Sequence to sequence

$$p(y_1, \dots y_J | x_1, \dots x_I) = \prod_{j=1}^{J} p(y_j | v, y_1, \dots y_{j-1})$$

- **Encoder:** maps the source sequence to the hidden vector

  RNN: $h_i = f(h_{i-1}, x_i)$ $\qquad v = h_I$

- **Decoder:** performs language modeling given this vector

  RNN: $s_j = g(s_{j-1}, [y_{j-1}, v])$

- **Prediction** (the simplest way):

  $p(y_j | v, y_1, \dots y_{j-1}) = softmax\,(U s_j + b)$