Attention and transformers

Attention and transformers

- Alternative to convolutions
  - Flexible in time
  - Popular in natural language processing

Attention

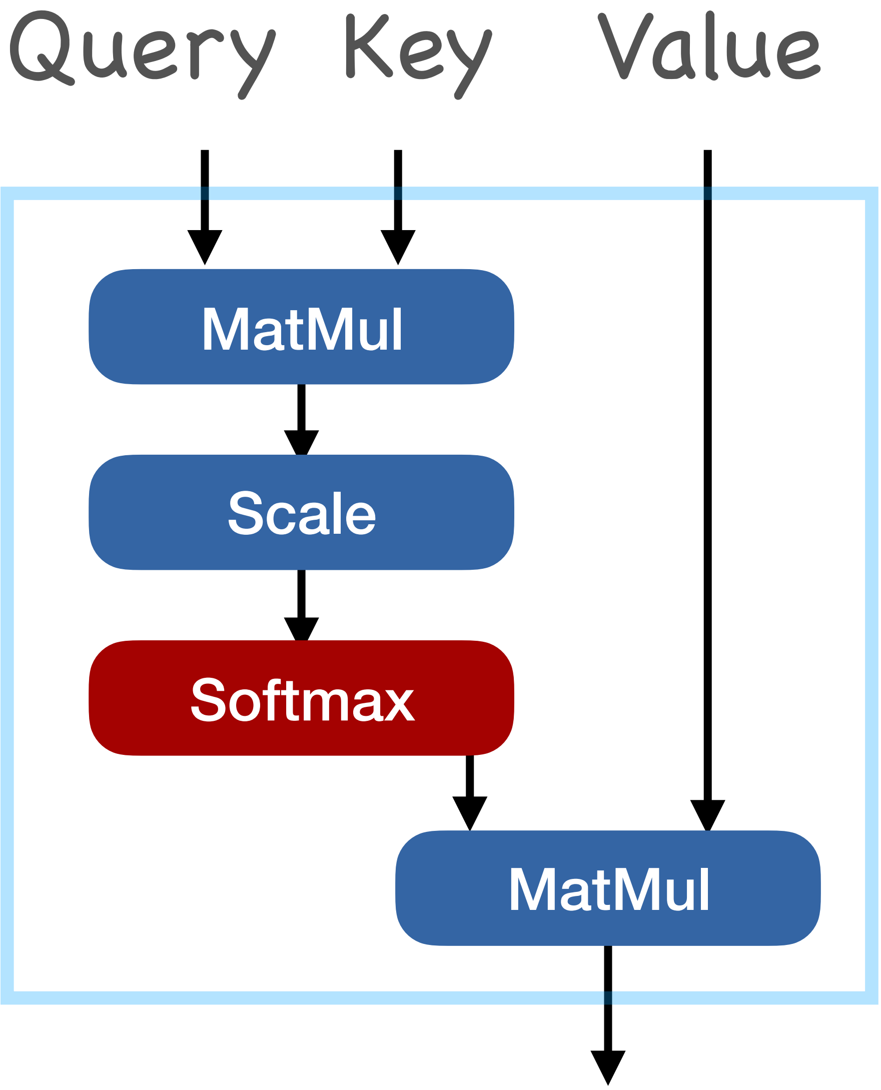
Attention Is All You Need, Vaswani et al., NIPS 2017

Attention

- Weighted average

$$\text{attention}\left(\mathbf{q}, \{\mathbf{k}_0, \mathbf{k}_1, \ldots\}, \{\mathbf{v}_0, \mathbf{v}_1, \ldots\}\right)$$

