Problem: Attention is blind to position

Solution 1: (Shaw et al., 2018)* *simplified

- Encode **relative positions** via distance factors in the weights
- Weights: $A = \mathrm{Softmax}(X \cdot W^{A} \cdot X^{\top} + D)$ $D_{i,j} = d(i - j) \in R$

  $D \in R^{N \times N}$

Solution 2: (Vaswani et al., 2017)

- Encode **absolute positions** via positional embeddings
- Input: $X^{P} = X + E_{P} = x_{1} + e_{1}, \ldots, x_{n} + e_{n}$

- $E_{P} \in R^{N \times d}$ can be a decomposition of D $E_{P} \cdot E_{P}^{\top} \approx D$