weights.$

Multi-Head Self-Attention



Attention used for interpretation in BERT

$A: (d_s, d_s)$
(Attention weight)

	I	$miss$	you
I			
$miss$			
you			

Different attention weights for each head and each layer.

ex) BERT-base has 12 heads and 12 layers => 144 different attention weights

On Identifiability in Transformers

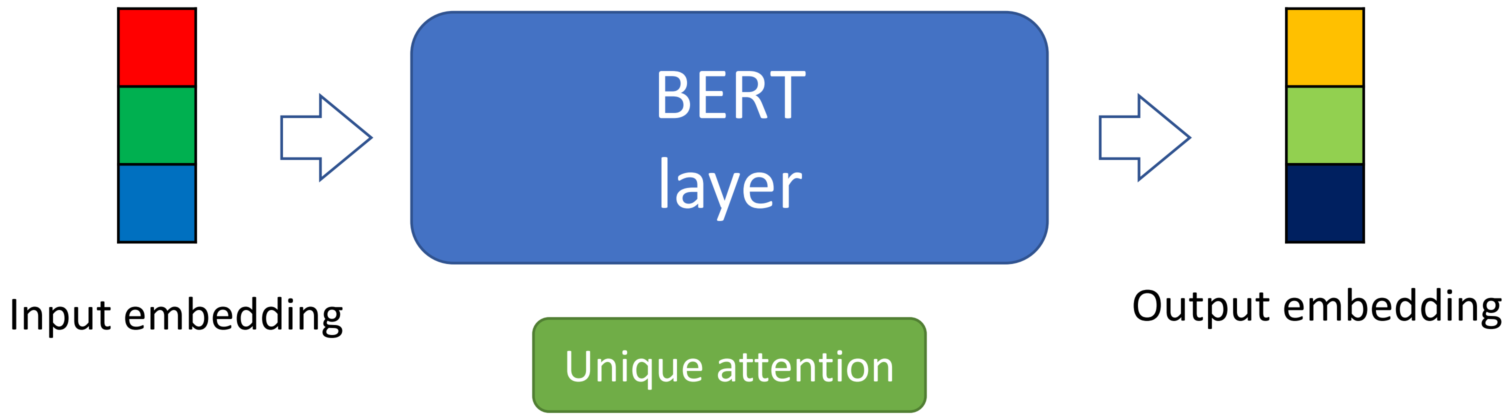
Contents

- Attention Identifiability
- Token Identifiability
- Token Mixing

Attention Identifiability

- Attention weights of an attention head are *identifiable* if they can be **uniquely** determined from the head's output.

- *Jain and Wallace, 2019, Attention is Not Explanation* had questioned the identifiability of attention.



Attention weights are *NOT* identifiable

- If the sequence length (d_s) is larger than the attention head dimension (d_v), attention is ***not*** unique.
  - *Is this a special case?* No. For BERT-base, d_s could reach up to 512, while d_v equals 64.

- Theoretical proof based on basic linear algebra.

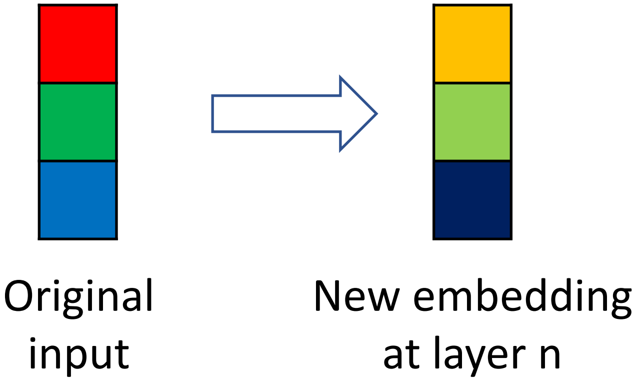
- Proposes ***effective attention***, which is part of the attention that actually affects the model output.

Token Identifiability

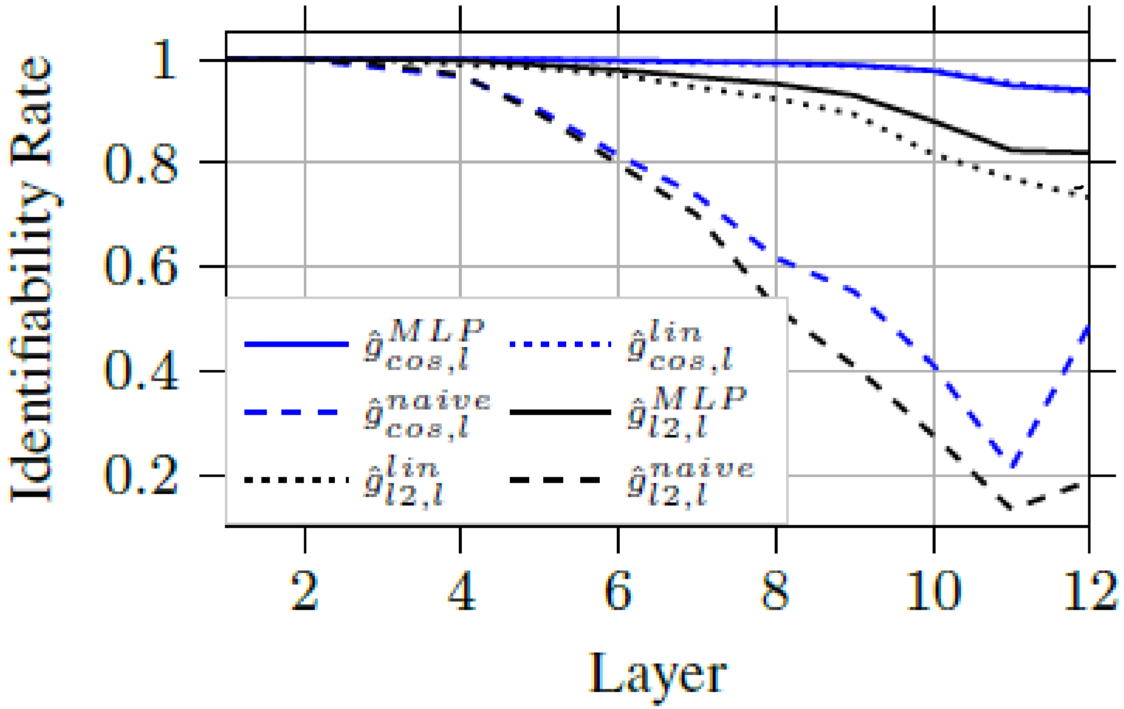
- Attention from the later layers are obtained from the new embedding, not the original input.
- If we were to use this attention to interpret the original input, we should ask:

