

$$e_{t,i,j} = f_{\text{att}}(s_{t-1}, h_{i,j})$$

$$a_{t,:} = \text{softmax}(e_{t,:})$$

$$c_t = \sum_{i,j} a_{t,i,j} h_{i,j}$$



$h_{1,1}$	$h_{1,2}$	$h_{1,3}$
$h_{2,1}$	$h_{2,2}$	$h_{2,3}$
$h_{3,1}$	$h_{3,2}$	$h_{3,3}$

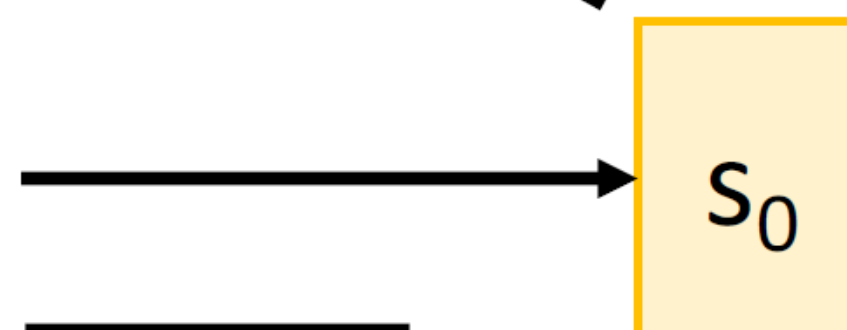
Alignment scores

$e_{1,1,1}$	$e_{1,1,2}$	$e_{1,1,3}$
$e_{1,2,1}$	$e_{1,2,2}$	$e_{1,2,3}$
$e_{1,3,1}$	$e_{1,3,2}$	$e_{1,3,3}$

softmax

Attention weights

$a_{1,1,1}$	$a_{1,1,2}$	$a_{1,1,3}$
$a_{1,2,1}$	$a_{1,2,2}$	$a_{1,2,3}$
$a_{1,3,1}$	$a_{1,3,2}$	$a_{1,3,3}$



Use a CNN to compute a grid of features for an image