Deep NLP

Applying Machine Learning To Text

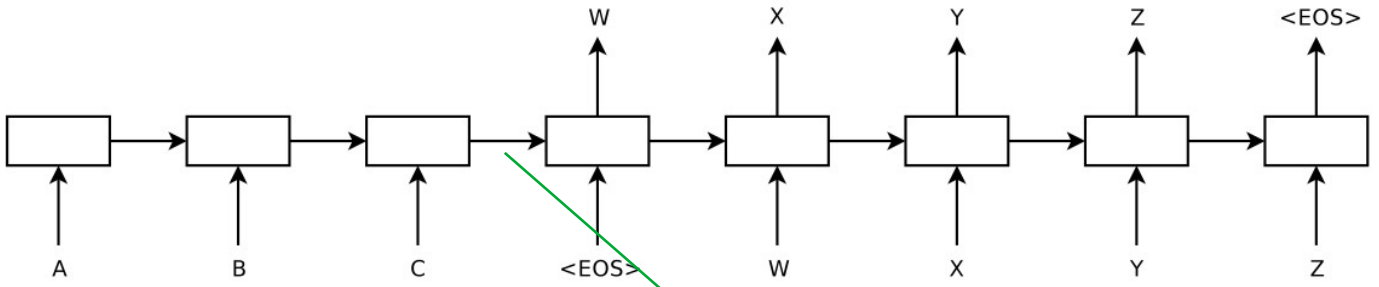
- For some tasks, we can get amazingly far with "bag-of-words" representations...
  - Document classification/clustering
  - Spam detection
- For some tasks, word order is crucial...

Some Notable ML related NLP Results

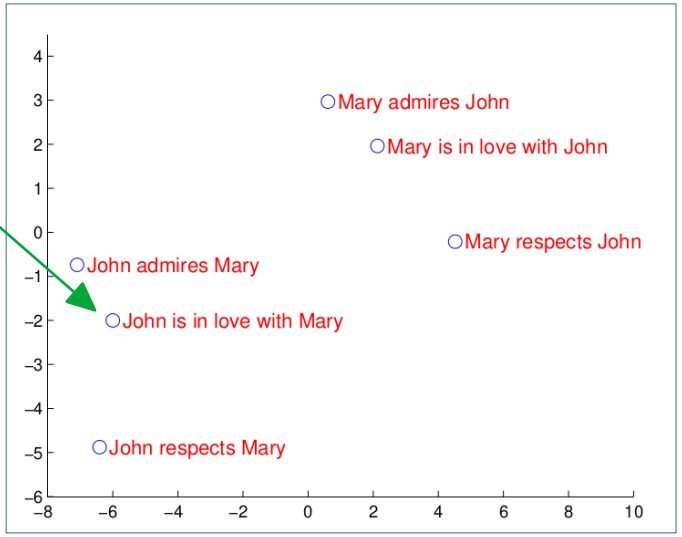
- 2016 – Google Translate 60% Reduction in translation errors
- Oct. 25th 2019 BERT used in Google search https://www.blog.google/products/search/search-language-understanding-bert/
- Nov, 2022 – ChatGPT released by OpenAI **https://openai.com/blog/chatgpt**

Translation: Encoder/Decoder Models

- Original encoder/decoder architecture:



- Both encoder and decoder are multi-layer LSTMs

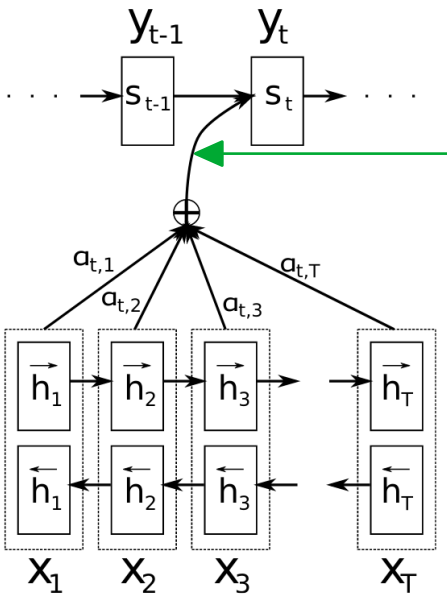


I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to Sequence Learning with Neural Networks,” in Advances in Neural Information Processing Systems 27, 2014, pp. 3104–3112.

