Lecture Notes for

Neural Networks and Machine Learning

Practical Transformers
Vision Transformers

Logistics and Agenda

- Logistics
  - None!
- Agenda
  - Paper Presentation
  - Practical Transformers
  - Vision Transformers

Paper Presentation

Published as a conference paper at ICLR 2018

mixup: Beyond Empirical Risk Minimization

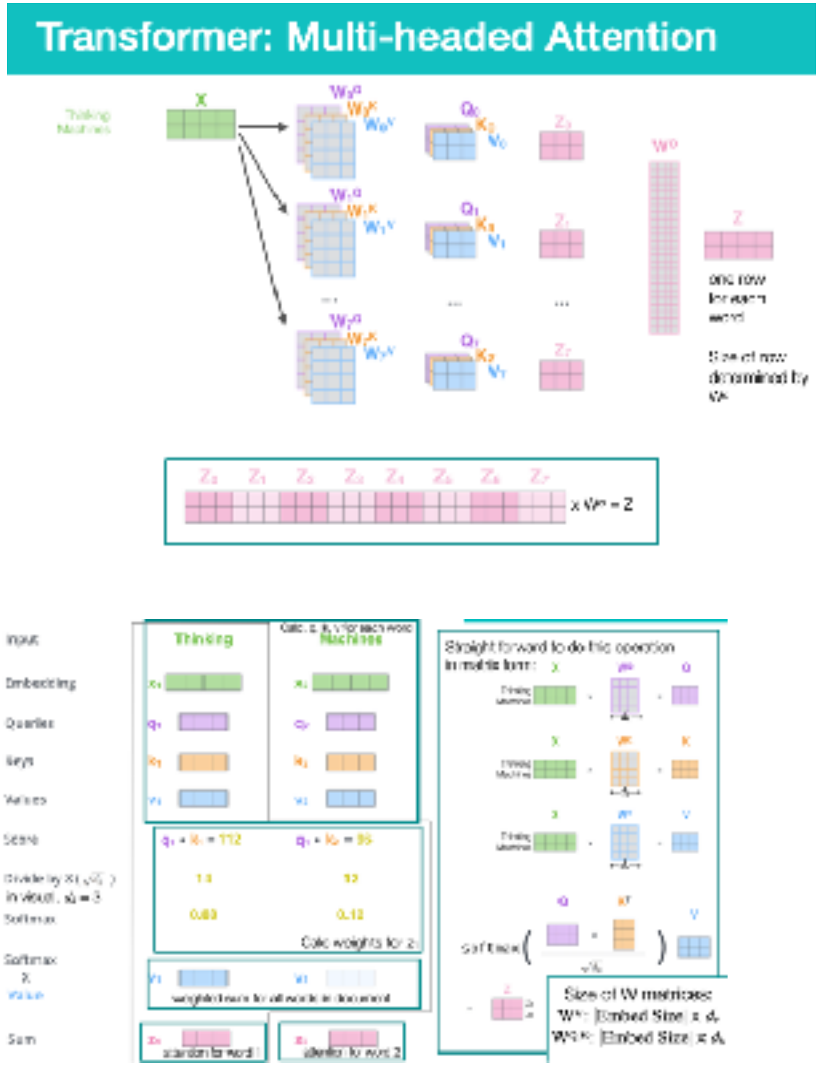
Hongyi Zhang
MIT

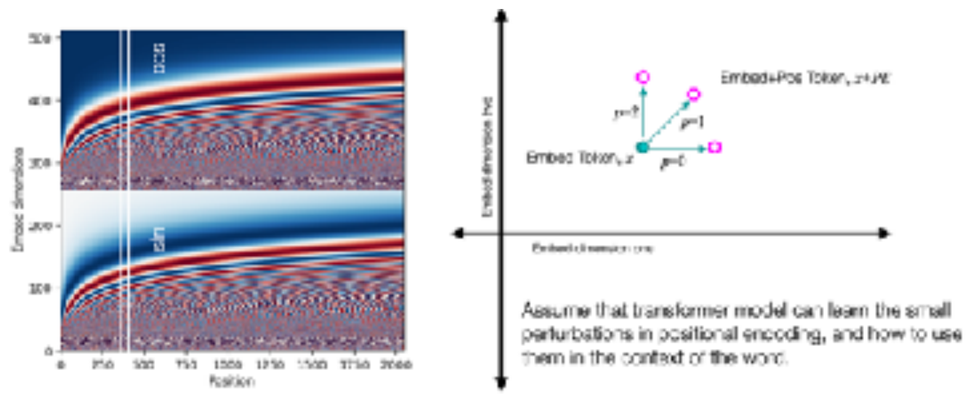
Moustapha Cisse, Yann N. Dauphin, David Lopez-Paz*
FAIR

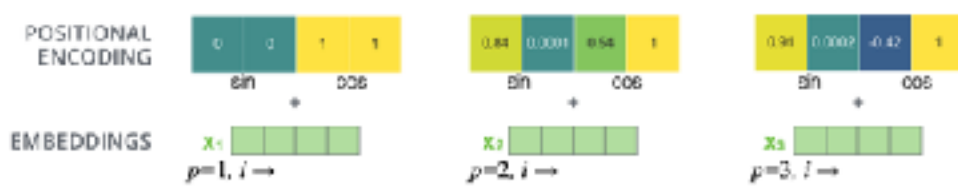
Abstract

Large deep neural networks are powerful, but exhibit undesirable behaviors such as memorization and sensitivity to adversarial examples. In this work, we propose *mixup*, a simple learning principle to alleviate these issues. In essence, *mixup* trains a neural network on convex combinations of pairs of examples and their labels. By doing so, *mixup* regularizes the neural network to favor simple linear behavior in-between training examples. Our experiments on the ImageNet-2012, CIFAR-10, CIFAR-100, Google commands and UCI datasets show that *mixup* improves the generalization of state-of-the-art neural network architectures. We also find that *mixup* reduces the memorization of corrupt labels, increases the robustness to adversarial examples, and stabilizes the training of generative adversarial networks.

