Want to get more attention on social network? Try our **A**ttention-**B**ased **B**ilingual **I**mage2caption **E**moji Model (**ABBIE**)

Yi-Hui Lee

yi-hui.lee@utdallas.edu

Outline

- Motivation
- Brief summary of previous work / background slide(s)
  - Transformer
  - Bilingual Transformer
  - Image Captioning through Image Transformer
- Topic summary covering key ideas slides
  - **A**ttention-**B**ased **B**ilingual **I**mage2caption **E**moji Model (**ABBIE**)
- References (not included in the <= 20 slide limit)

Motivation

- Transformer (Attention is all you need) utilize self-attention mechanism to pre-train the seq-2-seq model
    - It’s the fundamental component (seq-2-seq) of our ABBIE model
    - We will talk more about different transformer models on the following slides
- Bilingual transformer allow model to share latent vector in different language which is very helpful for our ABBIE model.
- Image Captioning through Image Transformer builds the model from image to text and emoji

Motivation

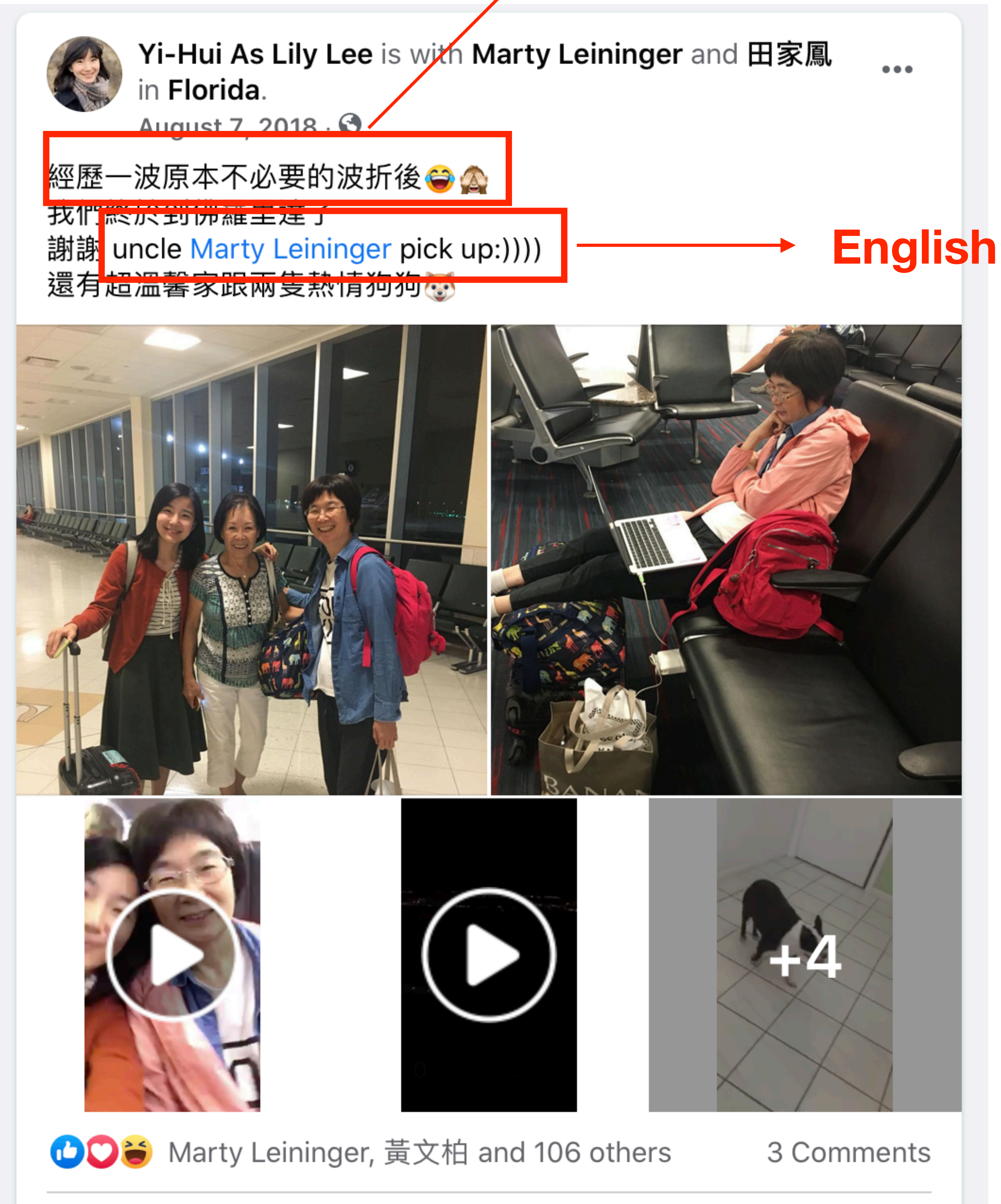
- Let't talk about the idea of **ABBIE**

- We share our life experience on social network and often with a nice picture

- It's not always easy to come up with a nice caption that catch friends and family's eyes

- We loved to announced our super helpful model **A**ttention-**B**ased **B**ilingual **I**mage2caption **E**moji Model (**ABBIE**), that transform the image pixels into the bilingual caption and also with the corresponding emoji.

