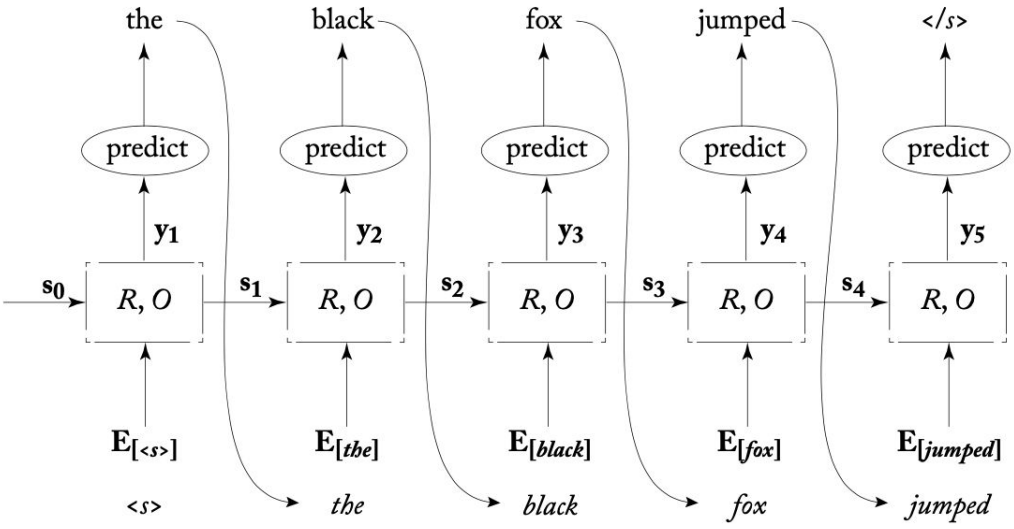
Simple sequences: RNN for generation



Similar to the case of generation from an *n*-gram language model, when generating from a trained RNN transducer one can either choose the highest probability item at each step, sample an item according to the model's predicted distribution, or use beam-search for finding a globally high-probability output.

Sequences generation

$$f : X \rightarrow Y$$

Sequence Sequence

x: x_1 x_2 x_3 ..... x_L (sequence)

y: y_1 y_2 y_3 ..... y_L (sequence)

http://speech.ee.ntu.edu.tw/~tlkagk/courses/MLDS_2015_2/Lecture/Sequence%20(v4).pdf

word	PTB tag
The	DT
German	JJ
Expressionist	NN
movement	NN
was	VBD
destroyed	VBN
as	IN
a	DT
result	NN
.	.

(8.20) a. *The U.S. Army captured Atlanta on May 14 , 1864*
B-ORG I-ORG I-ORG O B-LOC O B-DATE I-DATE I-DATE I-DATE

b. *Number of glucocorticoid receptors in lymphocytes and . . .*
O O B-PROTEIN I-PROTEIN O B-CELLTYPE O . . .

Examples from

https://github.com/jacobeisenstein/gt-nlp-class/blob/master/notes/eisenstein-nlp-notes-10-15-2018.pdf

Encoder-Decoder or seq2seq

$$\hat{t}_{j+1} \sim p(t_{j+1} = k \mid \hat{t}_{1:j}, c).$$

Where c can be **anything**.