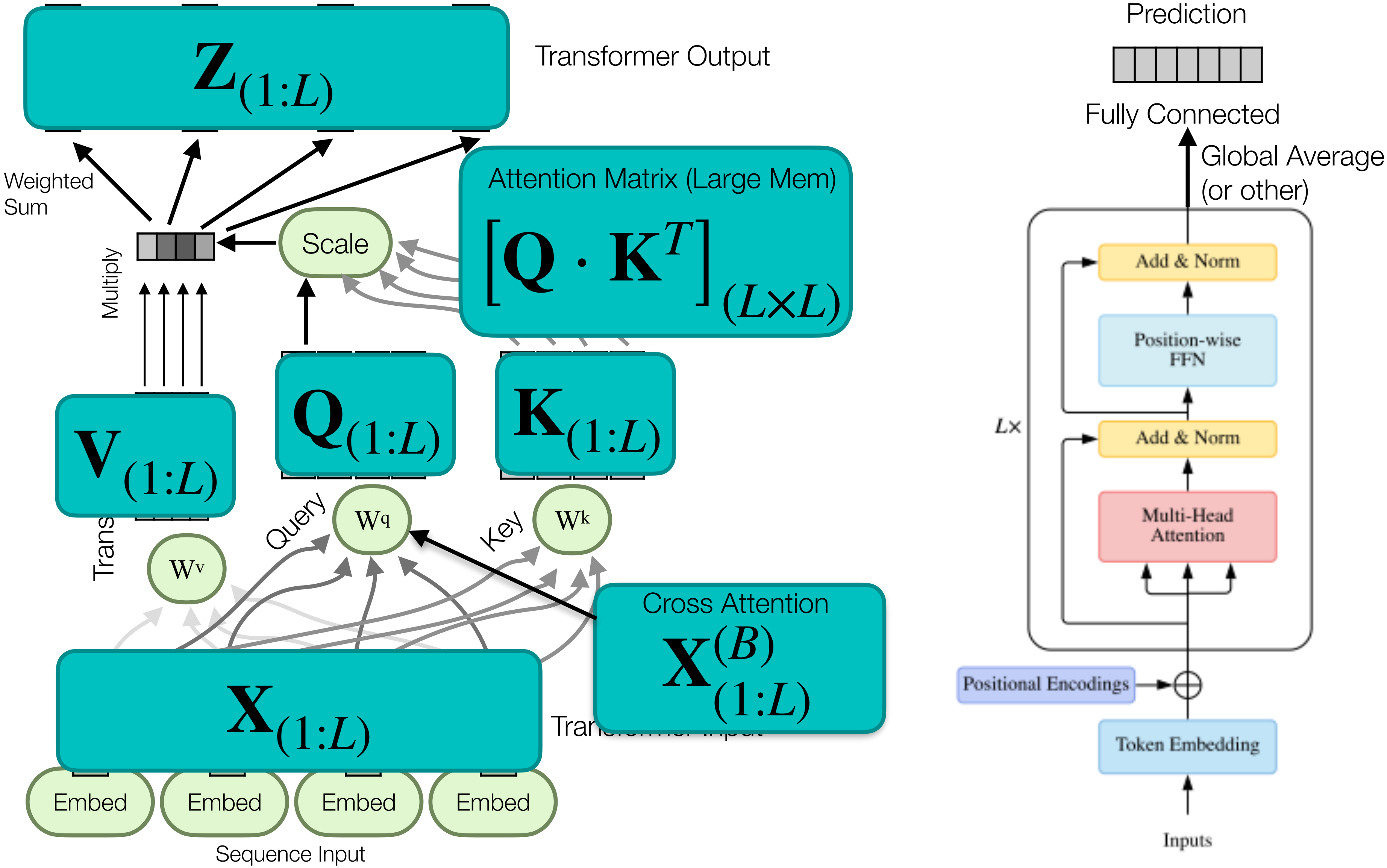
Last Time: Transformers





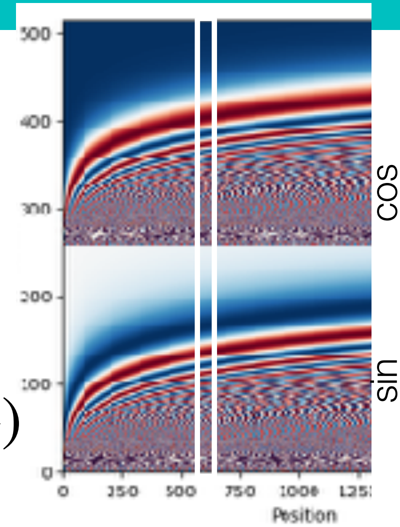


Transformer Review



Transformer: Positional Encoding

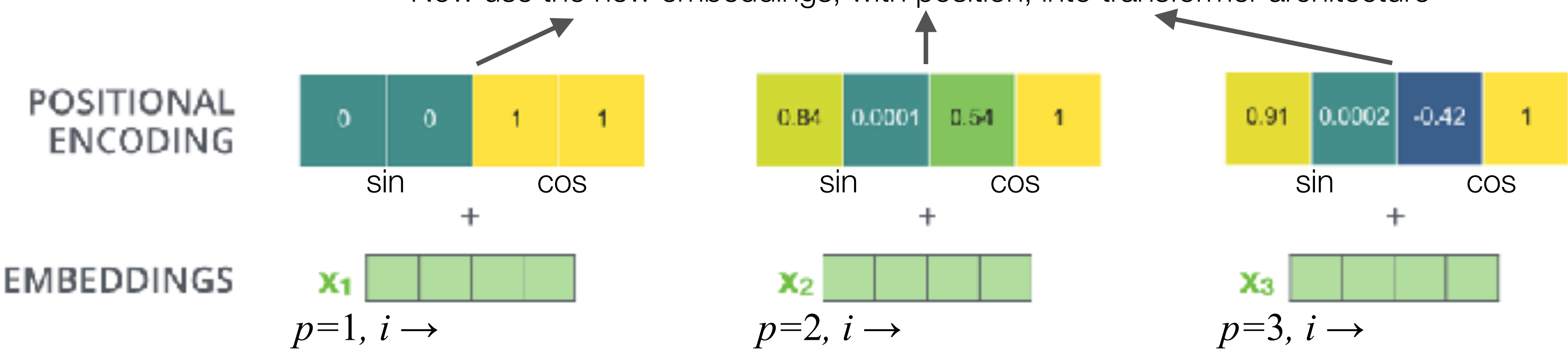
- Objective: add notion of position to embedding
- Attempt in paper: add sin/cos to embedding



p: in sequence
d_m: 1/2 dim of embed
i = index in vector

$$PE_{(p,i\in 0\ldots d_m-1)} = \sin(p/10000^{i/d_m})$$
$$PE_{(p,i\in d_m\ldots 2d_m)} = \cos(p/10000^{(i-d_m)/d_m})$$

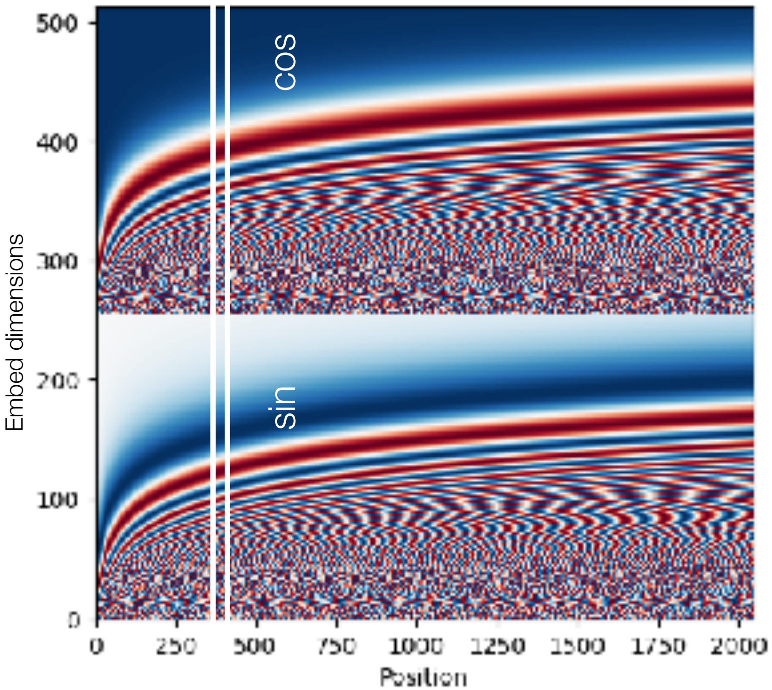
Now use the new embeddings, with position, into transformer architecture