- To make our dream come true, we top-down our model into three big substructure: transformer, bilingual language model, and image caption model.



Yi-Hui As Lily Lee is with Marty Leininger and 田家鳳 in Florida.
August 7, 2018 · 🌐

經歷一波原本不必要的波折後😂🙈
我們終於到佛羅里達了
謝謝 uncle Marty Leininger pick up:))))
還有超溫馨家跟兩隻熱情狗狗🐶

+4

Marty Leininger, 黃文柏 and 106 others 3 Comments

Chinese Caption with corresponding emoji

English Caption

Figure from:
https://www.facebook.com/amy030619/posts/1791084910938787

Transformer

- BERT: a deep bidirectional transformers that utilize the idea of Cloze task (Masking input)

- Using context to predict the missing token

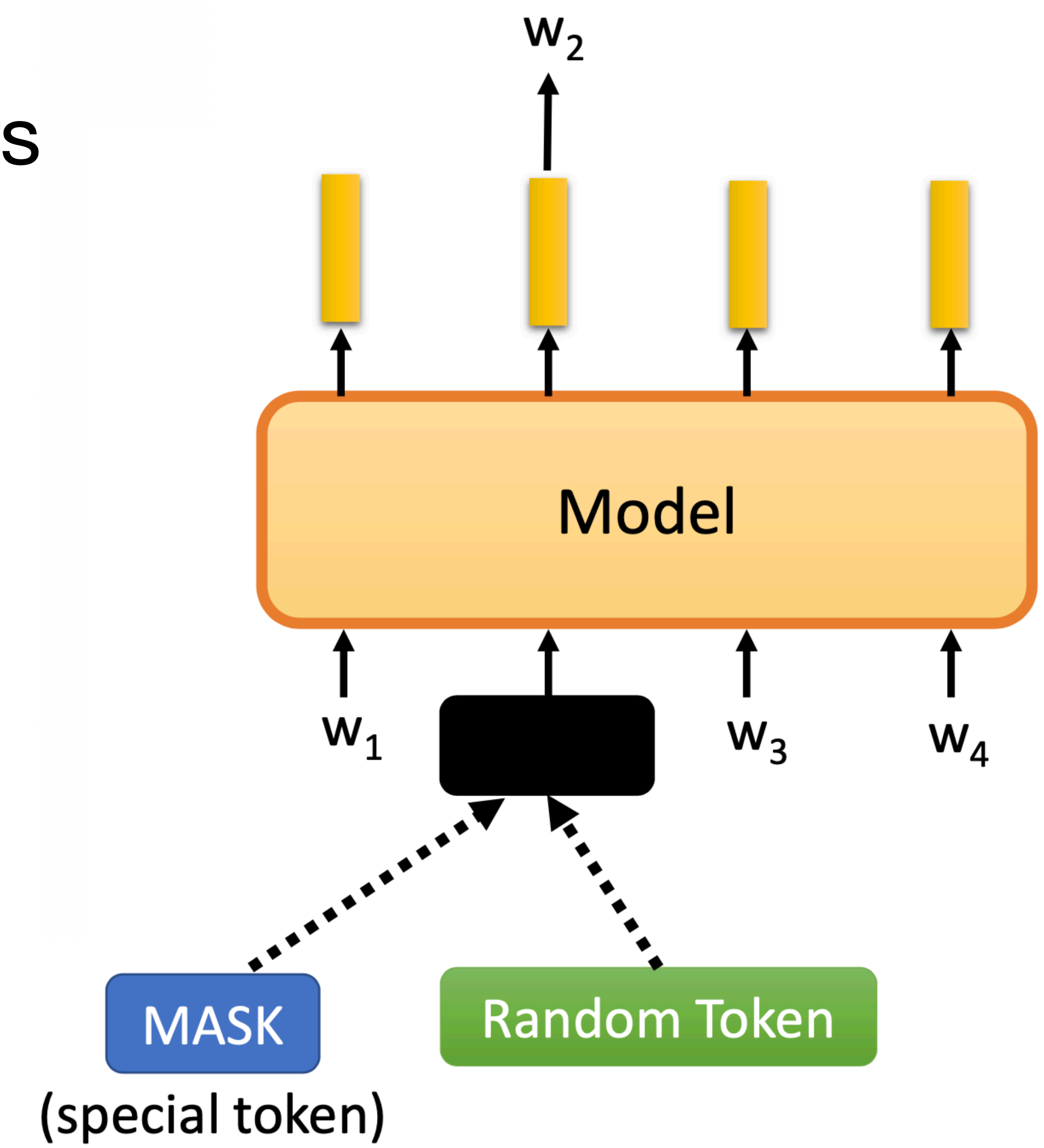


Figure from:
http://speech.ee.ntu.edu.tw/~tlkagk/courses/DLHLP20/BERT%20train%20(v8).pdf

