Compute (scalar) **alignment scores**
$$e_{t,i} = f_{att}(s_{t-1}, h_i) \qquad (f_{att} \text{ is an MLP})$$

Normalize alignment scores to get **attention weights**
$$0 < a_{t,i} < 1 \qquad \sum_i a_{t,i} = 1$$

Compute context vector as linear combination of hidden states
$$c_t = \sum_i a_{t,i} h_i$$

Use context vector in decoder: $s_t = g_U(y_{t-1}, s_{t-1}, c_t)$

From final hidden state: **Initial decoder state** s_0

This is all differentiable! Do not supervise attention weights – backprop through everything

Bahdanau et al, "Neural machine translation by jointly learning to align and translate", ICLR 2015