# Attention based seq2seq models
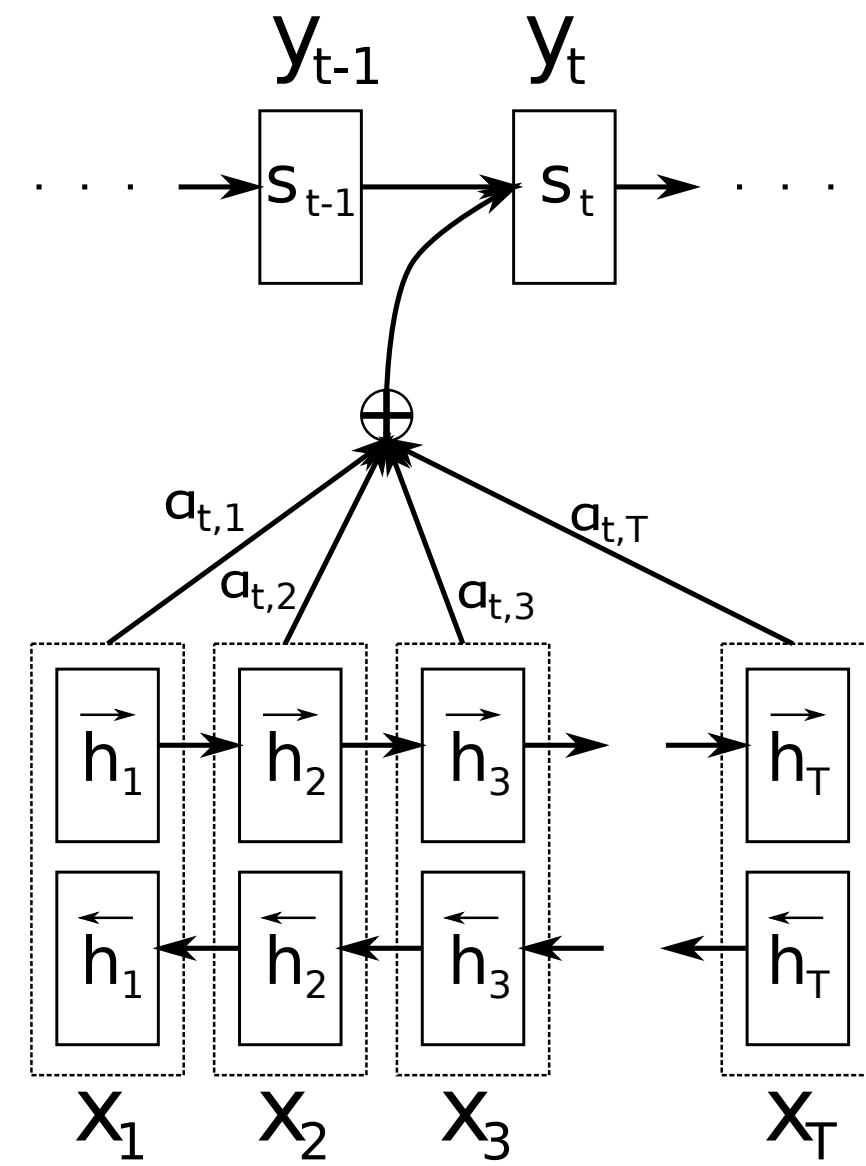


Bahdanau et al., *Neural Machine by Jointly Learning to Align and Translate,* 2015
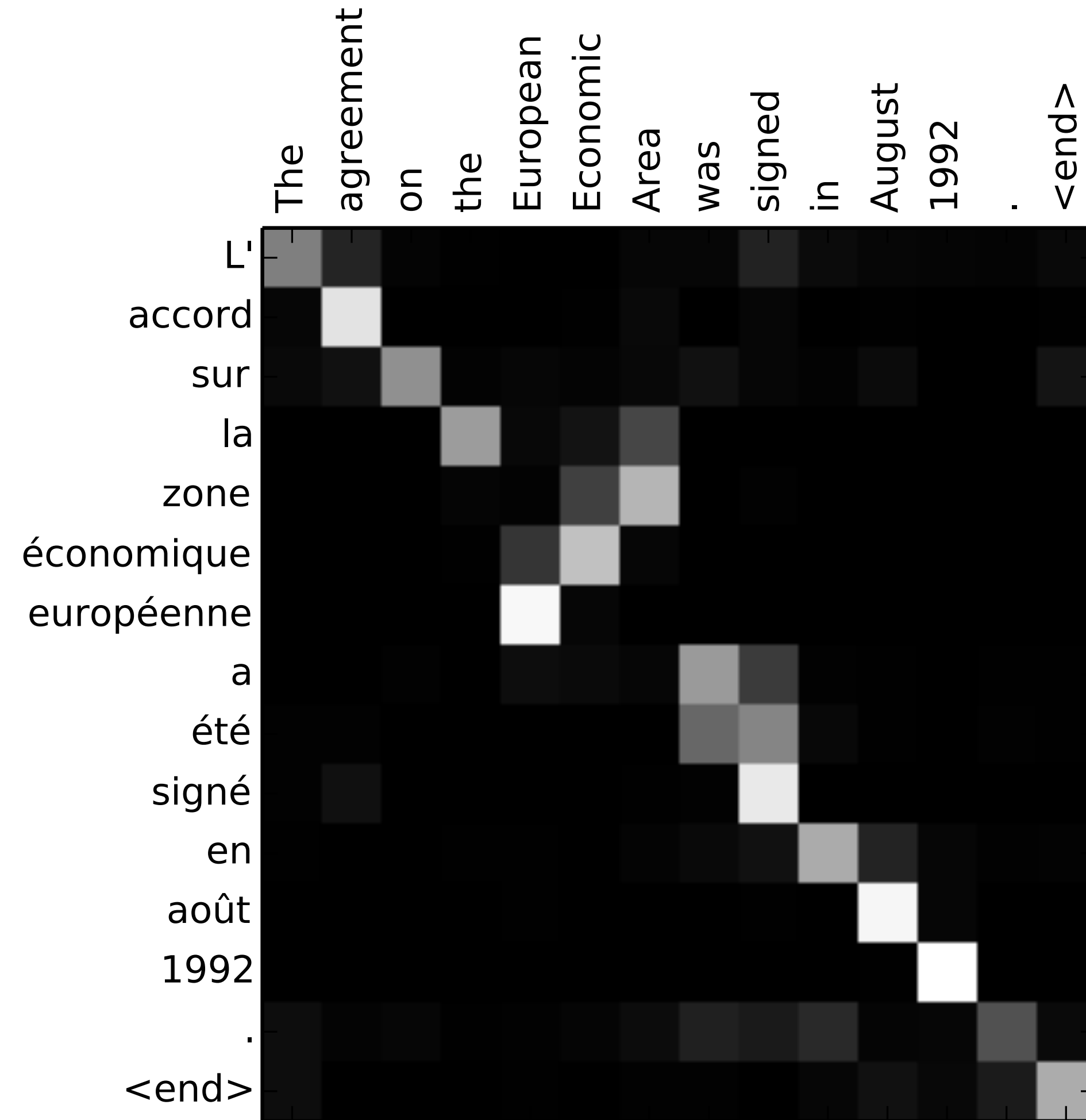
# Self Attention

Explicit representation of the pair-wise similarities between tokens in a sequence.

Similarity measures in include inner product, either between raw features or learnable transformations of them (i.e. with a small FFNN).

Interestingly, SA encodes the distributional hypothesis in the same way as word2vec.

In principle we can consider arbitrary amount of context, no "memory" required, all computation is parallelizable w.r.t. to sequence

