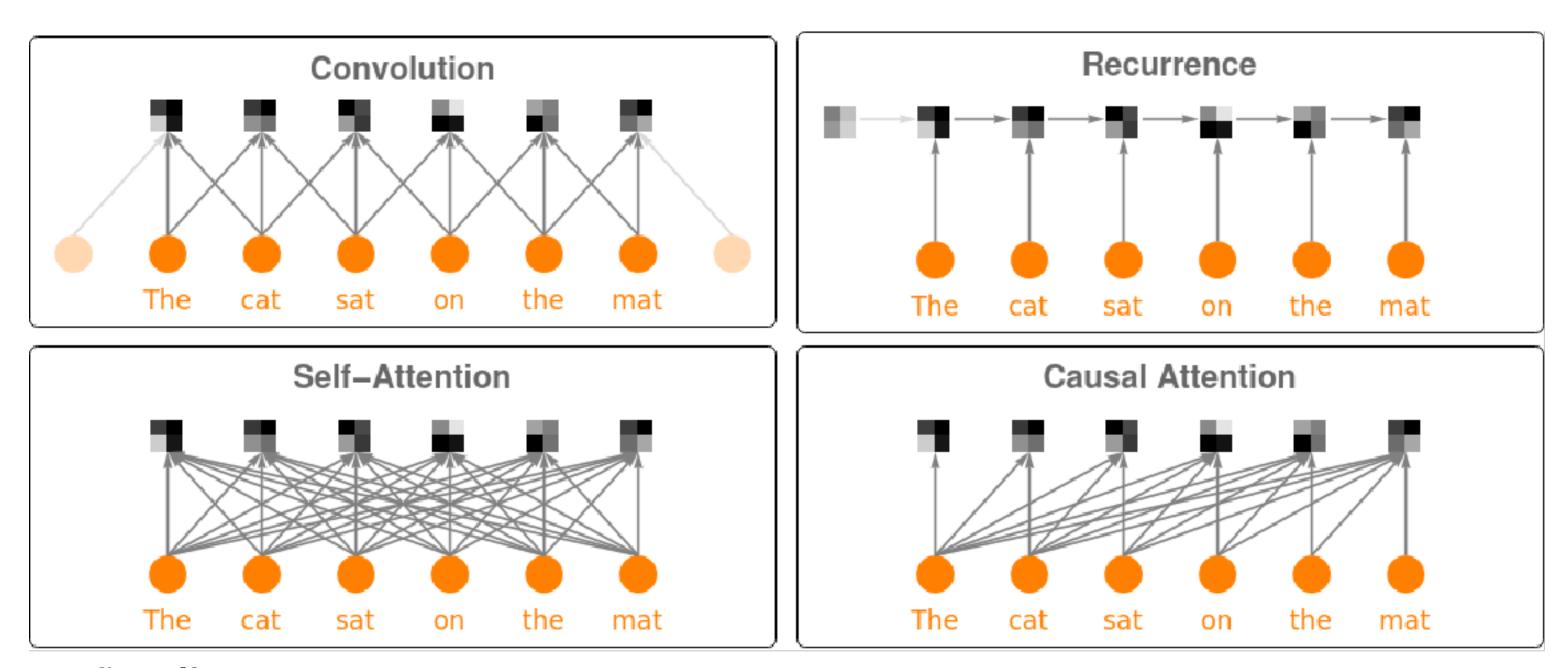
Neural language modeling: capturing context

- The three dominant classes of neural network architecture each differ in how context is captured:
 - Convolutional neural networks (CNNs)
 - Recurrent neural networks (RNNs)
 - Transformer networks (i.e., attention based)



Convolutional filtering

- Convolutional filtering is a general class of techniques used in signal processing for filtering continuous or discrete signals. It's based on the idea of a *finite impulse response filter* (FIR). In deep learning this was first applied to vision, wherein the convolutional filters are 2-dimensional patches (of varying sizes) that mimic the receptive fields in the human visual cortex.
- In NLP, convolutions are performed over the sequence dimension in embedding space, followed by a summation (or similar operation) to collapse the convolved features along the sequence dimension (this is called *pooling*), yielding a fixed size feature representation.
- Because of this pooling operation, convolutional filtering does not capture long range dependences well.