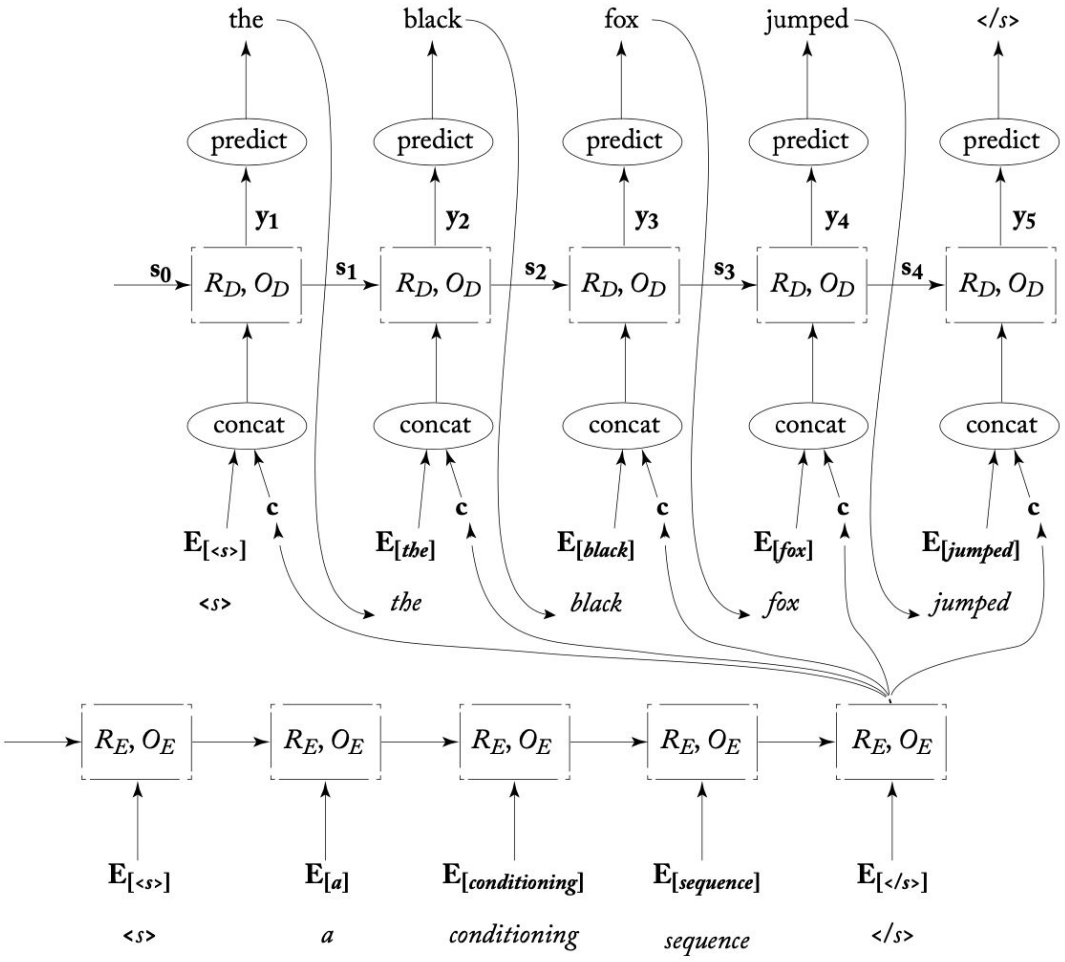
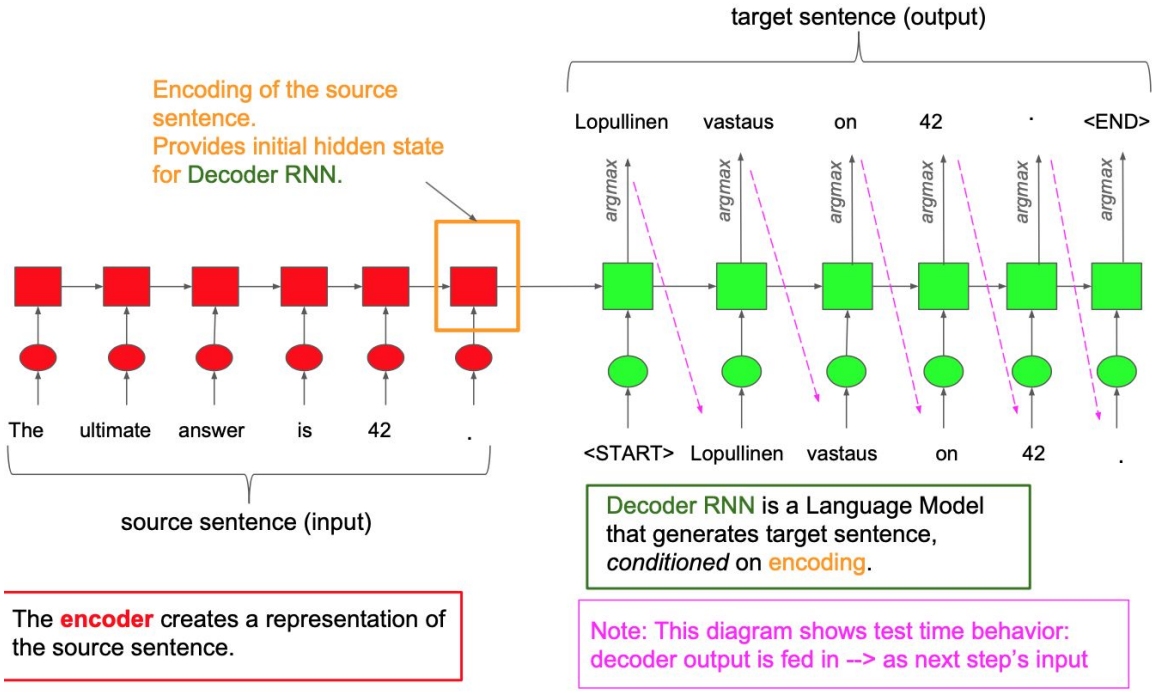