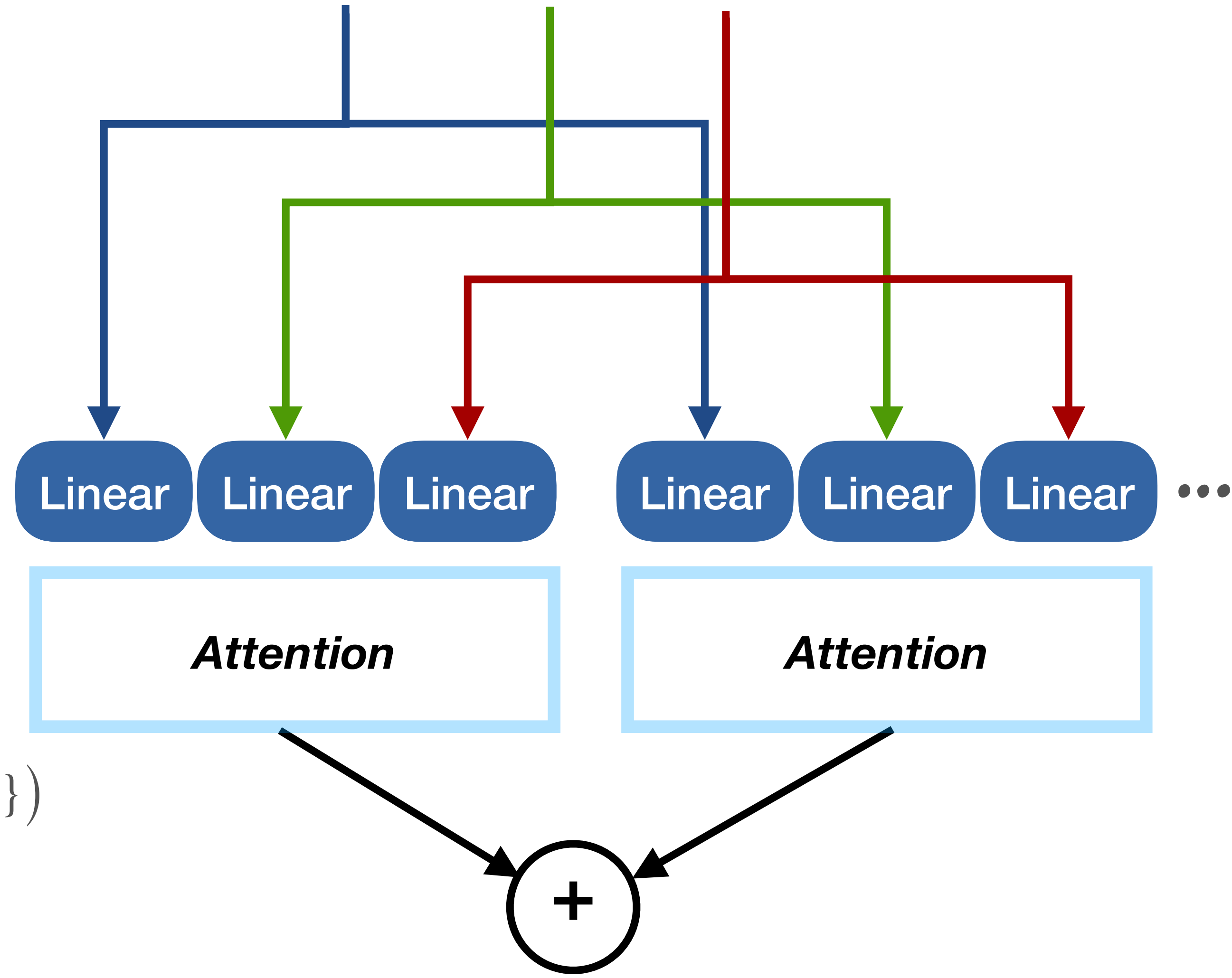
- $$= \frac{\sum_t \mathbf{v}_t \exp\left(\mathbf{k}_t^\top \mathbf{q}/\sqrt{d}\right)}{\sum_t \exp\left(\mathbf{k}_t^\top \mathbf{q}/\sqrt{d}\right)}$$