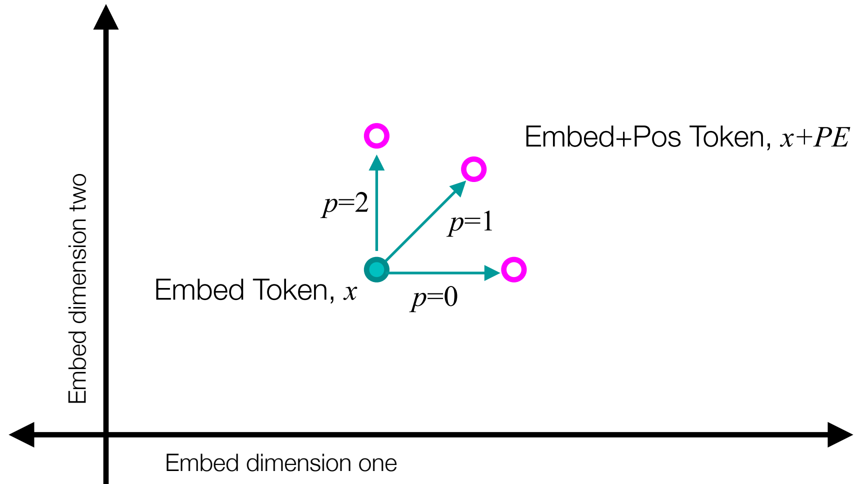


Hypothesis: Now the word proximity is encoded in the embedding matrix, with other pertinent information. Well, it does help… so it could be true that this is a good way to do it.

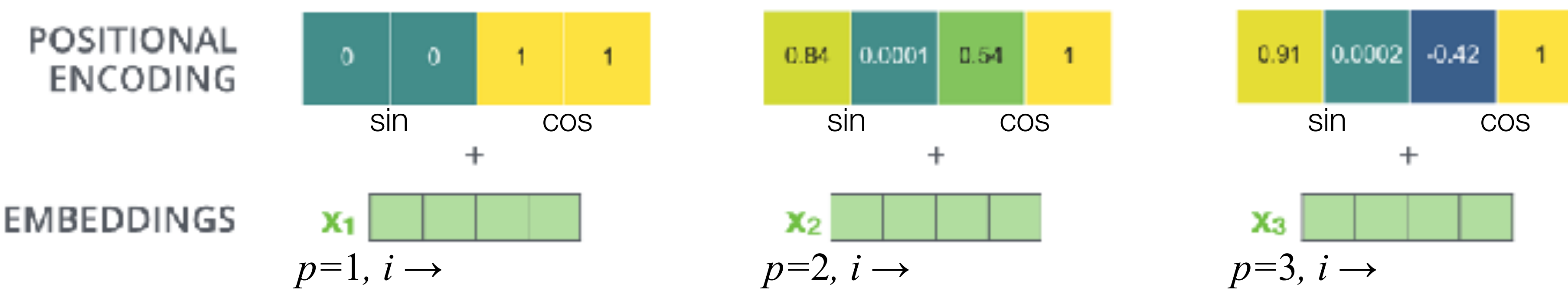
Excellent Blog on Transformers: http://jalammar.github.io/illustrated-transformer/

Positional Intuition, Geometrically



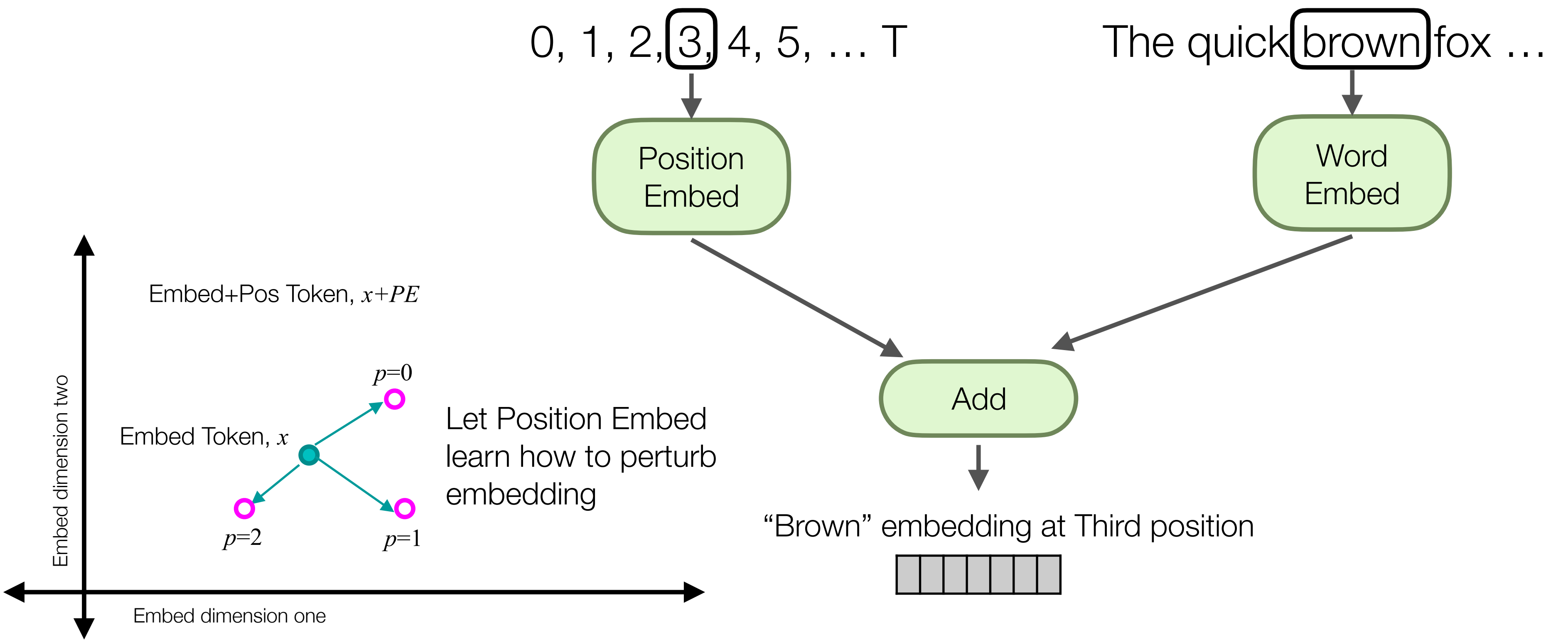


Assume that transformer model can learn the small perturbations in positional encoding, and how to use them in the context of the word.