Attention

Decoder

Encoder (Bi-directional LSTM)



$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j$$

Decoder receives weighted sum of *all* annotations

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})}$$

Weights are calculated using "softmax" over the output of an alignment model

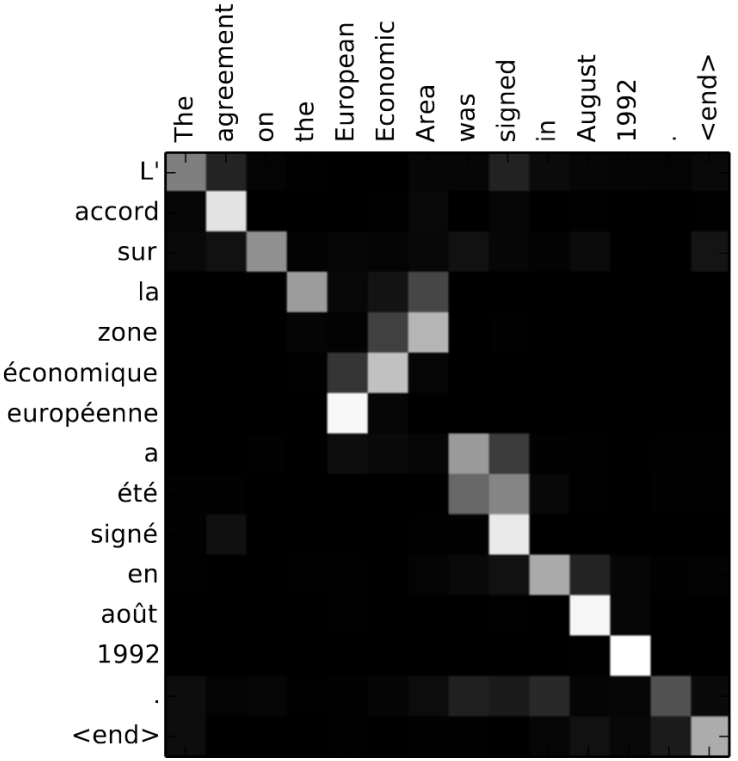
$$e_{ij} = a(s_{i-1}, h_j)$$

Alignment model is a simple feedforward network

Bahdanau, D., Cho, K. H., & Bengio, Y. (2015, January). Neural machine translation by jointly learning to align and translate. In 3rd International Conference on Learning Representations, ICLR 2015.

Attention: Alignment Example

- English to French translation
- Each “pixel” shows the corresponding α_{ij}



Bahdanau, D., Cho, K. H., & Bengio, Y. (2015, January). Neural machine translation by jointly learning to align and translate. In 3rd International Conference on Learning Representations, ICLR 2015.

Challenge For RNN's

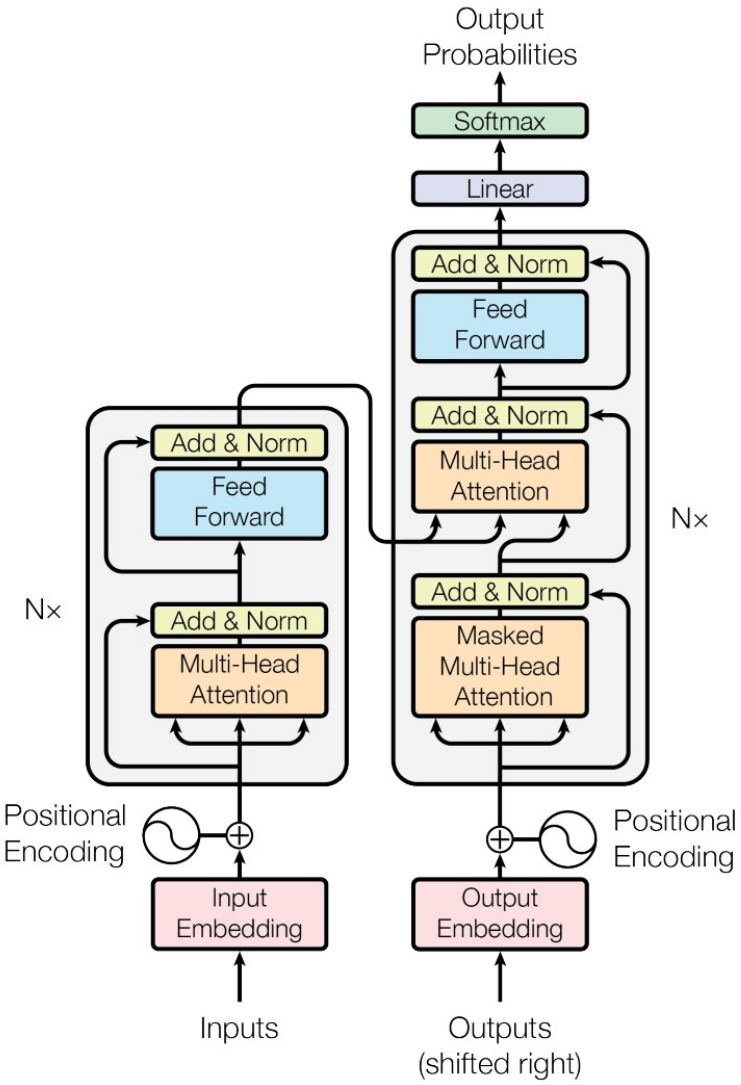
- Impossible to parallelize!
- One alternative is CNN's
- Another is transformer networks...