Multi-head attention

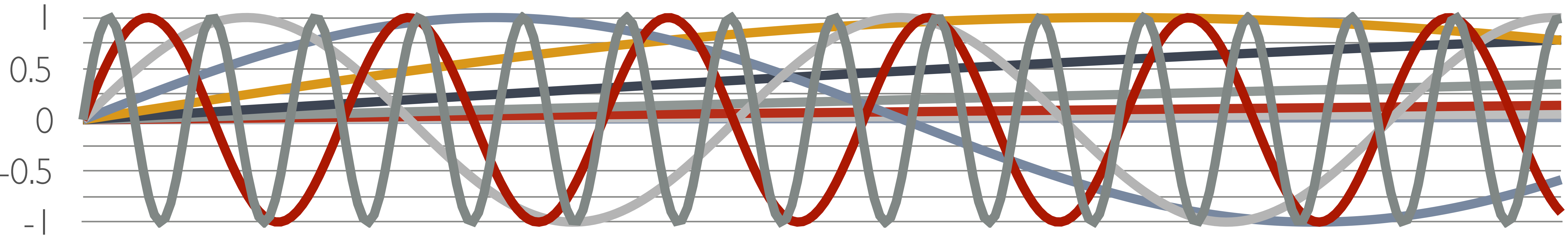
- Multiple attentions concatenated

- $$\begin{aligned}&\text{multihead}\left(\mathbf{q}, \{\mathbf{k}_0, \mathbf{k}_1, \ldots\}, \{\mathbf{v}_0, \mathbf{v}_1, \ldots\}\right)\\ &= \sum_i \text{attention}\left(\tilde{\mathbf{T}}_i\mathbf{q}, \{\mathbf{T}_i\mathbf{k}_0, \mathbf{T}_i\mathbf{k}_1, \ldots\}, \{\mathbf{W}_i\mathbf{v}_0, \mathbf{W}_i\mathbf{v}_1, \ldots\}\right)\end{aligned}$$



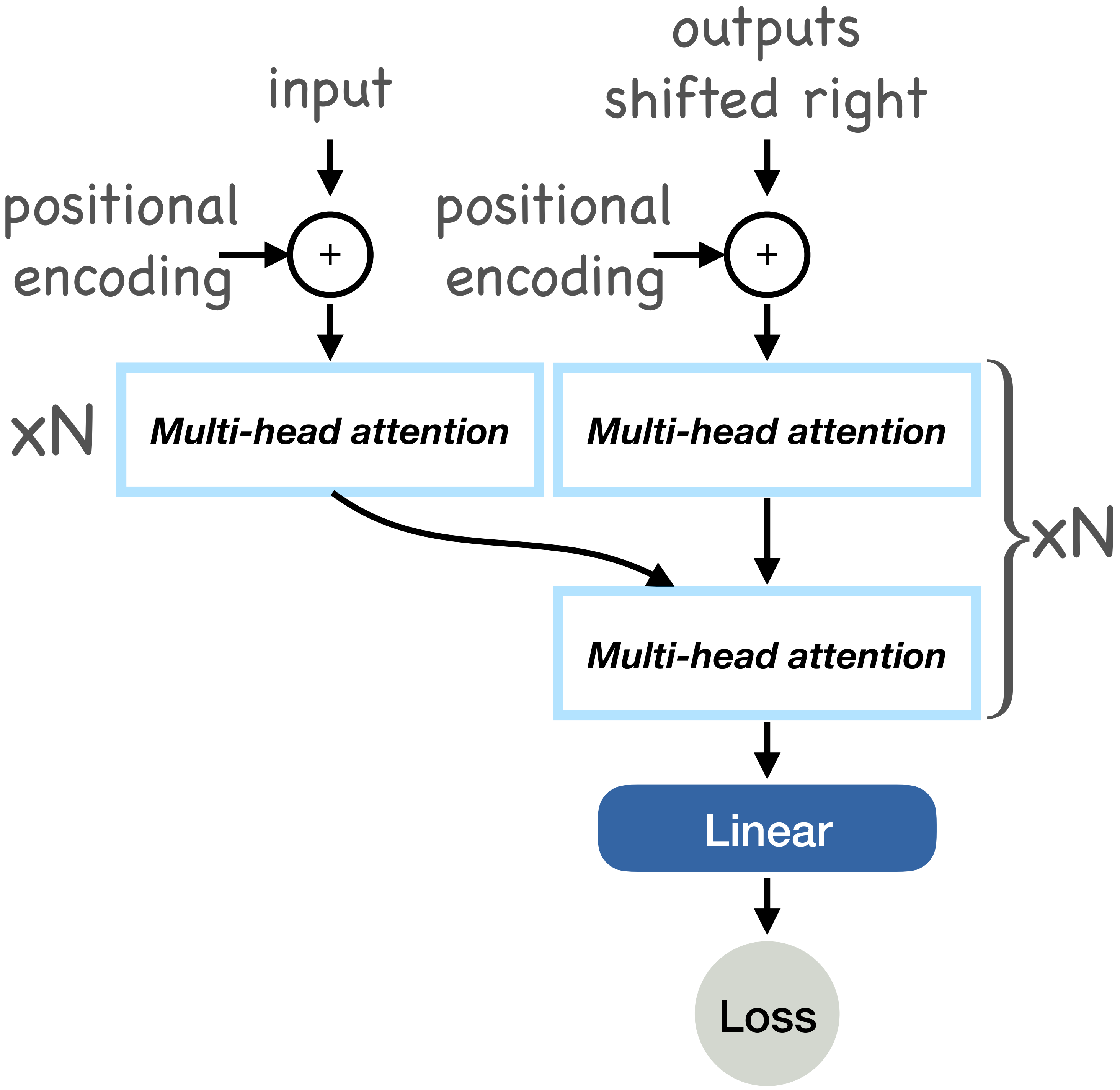
Positional encoding

- Attention is time-invariant
  - Add time back as a feature
    - sine and cosine of position