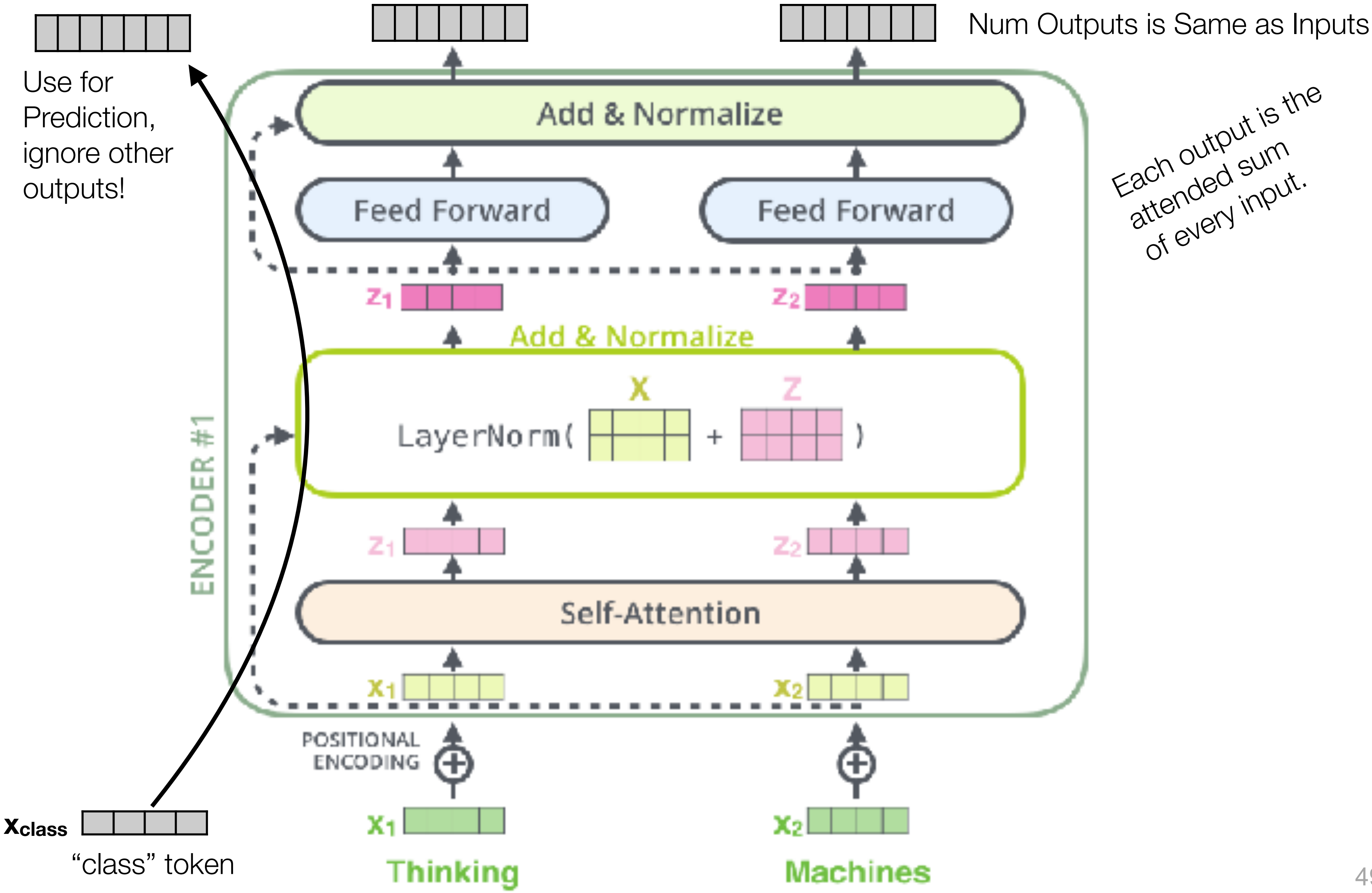
Transformer: Positional Encoding

- Objective: add notion of position to embedding
- Attempt in original paper: add sin/cos to embedding
- **But could be anything that encodes position, like:**

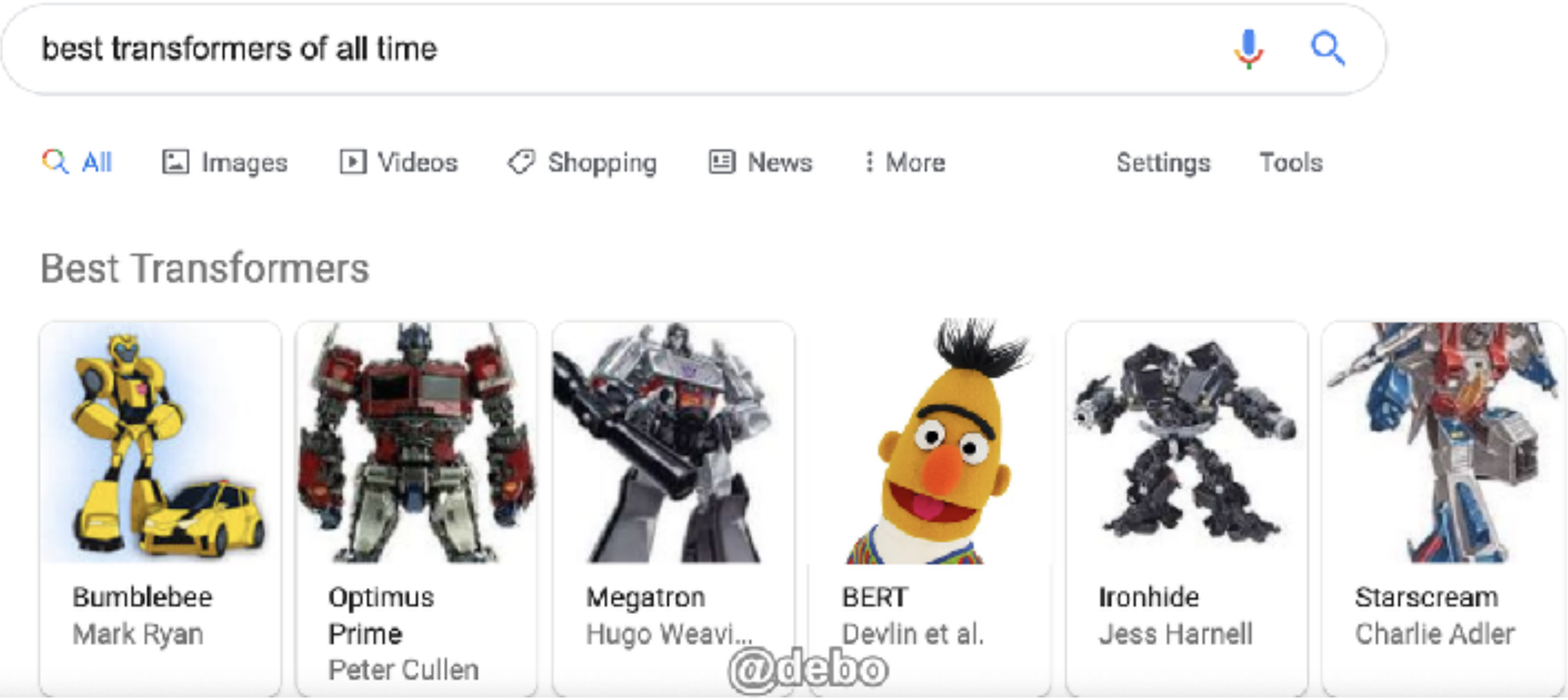


Excellent Blog on Transformers: http://jalammar.github.io/illustrated-transformer/

