

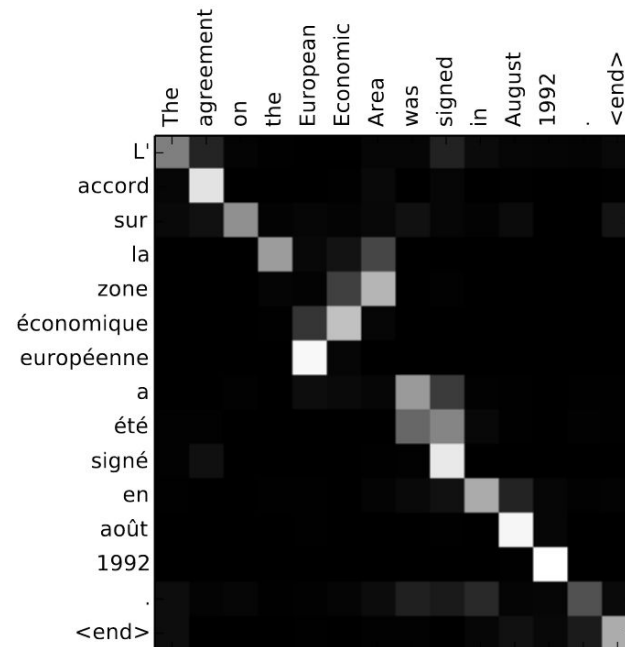
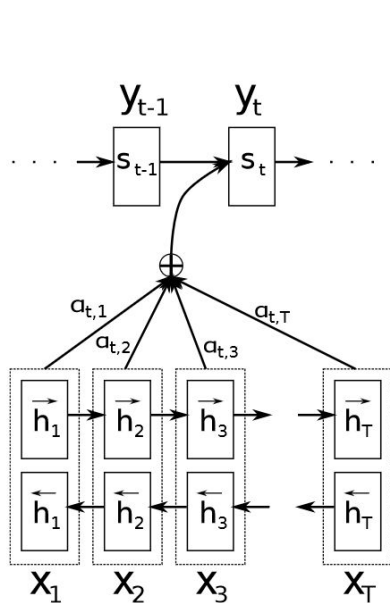
Advanced Language Models

Natural Language Processing

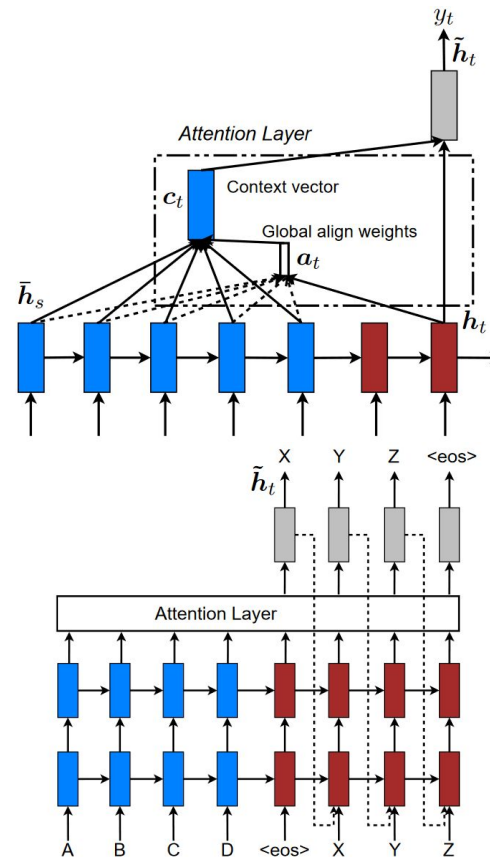
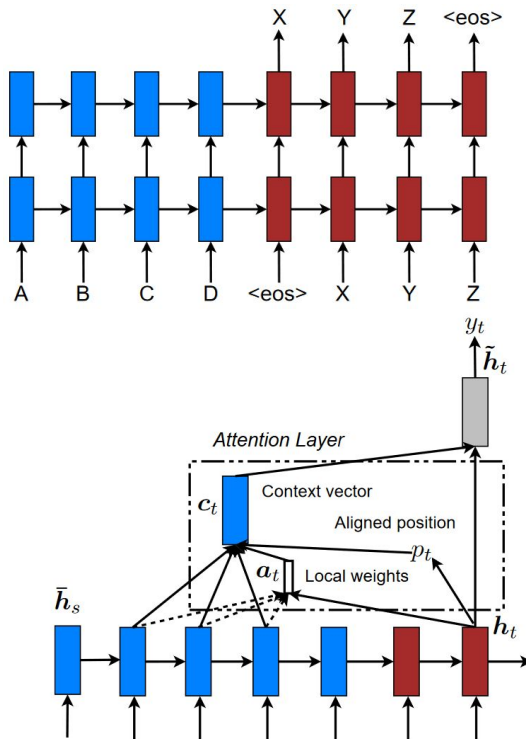
Objective