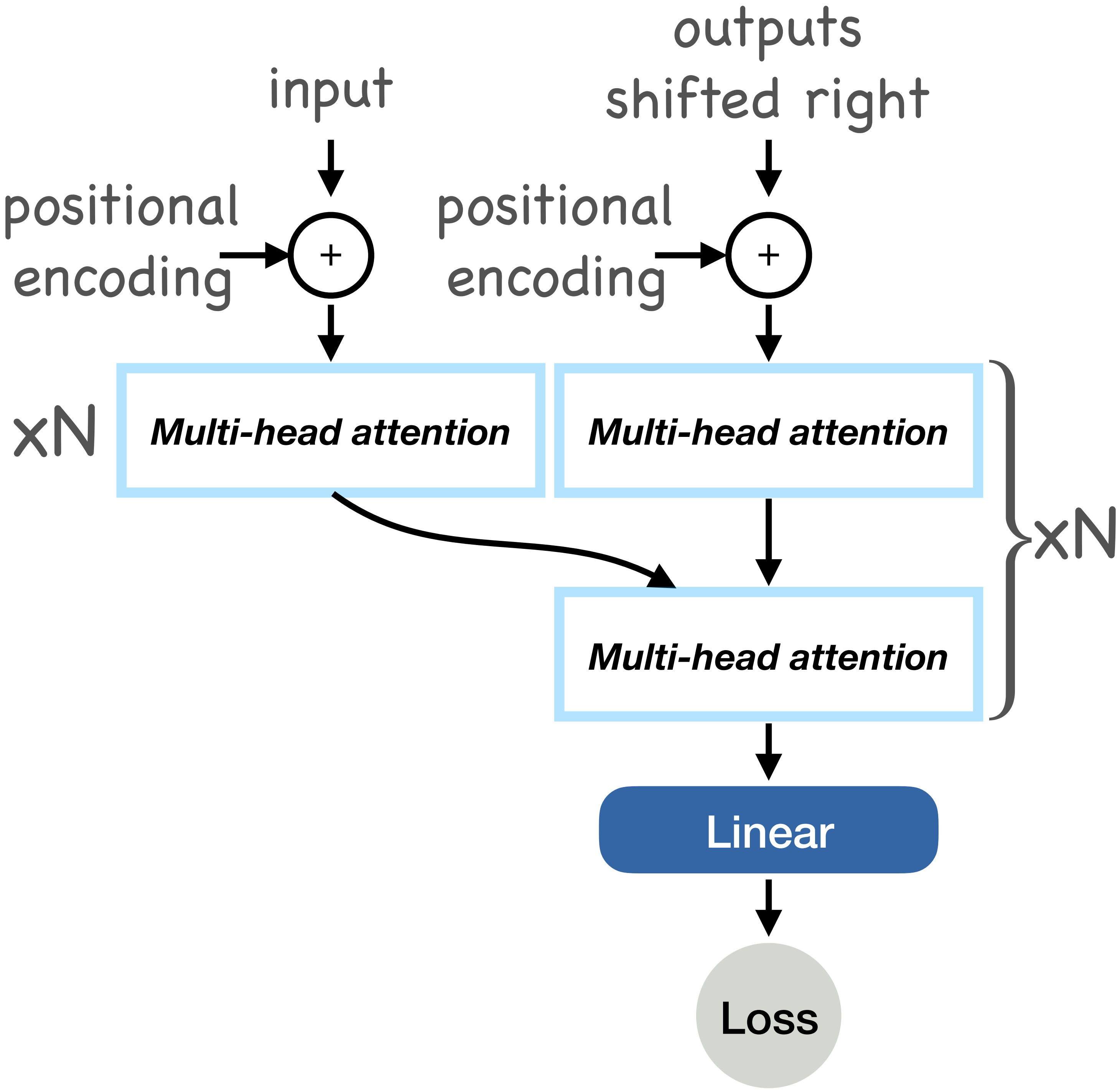
Transformer

- Feed forward
- Easy to train
  - Similar to Temporal CNN
- Causal attention
  - Auto-regressive