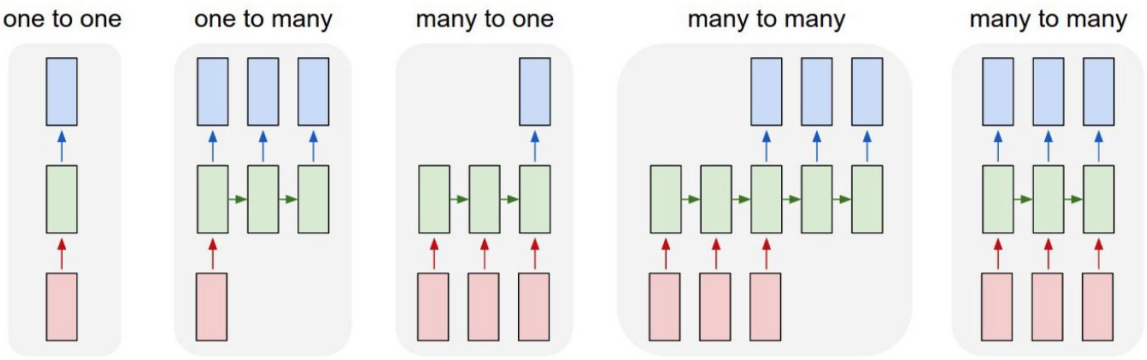
Figure 17.3: Sequence-to-sequence RNN generator.

- The neural network architecture is called sequence-to-sequence (aka seq2seq) or encoder-decoder and usually it involves *two* RNNs.



Besides machine translation, can you think of another application scenarios where an encoder-decoder architecture can fit?

What are concrete examples of the following architectures?



https://medium.com/cracking-the-data-science-interview/a-gentle-introduction-to-neural-networks-for-machine-learning-d5f3f8987786

😀An encoder processes the input sequence and compresses the information into a context vector of a fixed length. This representation is expected to be a good summary of the meaning of the whole source sequence.

😀A decoder is initialized with the context vector to emit the transformed output. The early work only used the last state of the encoder network as the decoder initial state.

😖 RNNs forget, in particular very long sentences.

https://lilianweng.github.io/lil-log/2018/06/24/attention-attention.html

⇒ use biLSTMs, stacked RNNs, etc. OR use attention.

Attention

Mechanism that can take the entire encoder context into account, that dynamically updates during the course of decoding, and that can be embodied in a fixed-size vector [JMch10].

This context vector, ci, is generated anew with each decoding step i and takes all of the encoder hidden states into account in its derivation.