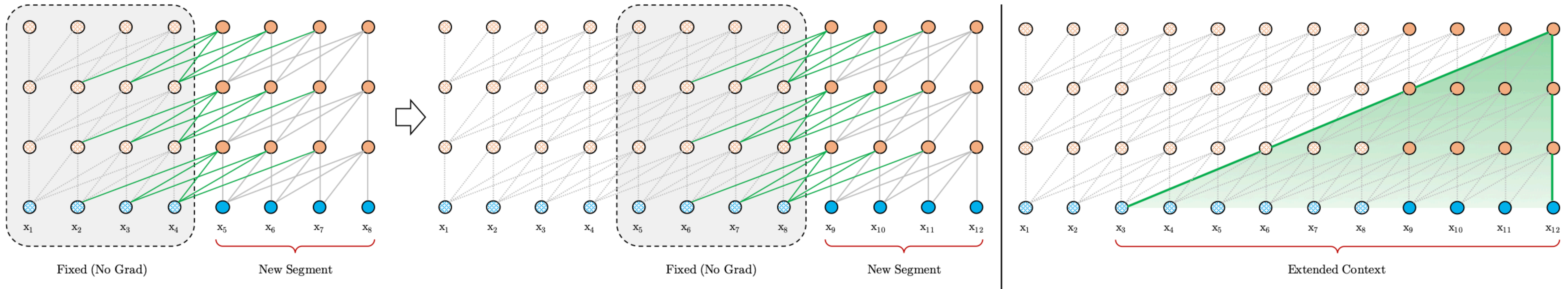
Transformer

- Other transformer:
  - Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context
  - XLNet: Generalized Autoregressive Pretraining for Language Understanding
  - RoBERTa: A Robustly Optimized BERT Pretraining Approach
  - ALBERT: A Lite BERT for Self-supervised Learning of Language Representations
  - DistilBERT: a distilled version of BERT: smaller, faster, cheaper and lighter
  - BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension
  - Reformer: The Efficient Transformer
  - Linformer: Self-Attention with Linear Complexity
- Helpful tools: Hugging face transformers

Transformer-XL

- Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context

- Segment-Level Recurrence with State Reuse


x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8
Fixed (No Grad)
New Segment
x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8 x_9 x_{10} x_{11} x_{12}
Fixed (No Grad)
New Segment
x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8 x_9 x_{10} x_{11} x_{12}
Extended Context

(a) Training phase.

(b) Evaluation phase.

Figure 2: Illustration of the Transformer-XL model with a segment length 4.

Figure from:
https://arxiv.org/pdf/1901.02860.pdf

XLNet

- XLNet: Generalized Autoregressive Pretraining for Language Understanding

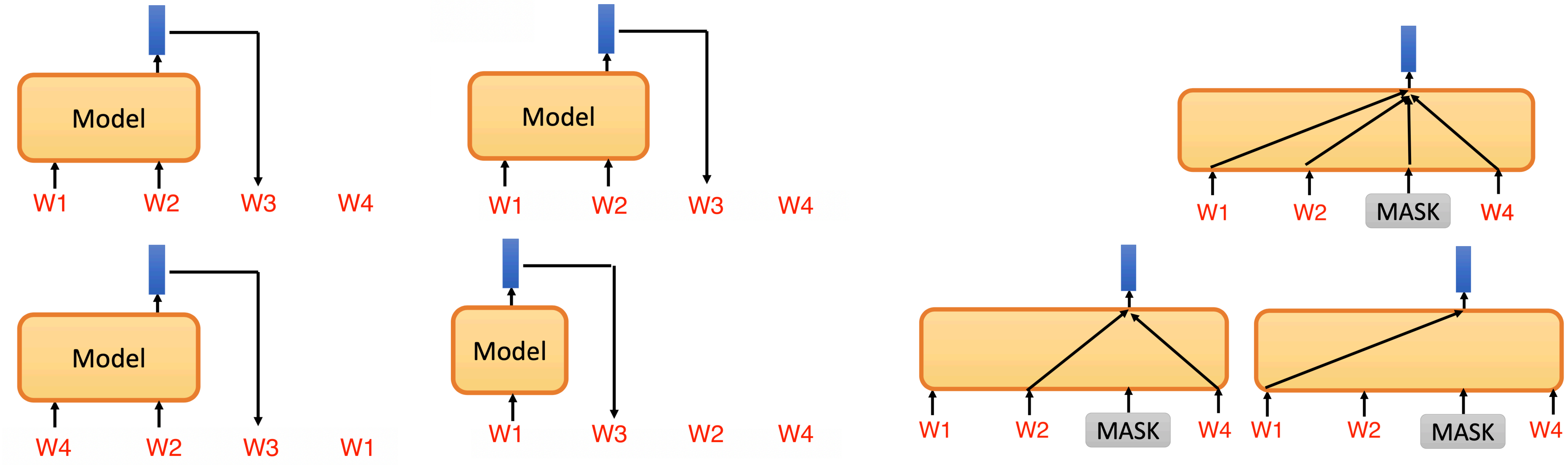


Figure edit from:
http://speech.ee.ntu.edu.tw/~tlkagk/courses/DLHLP20/BERT%20train%20(v8).pdf