- Revisit attention mechanisms
- Explain transformers
- Explain BERT
- Explain XLNet

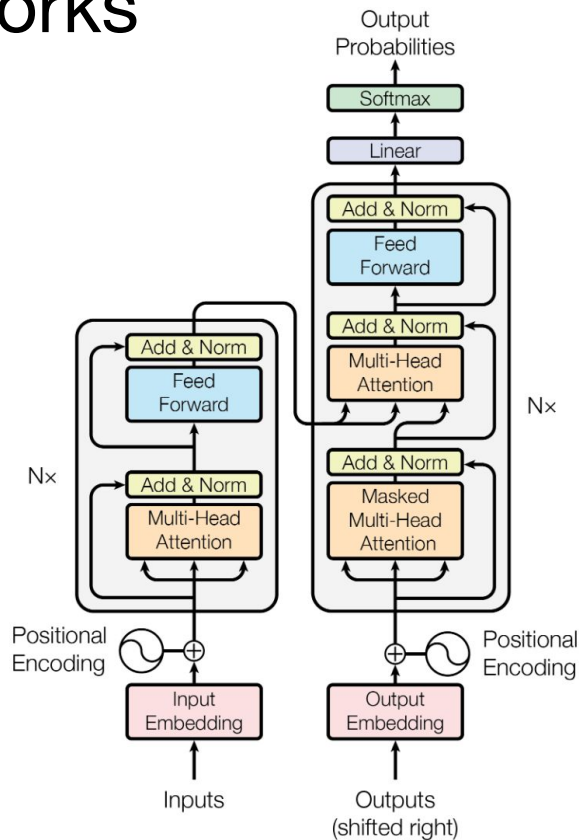
Attention between encoder and decoder



No vs. global vs. local attention

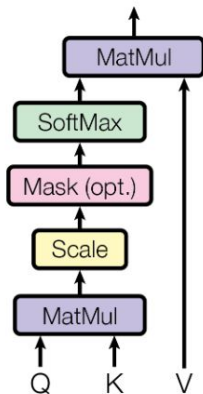


Transformer networks

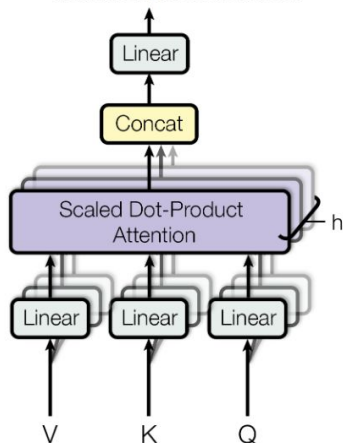