Is the attention a parametrized mini-model in its own?

https://web.stanford.edu/~jurafsky/slp3/10.pdf

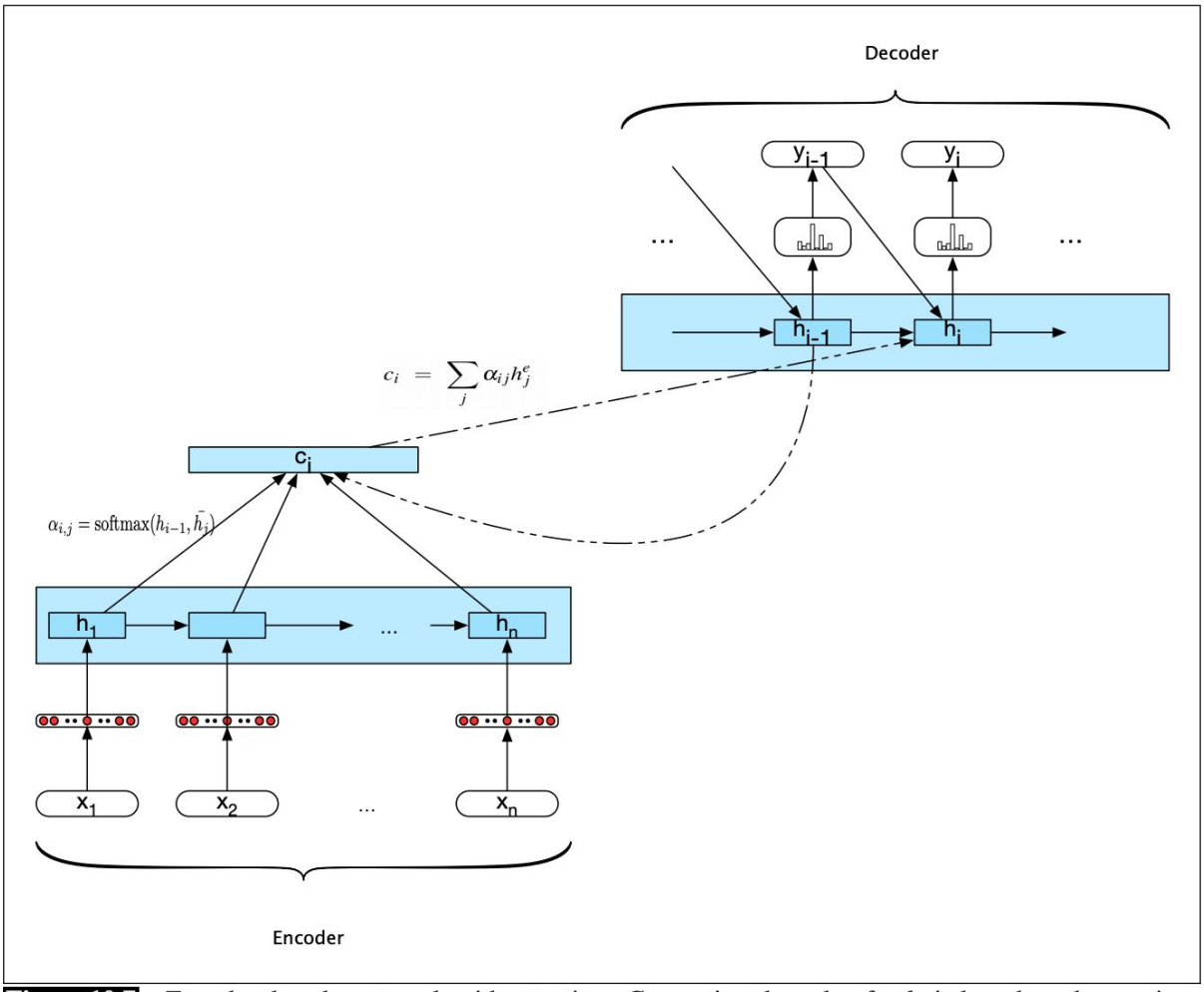


Figure 10.7 Encoder-decoder network with attention. Computing the value for h_i is based on the previous hidden state, the previous word generated, and the current context vector c_i. This context vector is derived from the attention computation based on comparing the previous hidden state to all of the encoder hidden states.