Is token identifiable?

=> Can we recover the original input from its embedding?



Tokens are mostly identifiable

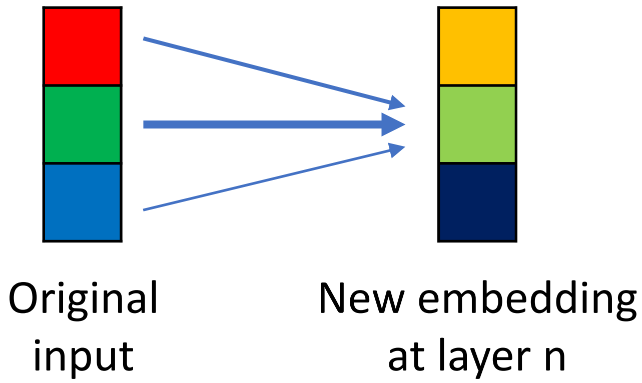


(Look at the solid and dotted lines)

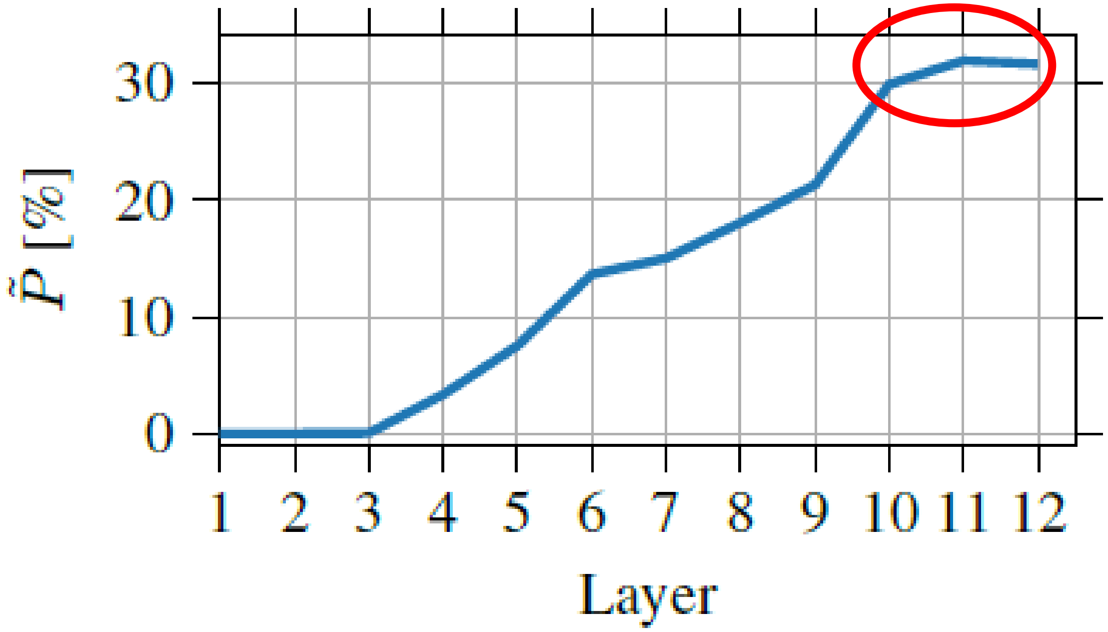
Token identifiability rate remains **high** throughout all the layers.

Token Mixing

- How much of the original input is still contained in the embedding?
- Proposes *hidden token attribution,* which is a gradient-based method.



Tokens are strongly mixed yet preserves some identity information



In the last layers, 30% of the tokens are ***not*** the highest contributor to their hidden embedding. => Quite high!

Ex) (Input: “Now almost done”) The word “done” is ***not*** the highest contributor to the embedding of “done” in the last layers, by 30% chance.

Follow-up Works

Follow-up Works

- Bhardwaj et al., 2021, More Identifiable yet Equally Performant Transformers for Text Classficiation, ACL – ***Attention identifiability***
  - Attention weights are more identifiable than previously claimed.
  - A variant of encoder layer which provides identifiability

- Pascual et al., 2021, Telling BERT's Full Story: from Local Attention to Global Aggregation, NAACL – ***Token mixing***
  - Distinction between local attention and global aggregation
  - More research on token mixing

- Sun et al., 2021, Effective Attention Sheds Light On Interpretability, Findings of ACL – ***Effective attention***
  - Interpretation based on effective attention vs raw attention

Thank You