But, attention of this form considers the input as a set, not a sequence ... this problem has been addressed with transformer models.
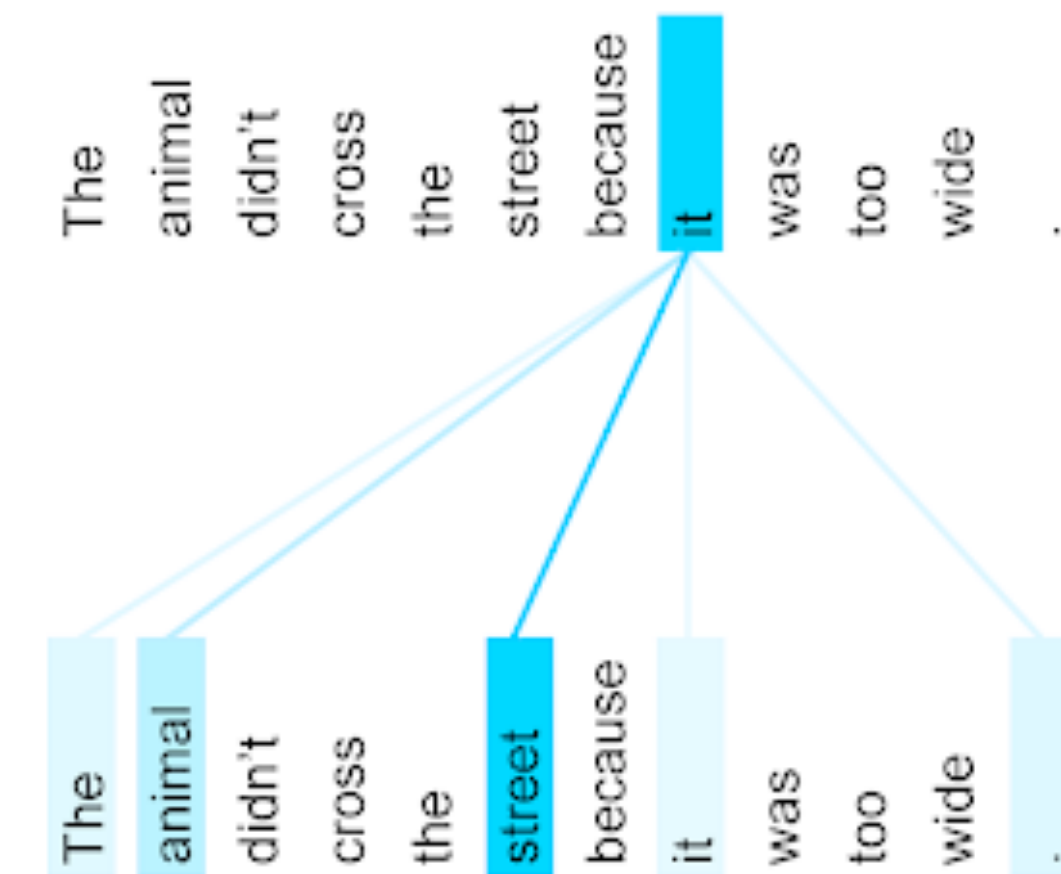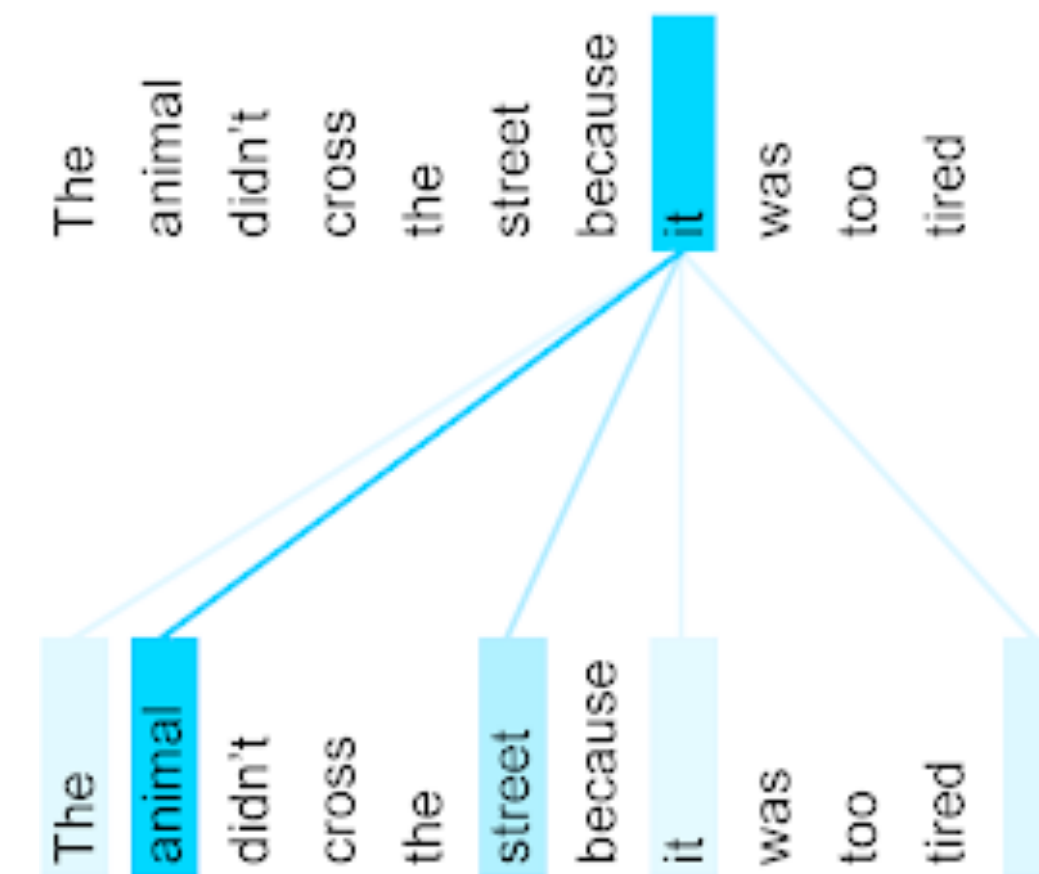


Taken from this Google AI blog post