Transformer Networks

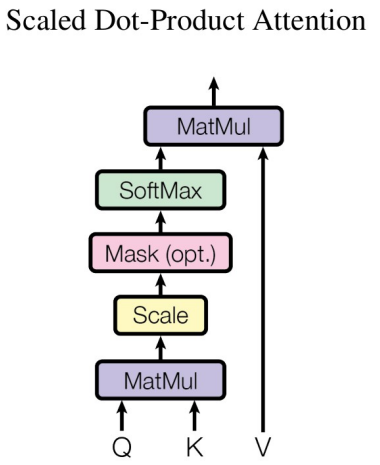
- Introduced in:

A. Vaswani et al., “Attention is All you Need,” in Advances in Neural Information Processing Systems 30, 2017.

- Nice visualizations:
  - http://jalammar.github.io/illustrated-transformer/



Transformer Attention



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

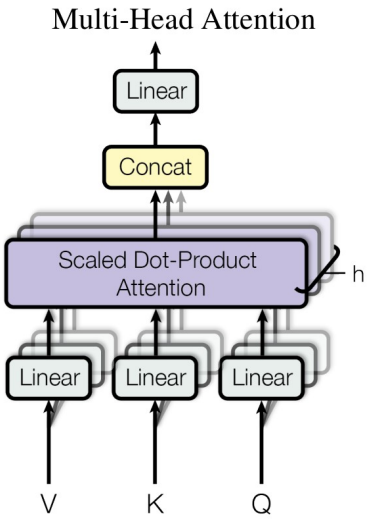
Think of Q, K and V as projected versions of the activations from the previous layer: if Q_i "matches" Kj, then V_j is selected as the output.

A. Vaswani et al., "Attention is All you Need," in Advances in Neural Information Processing Systems 30, 2017.

Transformer: Multi-Head Attention

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_\text{h})W^O$$
$$\text{where head}_\text{i} = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

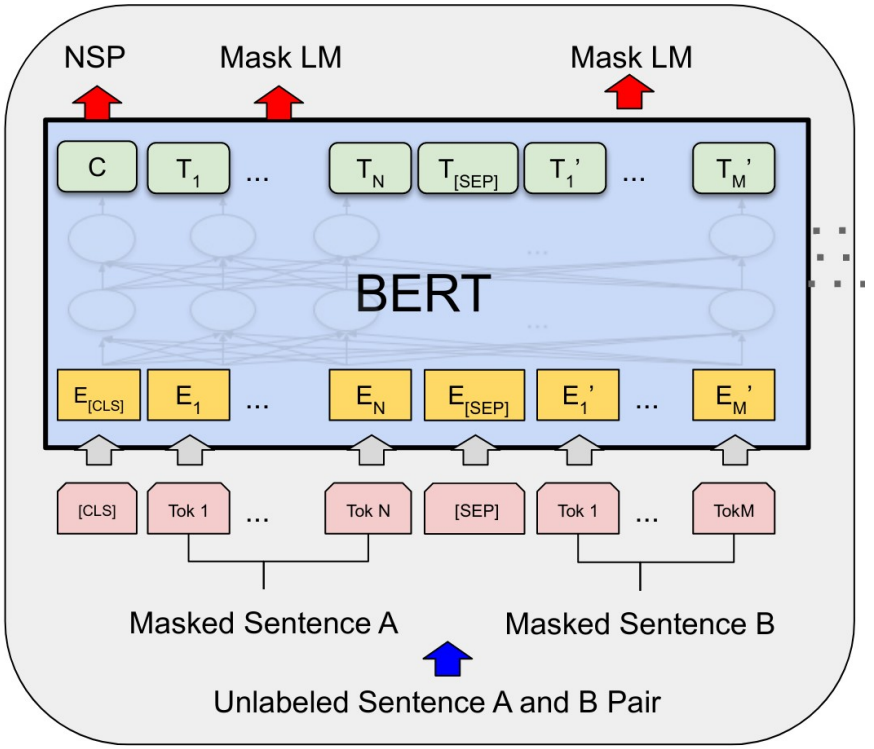
Where the projections are parameter matrices $W_i^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_v}$ and $W^O \in \mathbb{R}^{hd_v \times d_{\text{model}}}$.