https://lilianweng.github.io/lil-log/2018/06/24/attention-attention.html

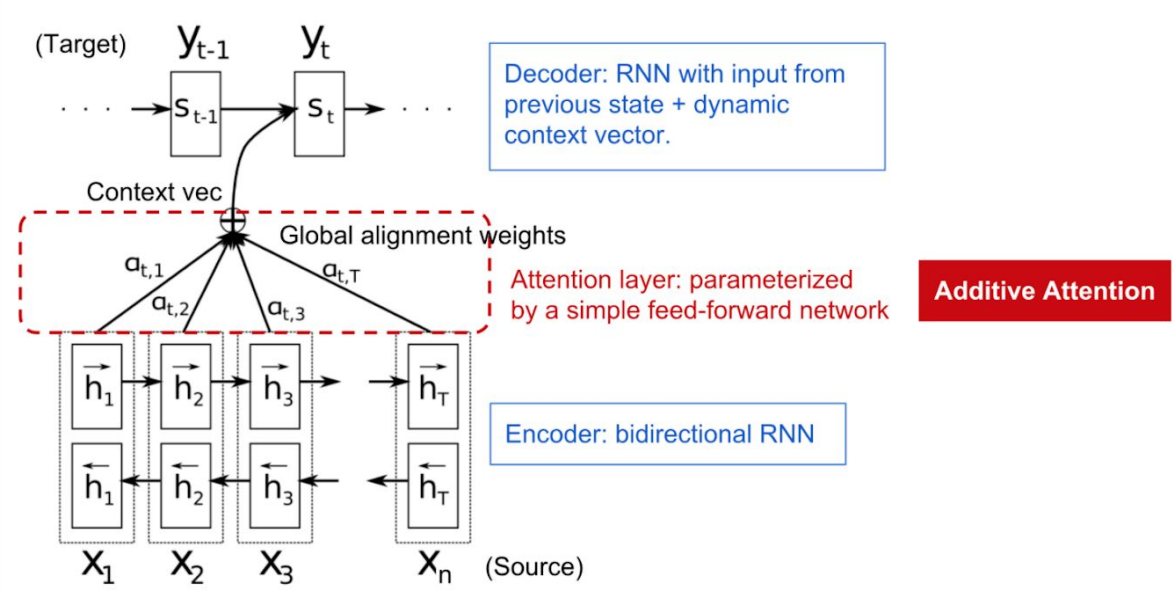


Fig. 4. The encoder-decoder model with additive attention mechanism in Bahdanau et al., 2015.

Many types of attention out there but attention and self-attention are among the main ones.

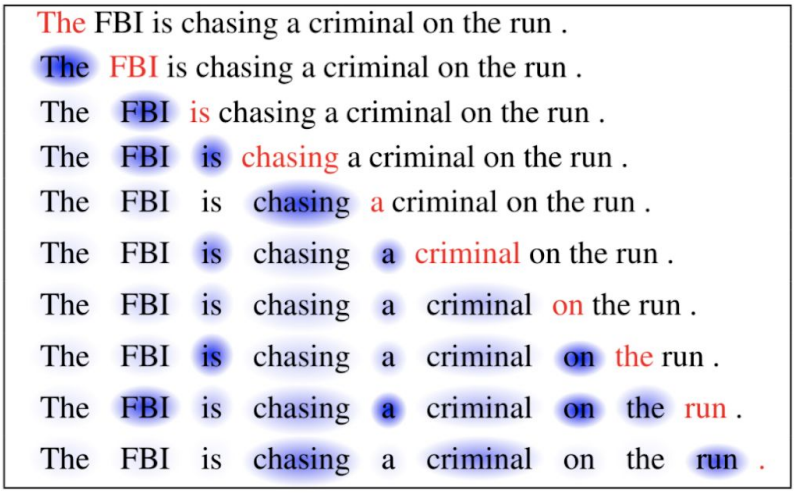


Fig. 6. The current word is in red and the size of the blue shade indicates the activation level. (Image source: Cheng et al., 2016)