RoBERTa

- RoBERTa: A Robustly Optimized BERT Pretraining Approach
- RoBERTa has a slightly change on training BERT, that trained on more time and on different hyper parameters
  - Train longer, bigger batches, bigger training data
  - Removing the next sentence prediction objective
  - Training on longer sequences
  - Dynamically changing the masking pattern applied to the training data

ALBERT

- ALBERT: A Lite BERT for Self-supervised Learning of Language Representations
- Lower down the parameters in the traditional BERT architecture
  - Factorized embedding parameterization: separate the size of the hidden layers from the size of vocabulary embedding by decomposing the large vocabulary embedding matrix into two small matrices
  - The separation makes it easier to grow the hidden size without significantly increasing the parameter size of the vocabulary embeddings.
- Cross-layer parameter sharing
  - Prevents the parameter from growing with the depth of the network

DistilBERT

- DistilBERT: a distilled version of BERT: smaller, faster, cheaper and lighter

- Pre-trained with knowledge distillation that makes DistilBERT becomes much smaller language models

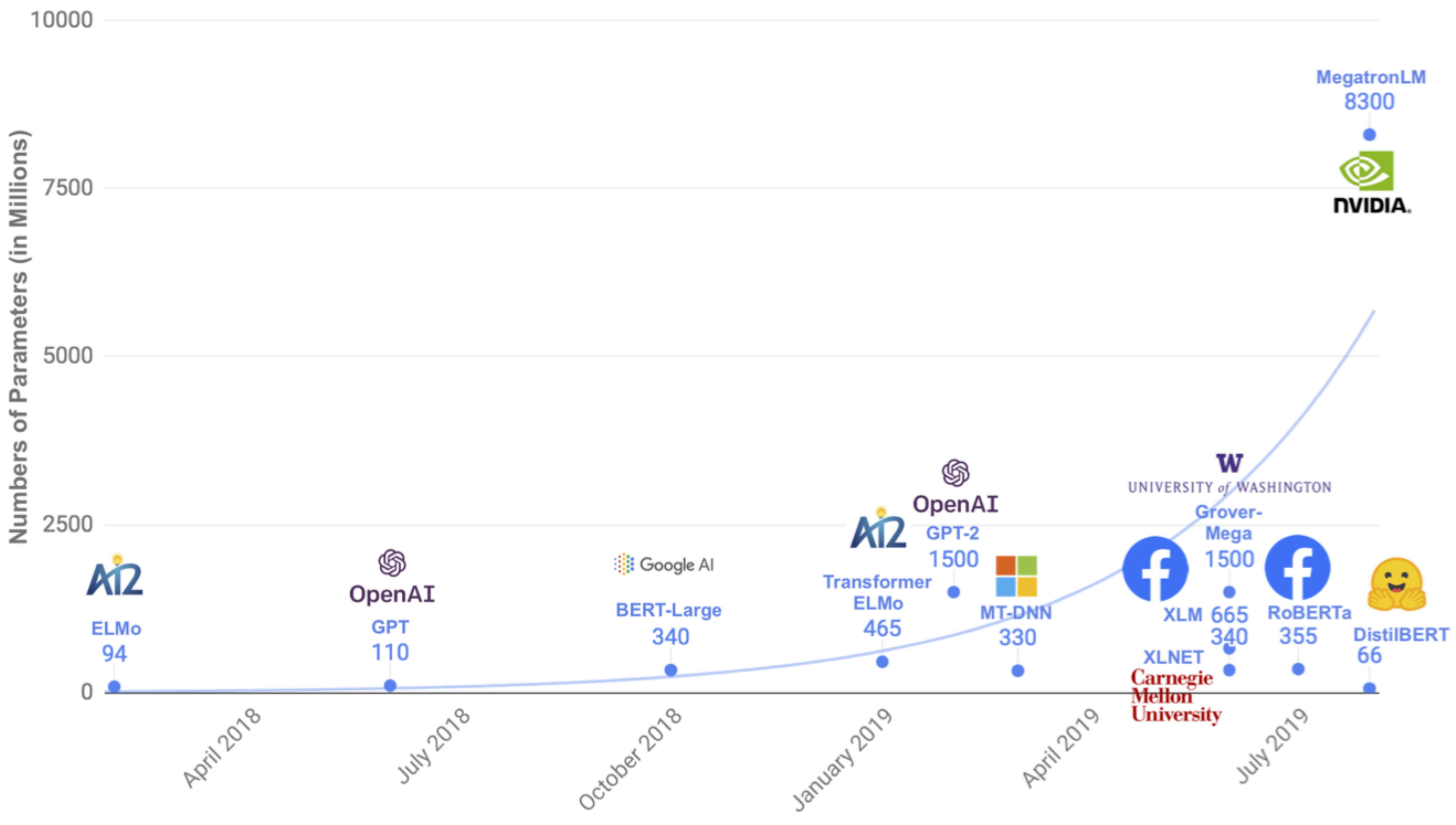
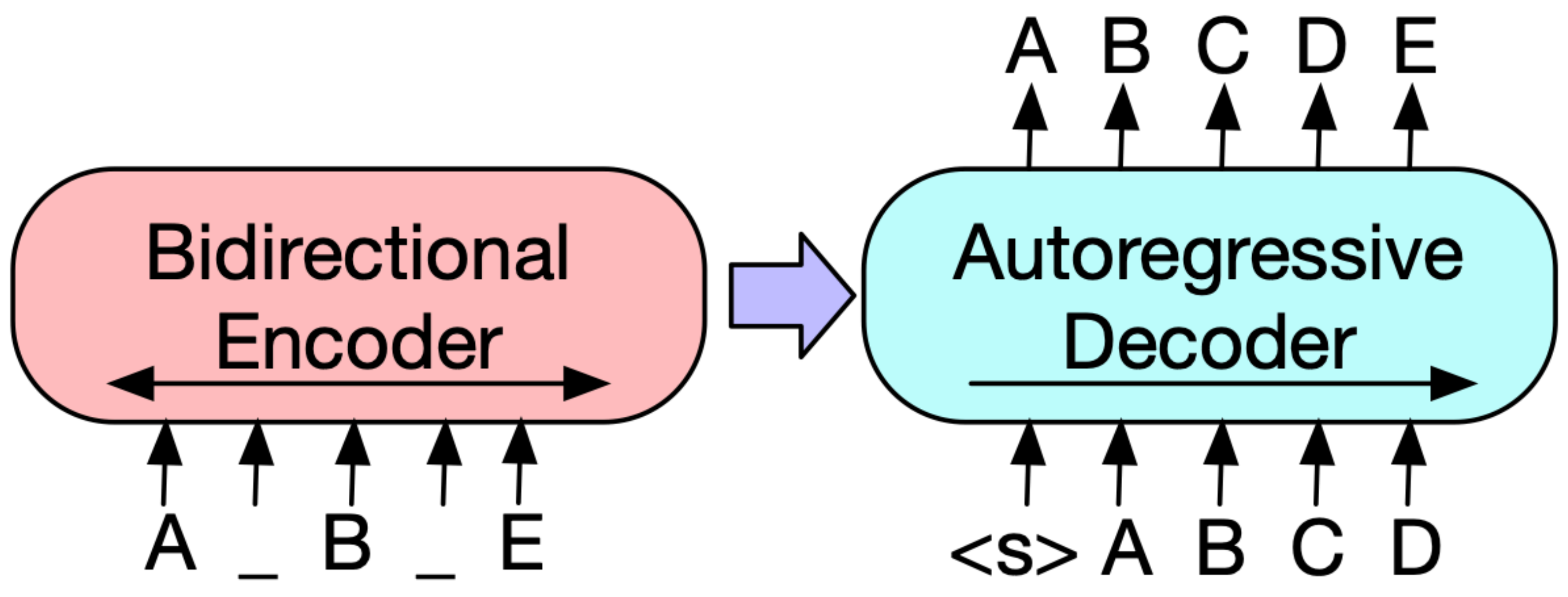


Figure 1: **Parameter counts of several recently released pretrained language models.**

Figure from:
https://arxiv.org/pdf/1910.01108.pdf

BART

- BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension



(c) BART: Inputs to the encoder need not be aligned with decoder outputs, allowing arbitary noise transformations. Here, a document has been corrupted by replacing spans of text with mask symbols. The corrupted document (left) is encoded with a bidirectional model, and then the likelihood of the original document (right) is calculated with an autoregressive decoder. For fine-tuning, an uncorrupted document is input to both the encoder and decoder, and we use representations from the final hidden state of the decoder.

Figure from:
https://arxiv.org/pdf/1910.13461.pdf

Reformer

- Reformer: The Efficient Transformer

- Faster and more memory-efficient when train on long sequences

    - Reduce the complexity of self-attention: replace dot-product attention by one that uses locality-sensitive hashing

    - Utilize reversible residual layers instead of the standard residuals: allows storing activations only once in the training process instead of N times (N = number of layers)

Linformer

- Linformer: Self-Attention with Linear Complexity
- More time-efficient and more memory-efficient
  - Self-attention mechanism be approximated by a low-rank matrix