Attention in Transformer networks

Scaled Dot-Product Attention



Multi-Head Attention

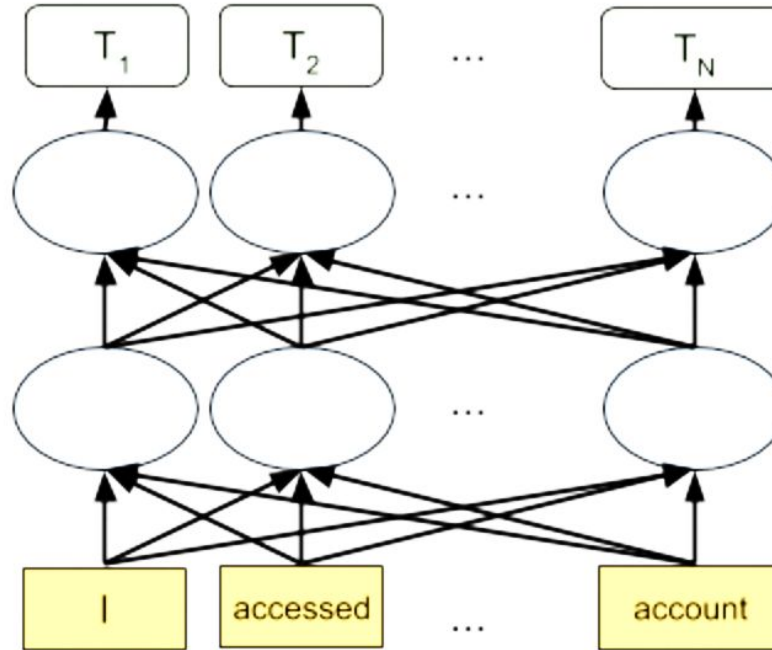


$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

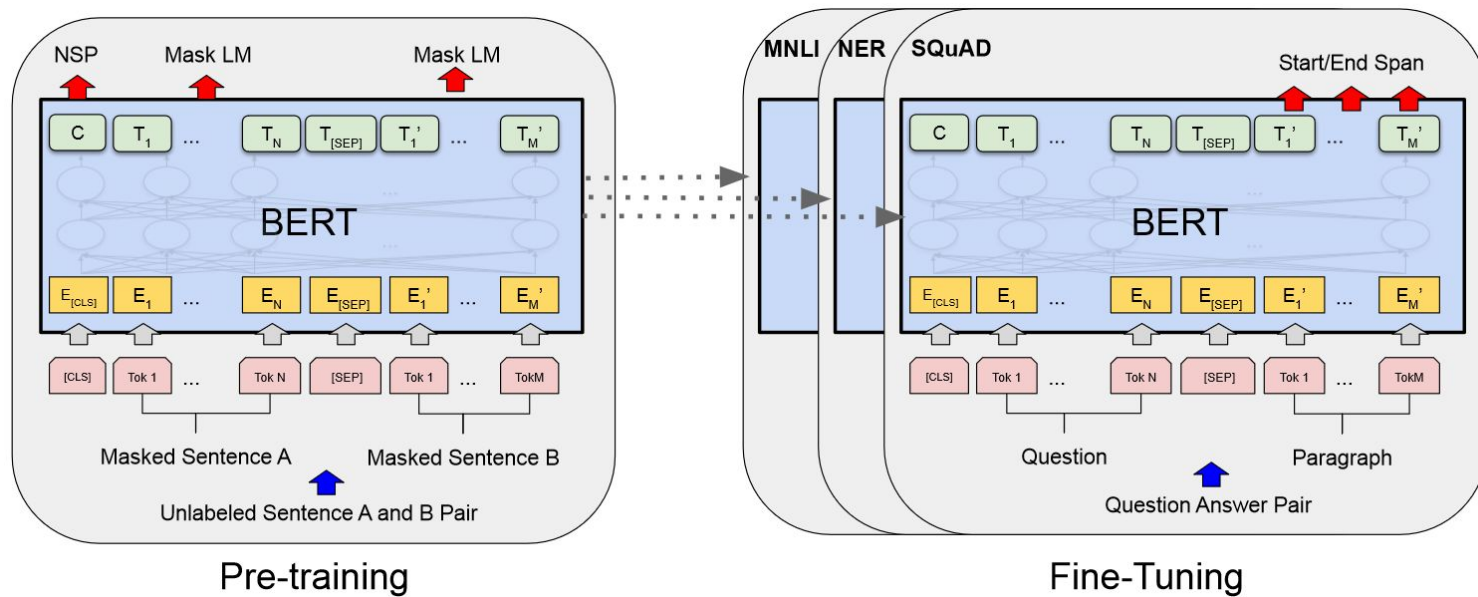
$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O$$