A. Vaswani et al., “Attention is All you Need,” in Advances in Neural Information Processing Systems 30, 2017.

BERT

- Pre-train a transformer on unsupervised language tasks:
  - Predicting masked words
  - Next sentence prediction
- Fine tune on the supervised task of interest



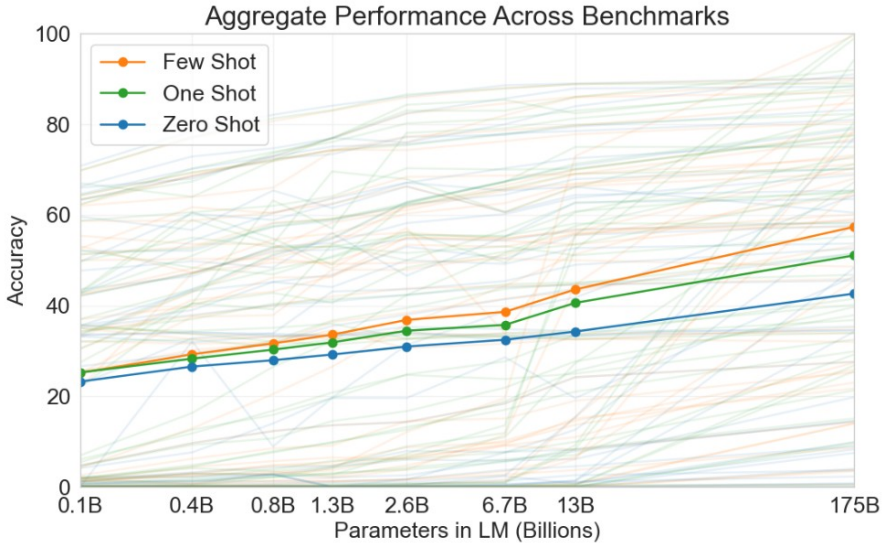
J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” 2019

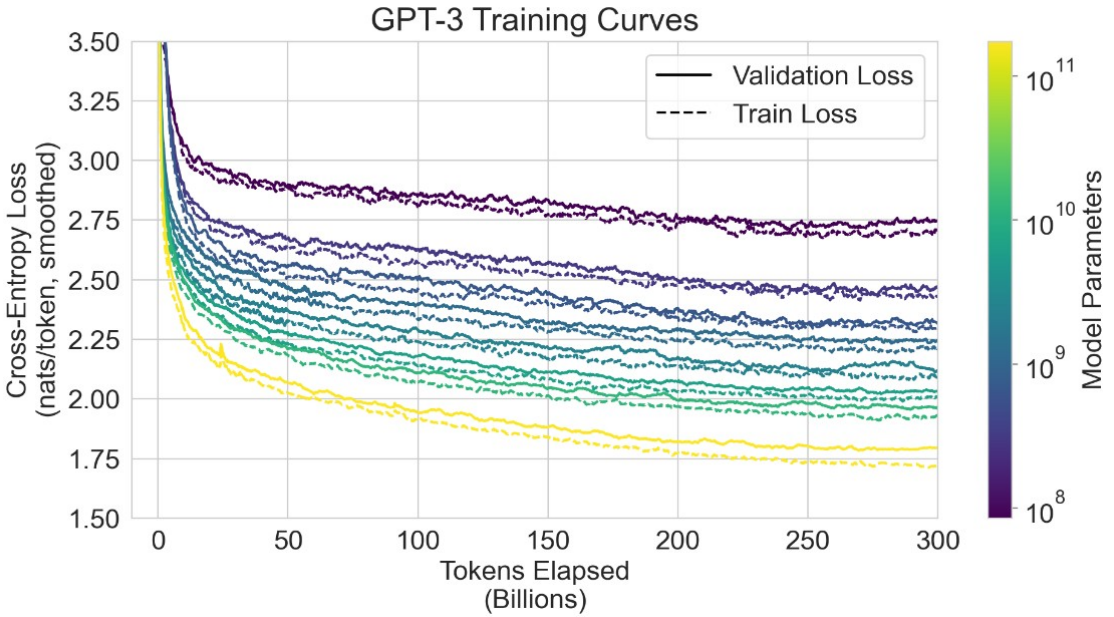
GPT-3

- GPT-3 architecture is similar to BERT, except it is *not* bi-directional: only predicts future symbols.
- Conclusion: If the model is big enough, fine-tuning is less important, or not needed.
  - They describe this as “one shot” or “few shot” learning. No weight updates, “learning” from the text of the query.
- Bigger is better...

Brown, Tom, et al. "Language models are few-shot learners." Advances in neural information processing systems 33 (2020)

GPT-3





Brown, Tom, et al. "Language models are few-shot learners." Advances in neural information processing systems 33 (2020)

GPT-4 (ChatGPT)

“This report focuses on the capabilities, limitations, and safety properties of GPT-4. GPT-4 is a Transformer-style model [39] pre-trained to predict the next token in a document, using both publicly available data (such as internet data) and data licensed from third-party providers. The model was then fine-tuned using Reinforcement Learning from Human Feedback (RLHF) [40]. Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar.”

GPT-4 Technical Report, OpenAI (2023)