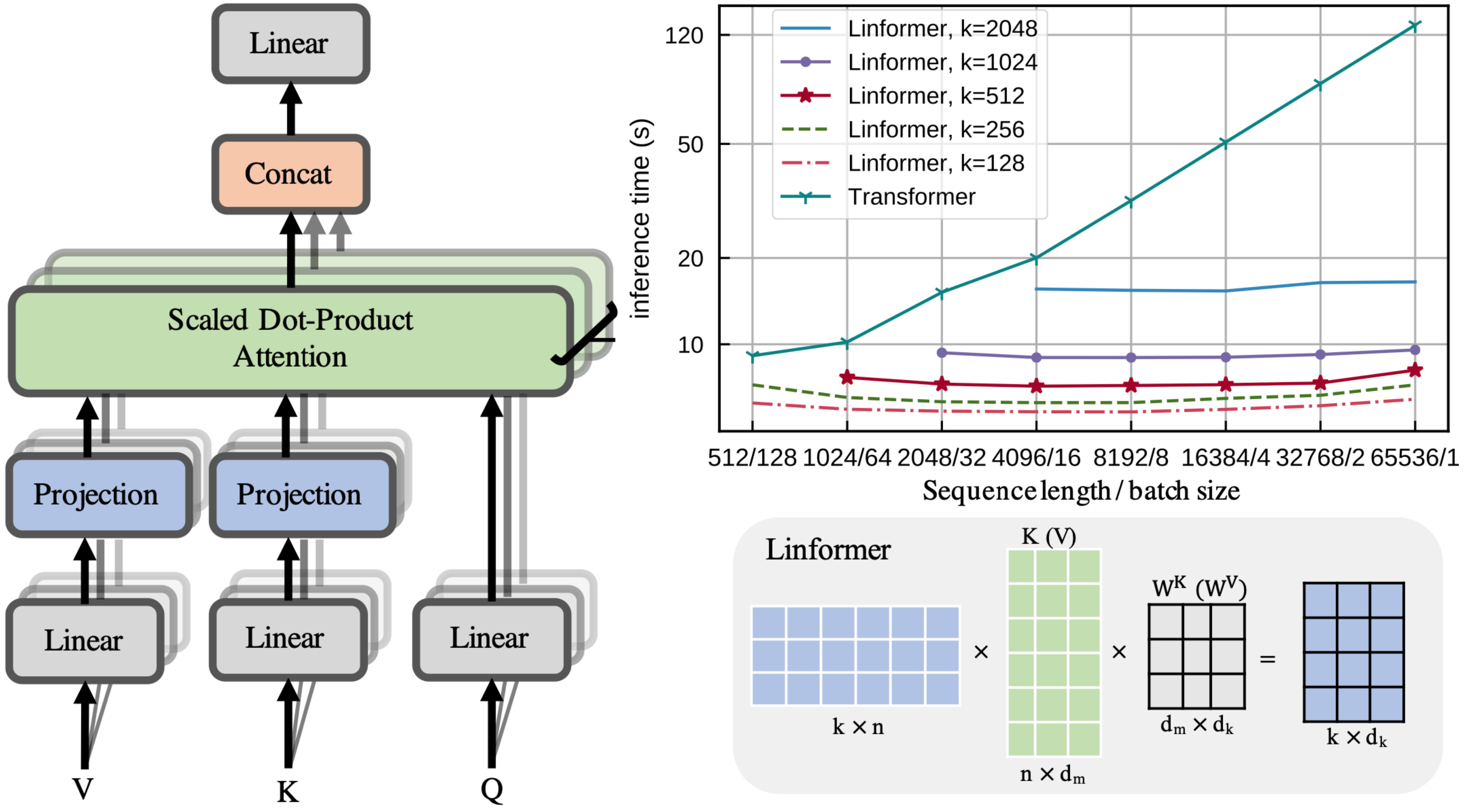


Figure 2: Left and bottom-right show architecture and example of our proposed multihead linear self-attention. Top right shows inference time vs. sequence length for various Linformer models.

Figure from:
https://arxiv.org/pdf/2006.04768.pdf

Hugging face transformers

- Hugging Face is the state-of-the-art Natural Language Processing tool for Pytorch and TensorFlow 2.0.

- Provides lots of architectures we mentioned in the previous slides such as: BERT, RoBERTa, DistilBert, XLNet, etc.

- Hugging Face provides lots of official notebooks for the beginner to start from: https://huggingface.co/transformers/notebooks.html

Bilingual Transformer

- Most of the transformers we mentioned in the previous slides focus on the performance instead of the different language. To make our ABBIE woks, we need not only the efficient transformer but also a transformer that can underhand bilingual or even multi-language in the same time.