Transformer: Putting it all together

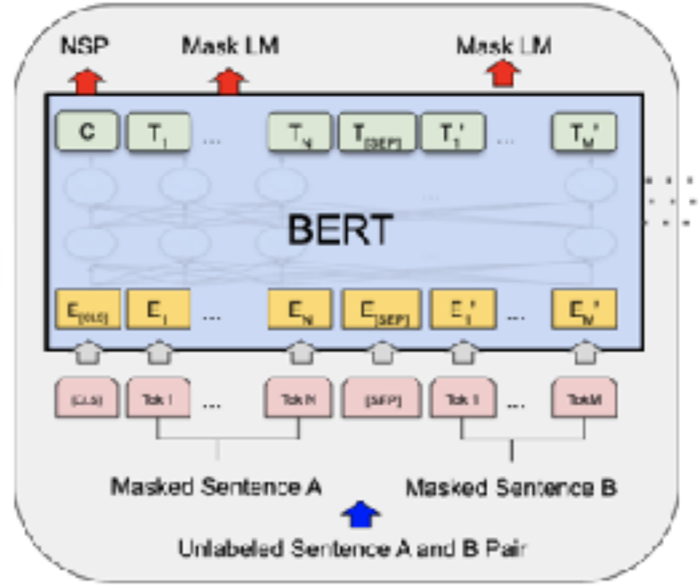


Encoder Transformers



Bidirectional Encoder Representation

- Google, 2018. Vocab: 30k words
- Bidirectional (non-causal attention)
- BERT$_{\text{Base}}$
  - 12 encoder layers, 12 heads/layer
  - 110M parameters
- BERT$_{\text{Large}}$
  - 24 encoder layers, 16 heads/layer
  - 340M parameters
- Various ways to fine tune



Pre-training

Masked Language Modeling (LM)

“I am **[MASK1]** in CS8321 at SMU. This class is **[MASK2]**”
MASK1: **“enrolled”**   MASK2: **“great”**

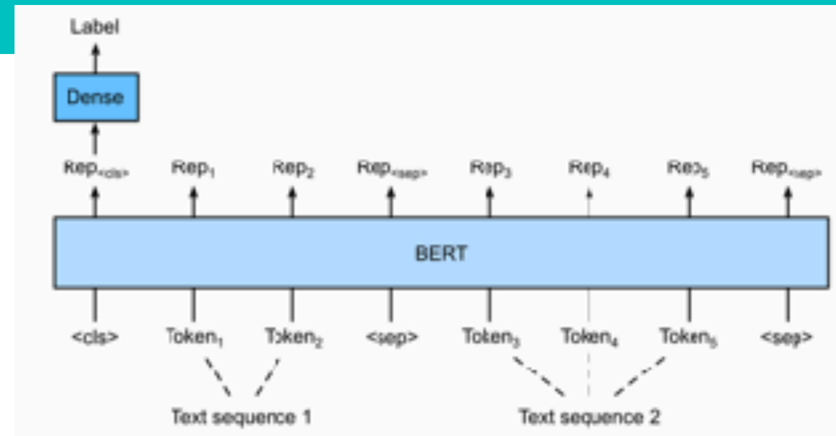
Next Sentence Prediction (NSP)

"[CLS] Dr. Larson is a professor [SEP] his class examples are great” → **Label “IsNext”**
"[CLS] Dr. Larson is a professor [SEP] do you like bread” → **Label “NotNext”**

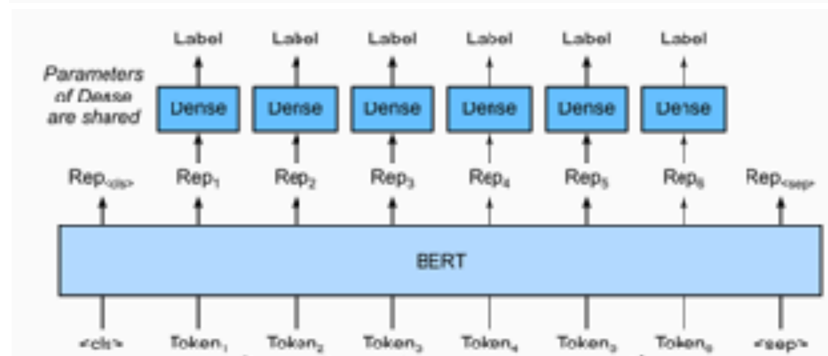
Devlin et al. ”BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". arXiv:1810.04805v2

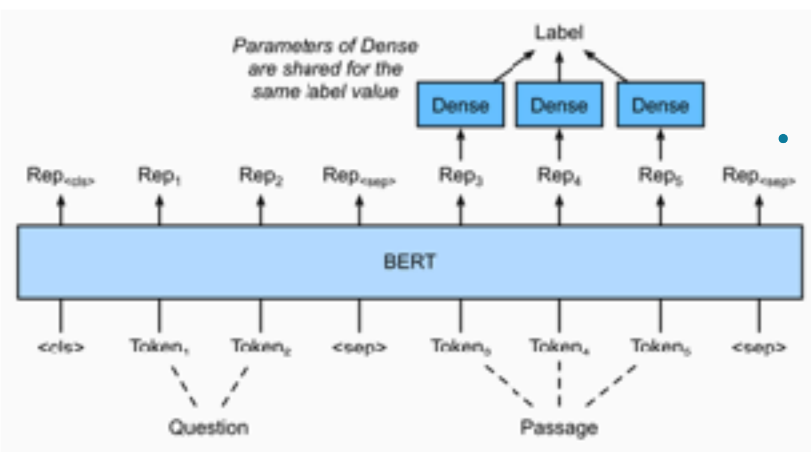
Fine Tuning BERT

- Sentence predict: like Text Similarity
  - Make use of NSP
  - Two sentences, do they belong?