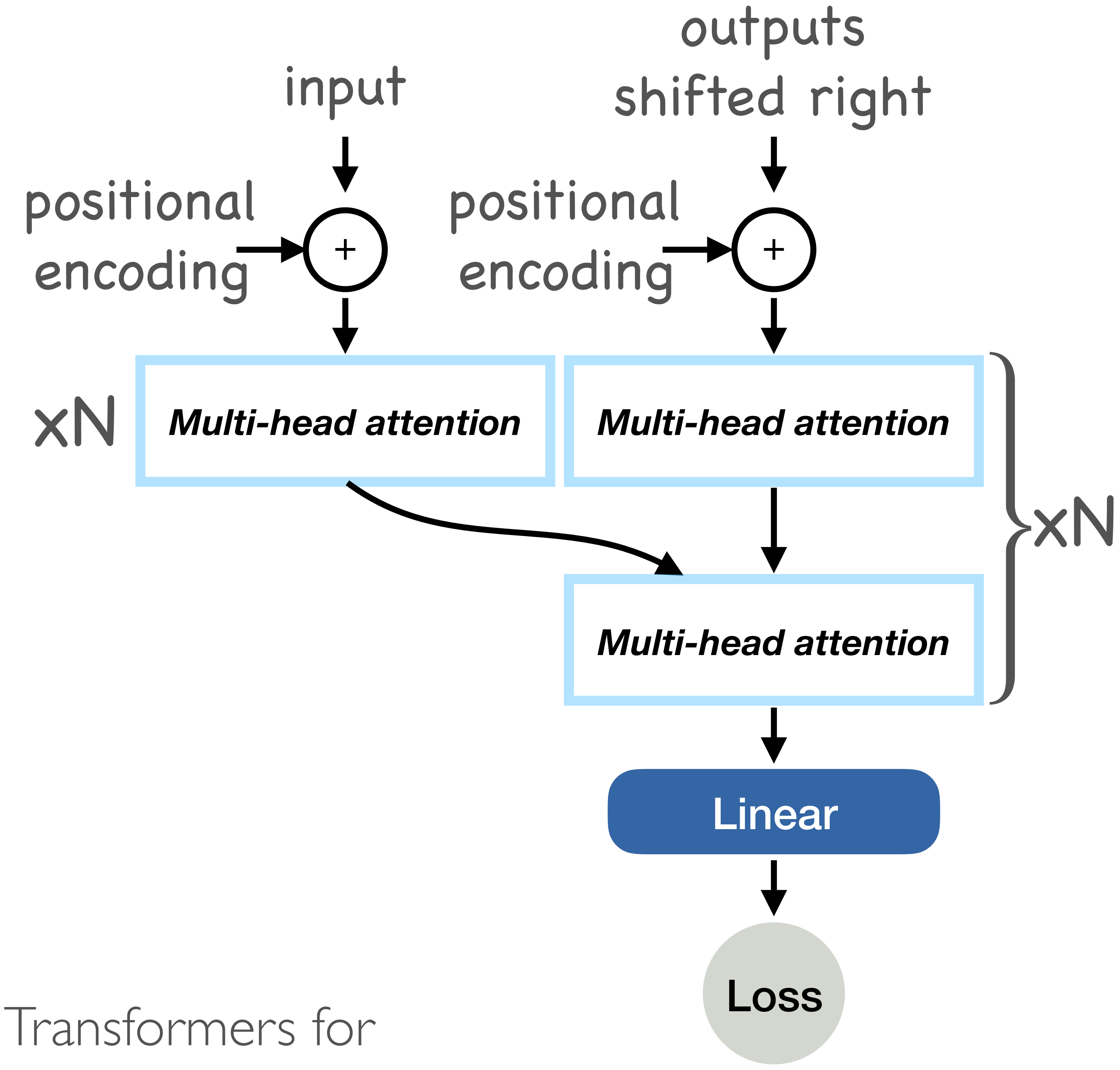
Transformer

- Faster to train
- Better performance
- State of the art performance



Bert

- Large transformer trained unsupervised
  - Predict masked out word
  - Predict next sentence
- Fine-tuned on NLP tasks
  - State-of-the-art for 6 month



BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Devlin et al., arXiv 2018