where $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$

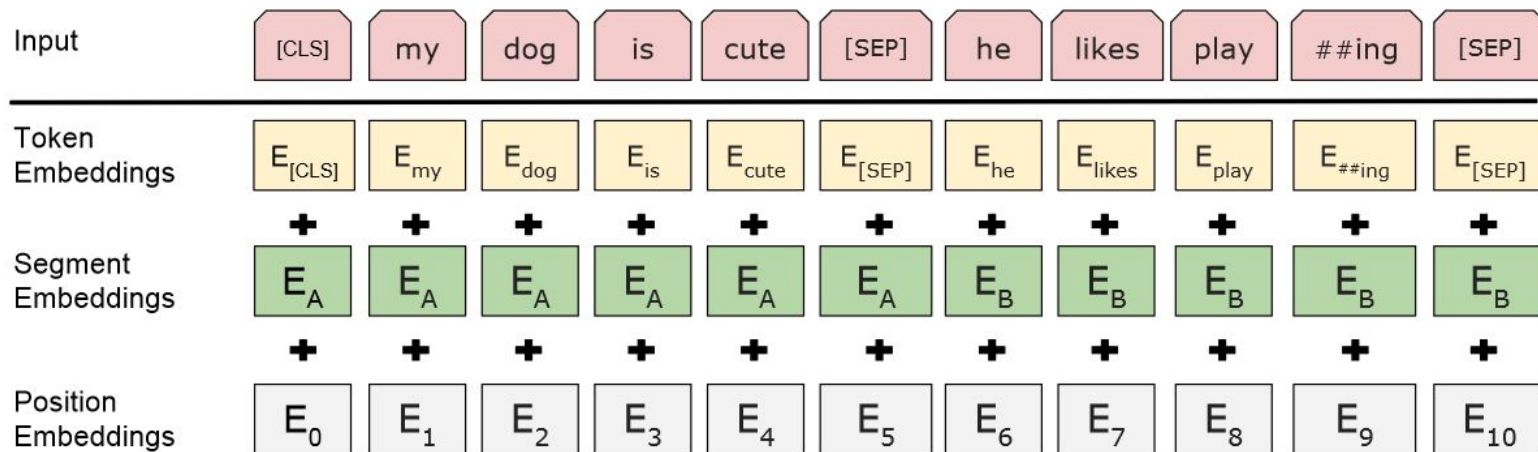
BERT



BERT



BERT



XLNet compared to previous models

LSTM and transformer's forward autoregressive factorization

$$\max_{\theta} \log p_{\theta}(\mathbf{x}) = \sum_{t=1}^T \log p_{\theta}(x_t \mid \mathbf{x}_{<t}) = \sum_{t=1}^T \log \frac{\exp(h_{\theta}(\mathbf{x}_{1:t-1})^{\top} e(x_t))}{\sum_{x'} \exp(h_{\theta}(\mathbf{x}_{1:t-1})^{\top} e(x'))}$$

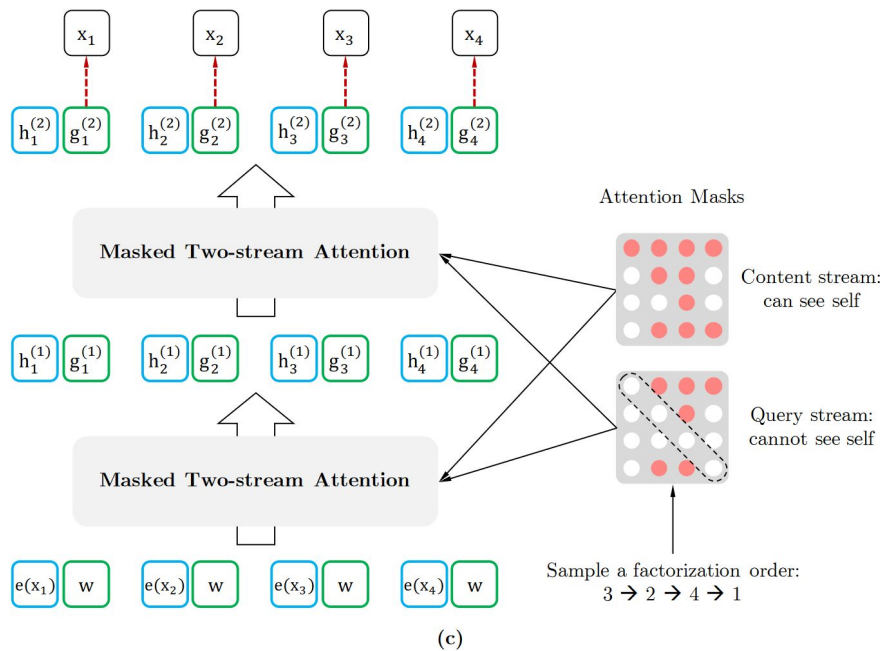
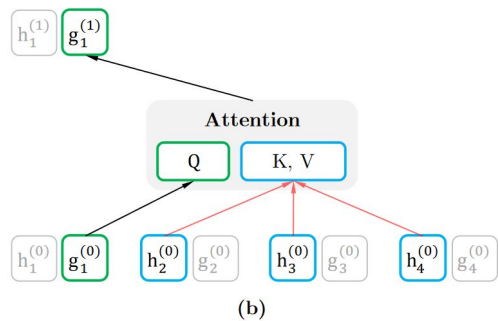
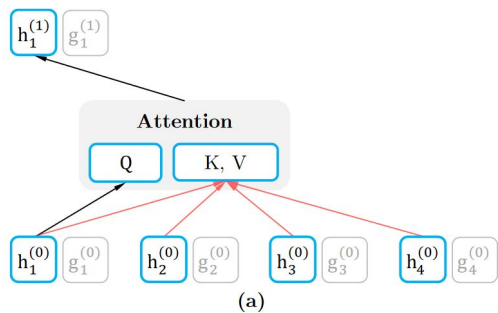
BERT's denoising auto-encoding

$$\max_{\theta} \log p_{\theta}(\bar{\mathbf{x}} \mid \hat{\mathbf{x}}) \approx \sum_{t=1}^T m_t \log p_{\theta}(x_t \mid \hat{\mathbf{x}}) = \sum_{t=1}^T m_t \log \frac{\exp(H_{\theta}(\hat{\mathbf{x}})_t^{\top} e(x_t))}{\sum_{x'} \exp(H_{\theta}(\hat{\mathbf{x}})_t^{\top} e(x'))}$$

XLNet's permutation language modeling

$$\max_{\theta} \mathbb{E}_{\mathbf{z} \sim \mathcal{Z}_T} \left[\sum_{t=1}^T \log p_{\theta}(x_{z_t} \mid \mathbf{x}_{\mathbf{z}_{<t}}) \right]$$

XLNet



Thank you!

Happy Learning :)