- Part of speech tagging
  - Make use of Masked LM
  - Shared dense layer for each Rep





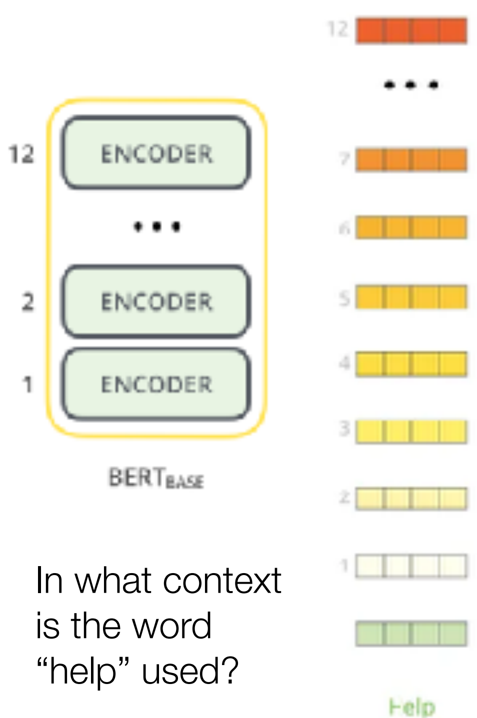
- Question Answering (Stanford QA Dataset, SQuaD)
  - Make use of Masked LM
  - Highlight passage text that answers given question

Q: Who currently teaches machine learning at SMU?
P: "Machine learning was first offered at SMU in the 1990's. **Dr. Larson** has been teaching the course since 2014 and has changed it into a neural networks course, despite its origins. "

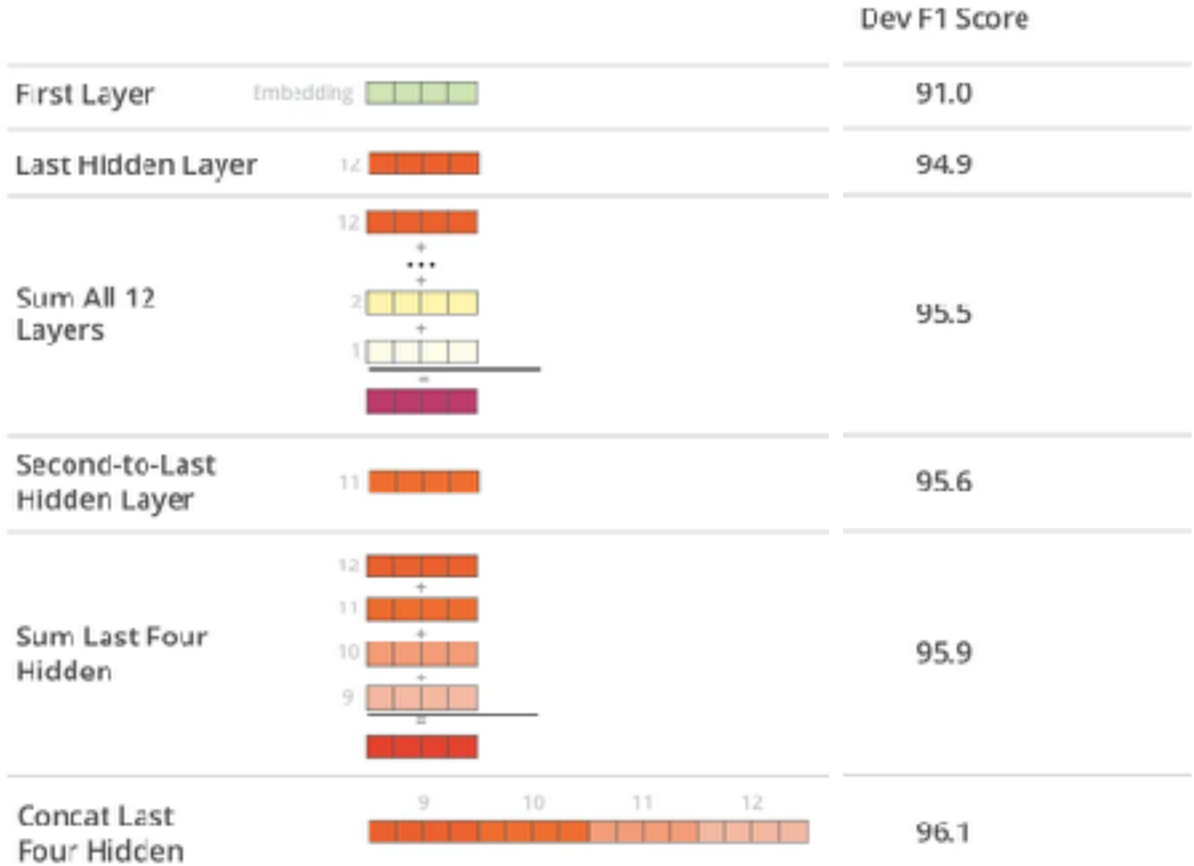
https://classic.d2l.ai/chapter_natural-language-processing-applications/finetuning-bert.html

Fine Tuning BERT

- Could we use more than just the final output layers?



Each word has 12 layer outputs in $\text{BERT}_{\text{Base}}$



In what context is the word "help" used?

Noun, verb, adverb, etc.

Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". arXiv:1810.04805v2

MELVA Results (my lab)

- Measuring English Language Vocabulary Acquisition
- Or results from my lab:
    - Using science terms in sentence?
    - Collect/transcribe responses
- Collected about 6000 sentences

- Transfer learn based upon LM output
    - Without transformer LM: ~75%
    - With transformer LM: ~84%

L@S '23, July 20–22, 2023, Copenhagen, Denmark

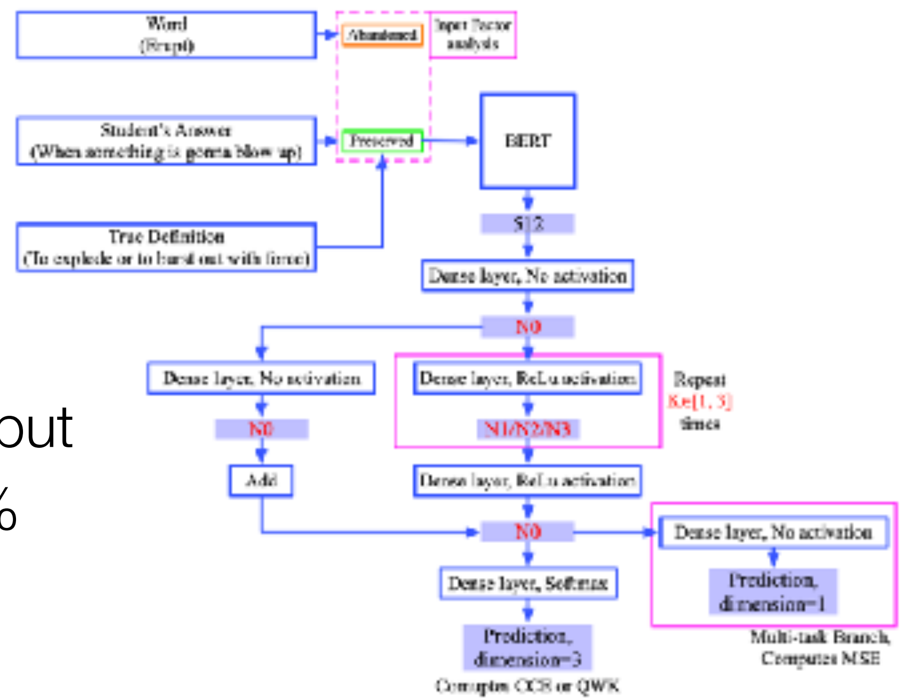
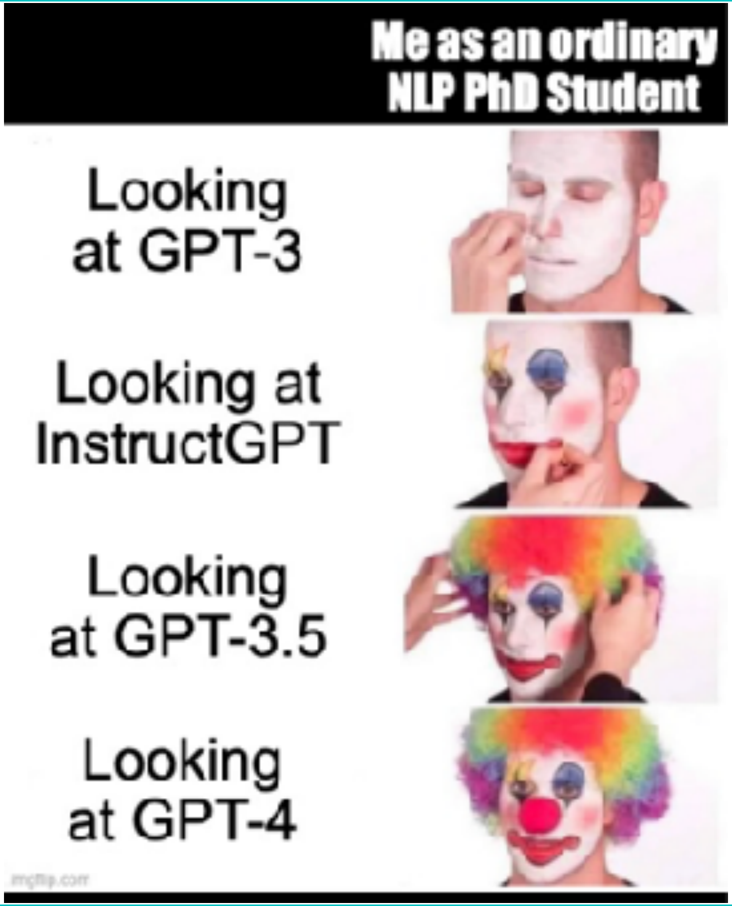