- Cross-lingual language model pretraining:

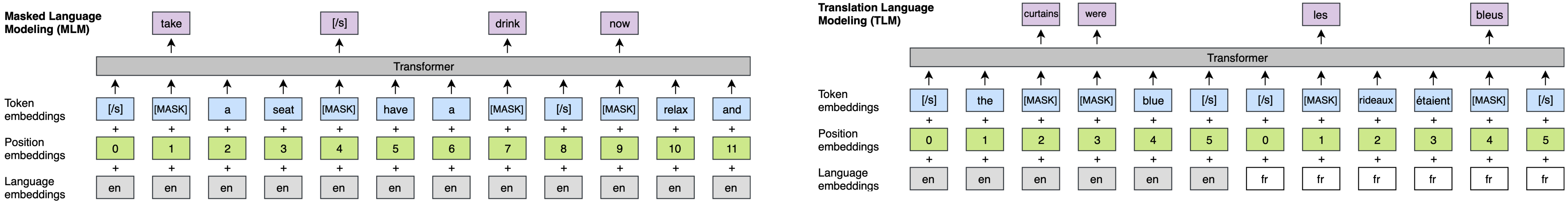

Masked Language Modeling (MLM)
take
[/s]
drink
now
Transformer
Token embeddings
[/s] [MASK] a seat [MASK] have a [MASK] [/s] [MASK] relax and
Position embeddings
0 1 2 3 4 5 6 7 8 9 10 11
Language embeddings
en en en en en en en en en en en en
Translation Language Modeling (TLM)
curtains
were
les
bleus
Transformer
Token embeddings
[/s] the [MASK] [MASK] blue [/s] [/s] [MASK] rideaux étaient [MASK] [/s]
Position embeddings
0 1 2 3 4 5 0 1 2 3 4 5
Language embeddings
en en en en en en fr fr fr fr fr fr

Figure 1: **Cross-lingual language model pretraining.** The MLM objective is similar to the one of Devlin et al. (2018), but with continuous streams of text as opposed to sentence pairs. The TLM objective extends MLM to pairs of parallel sentences. To predict a masked English word, the model can attend to both the English sentence and its French translation, and is encouraged to align English and French representations. Position embeddings of the target sentence are reset to facilitate the alignment.