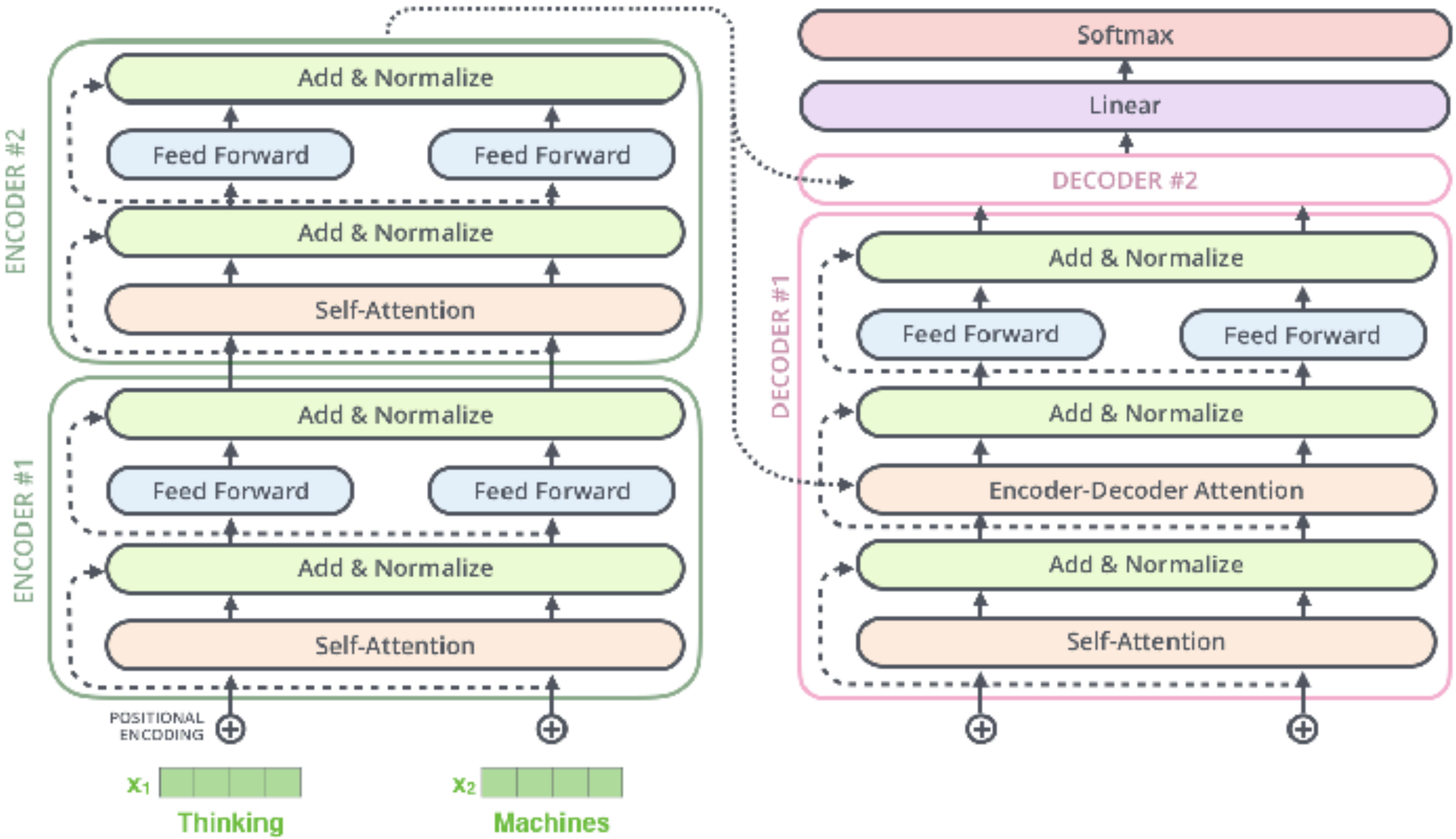
Figure 1: Example of the end-to-end pipeline of the network. Variables marked in red are found through hyperparameter search.