Figure from:
https://arxiv.org/pdf/1901.07291.pdf

Bilingual Transformer

- Bilingual Transformer

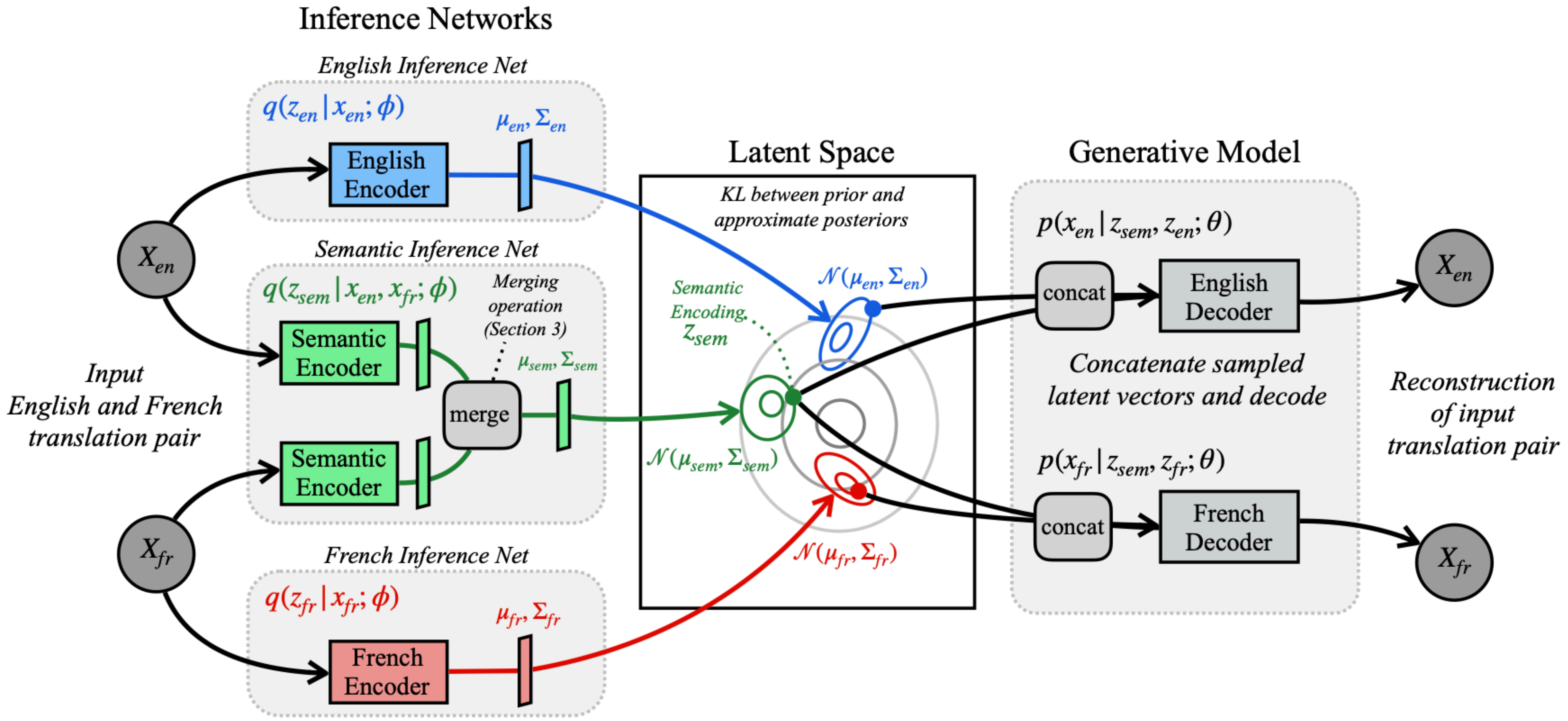

Inference Networks
English Inference Net
$q(z_{en}|x_{en};\phi)$
μ_{en}, Σ_{en}
English Encoder
Semantic Inference Net
$q(z_{sem}|x_{en}, x_{fr};\phi)$
Merging operation (Section 3)
μ_{sem}, Σ_{sem}
Semantic Encoder
merge
Semantic Encoder
French Inference Net
$q(z_{fr}|x_{fr};\phi)$
μ_{fr}, Σ_{fr}
French Encoder
X_{en}
Input English and French translation pair
X_{fr}
Latent Space
KL between prior and approximate posteriors
Semantic Encoding, z_{sem}
$\mathcal{N}(\mu_{en}, \Sigma_{en})$
$\mathcal{N}(\mu_{sem}, \Sigma_{sem})$
$\mathcal{N}(\mu_{fr}, \Sigma_{fr})$
Generative Model
$p(x_{en}|z_{sem}, z_{en};\theta)$
concat
English Decoder
Concatenate sampled latent vectors and decode
$p(x_{fr}|z_{sem}, z_{fr};\theta)$
concat
French Decoder
X_{en}
Reconstruction of input translation pair
X_{fr}

Figure 2: The computation graph for the variational lower bound used during training. The English and French text are fed into their respective inference networks and the semantic inference network to ultimately produce the language variables z_{fr} and z_{en} and semantic variable z_{sem}. Each language-specific variable is then concatenated to z_{sem} and used by the decoder to reconstruct the input sentence pair.

Figure from:
https://arxiv.org/pdf/1911.03895.pdf

Image Captioning through Image Transformer

- Image caption vs image emoji caption

TABLE IV
Top 5 Emojis Predicted in the context of an image using different emoji knowledge concepts

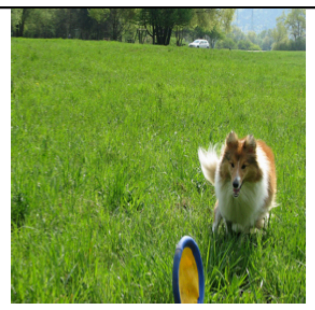
S.No	Image	Text Description	Using Processed sense definition	Using emoji senses	Using emoji names
1		A person looks down at something while sitting on a bike	🚵‍♀️ 🚵 🚴‍♀️ 🚲 🚴	🚴 🚴 🚴 🚴 🚴	🚲 🛵 🏍️ 🌀 🚴
2		The dog is playing with his toy in the grass	🐨 🐶 🐕 🐩 🐆	🐶 🐩 🐕 🐕 🐺	🐕 💠 🐶 🌭 😺
3		A tennis player in action on the court	🎾 🏌️‍♀️ 🏀 🏌️ 🏓	🎾 🏸 ⛹️‍♀️ ⛹️‍♀️ ⛹️‍♀️	🎾 ⛹️‍♀️ 🏸 ⛹️ ⛹️
4		Cup of coffee with dessert items on a wooden grained table	🍮 🍔 🍥 🥗 🍟	🍿 🍭 🍬 🍵 ☕	📤 📥 🥘 🇬🇼 🖍️

Figure from:
https://arxiv.org/pdf/1808.08891.pdf

ABBIE

- ABBIE: a **A**ttention-**B**ased **B**ilingual **I**mage2caption **E**moji Model

- Different transformer help us to pre-train the word and emoji embedding

- Bilingual language model allow ABBIE to understand the same semantic meaning in the different language

- Image caption help ABBIE to transfer the image pixels into word sequence that semantically related to the given image

ABBIE

- Input: The image you are ready to upload on social network
- Output: The bilingual caption with the corresponding emoji
- Method:
  - Choose one of the transformer architecture in the previous slides to pretrain the word and emoji embedding
  - Fine-tuned the transformer to the image-2-text task
  - ABBIE DONE!!!

References

- Cross-lingual Language Model Pretraining
  - https://arxiv.org/pdf/1901.07291.pdf
- A Bilingual Generative Transformer for Semantic Sentence Embedding
  - https://arxiv.org/pdf/1911.03895.pdf
- Which Emoji Talks Best for My Picture?
  - https://arxiv.org/pdf/1808.08891.pdf
- Image Captioning through Image Transformer
  - https://arxiv.org/pdf/2004.14231.pdf