Zhongdi Wu, Larson, E., Makoto Sano, Doris Baker, Akihito Kamata, & Nathan Gage (2023) Towards Scalable Vocabulary Acquisition Assessment with BERT. Learning at Scale, 5. 10.1145/3573051.3596170

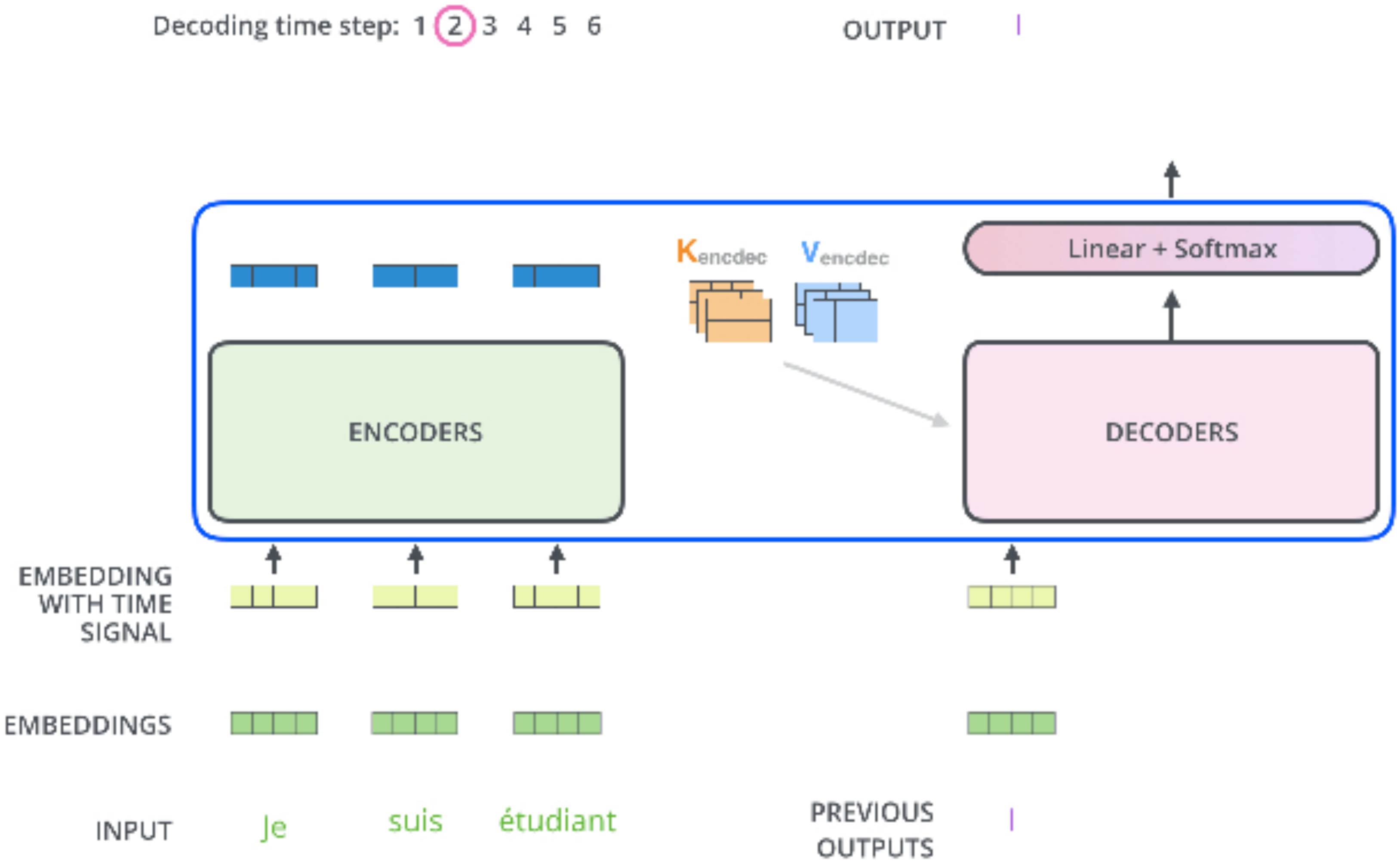
Encoder+Decoder Xformer



Transformer: Encoders and Decoders



Transformer: Putting it all together

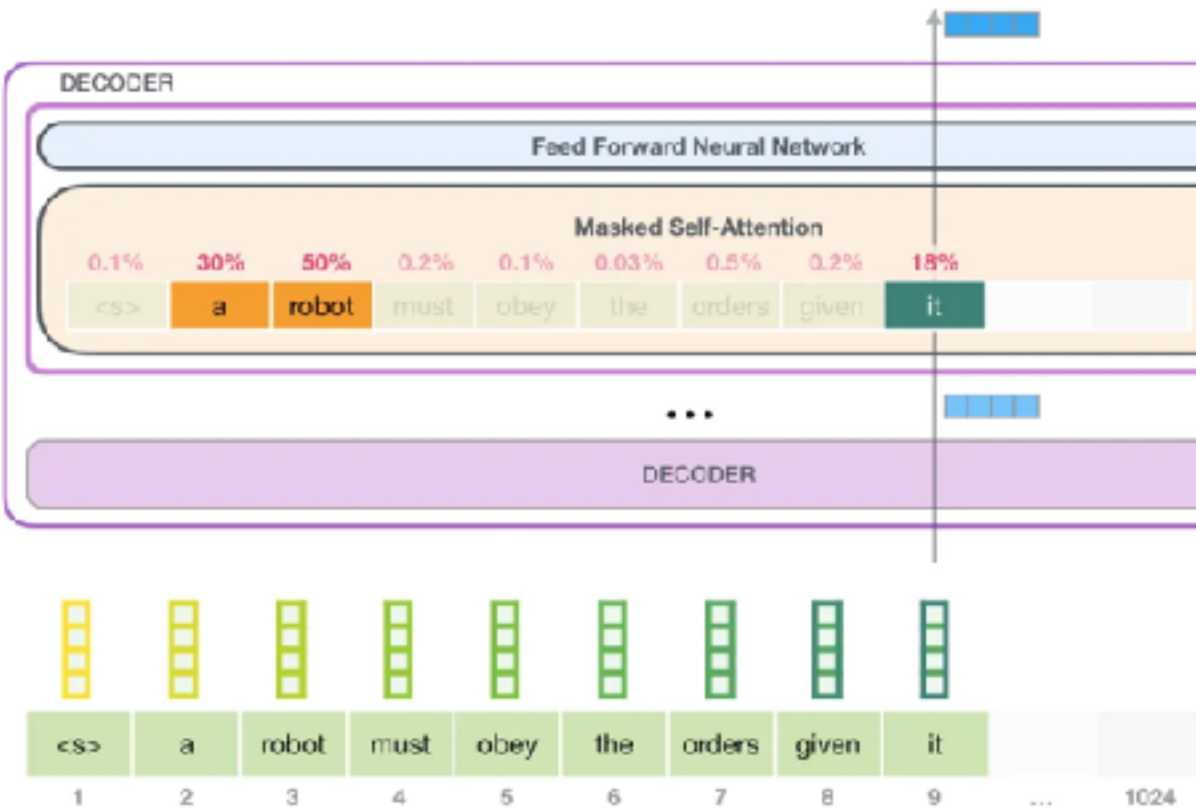


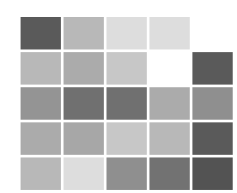
Auto-regressive Transformer

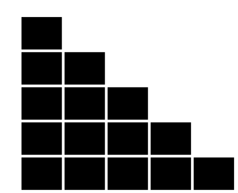
- Essentially: decoder only, text encoding happens in first attention layer
- Generative pre-training (GPT)

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

Predict the next word from unlabeled dataset




Attention Map


Mask

Radford, et al. Improving Language Understanding by Generative Pre-Training, ArXiV 20

Lecture Notes for

Neural Networks and Machine Learning

Transformers

Next Time:
SSL, Vision Transformers

Reading: None