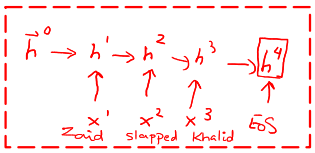
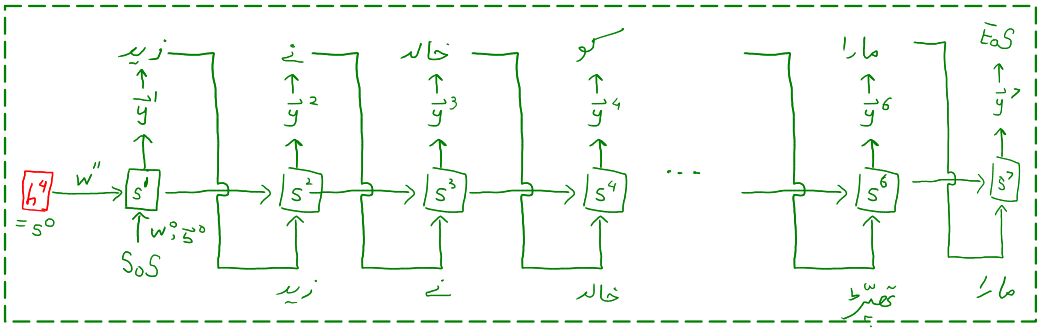
# Attention

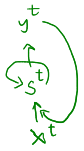
Continuing with our language translation problem.

Zaid slapped Khalid → زید نے خالد کو تھپڑ مارا



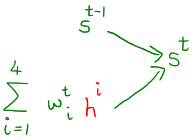




Only $s^1$ takes input from only $h^4$



## Decoder with attention

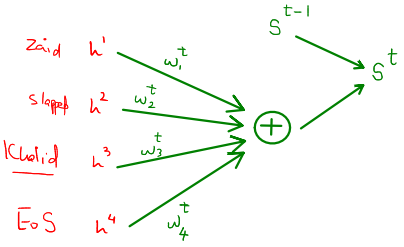
1. Replace encoding $h^4$ by a weighted sum of all encodings.
2. Feed weighted sum of encodings to each hidden state $s^t$.
3. Weights change for each time step.



$$\sum_{i=1}^{4} w_i^t h^i$$

- Decoder "looks at" encodings $h^1, h^2, h^3, h^4$ using weights specific to each time $t$.
- $w(t,i)$ is the weightage given to encoding $h^i$ when finding output at time $t$.
- In other words, $w(t,i)$ is the weightage given to input $i$ when producing output at time $t$.
- For example,
  - If $w(t=3, i=3)$ is high, then output خالد is produced by "looking at" or "attending to" input Khalid.
  - If $w(t=5, i=2)$ is high, then output تھپڑ is produced by attending to input slapped.

## How to get attention weights $w_i^t$?




- Make $w_i^t$ depend on $s^{t-1}$ and $h^i$.
- To ensure weighted average, compute $w_i^t$ via softmax to produce probability values.

$$w_i^t = \frac{\exp(u_i^t)}{\sum_{j=1}^{4} \exp(u_j^t)}$$

## Options for computing unnormalized weights $u_i^t$

1. $u_i^t = h^i \cdot s^{t-1}$ ← Favor input encoding similar to present hidden state.
2. If sizes of encoding and decoding are different, then use a learnable linear projection $W_a$.
   $$u_i^t = h^i \cdot (W_a s^{t-1})$$
3. $$u_i^t = v_a \cdot \tanh\left(W_a \begin{bmatrix} h^i \\ s^{t-1} \end{bmatrix}\right)$$
4. $$u_i^t = \text{MLP}\left(\begin{bmatrix} h^i \\ s^{t-1} \end{bmatrix}\right)$$ with a single linear output neuron.

} Learning a model for computing weights $w_i^t$.

Now let's consider a different problem with a static encoder but dynamic decoder.

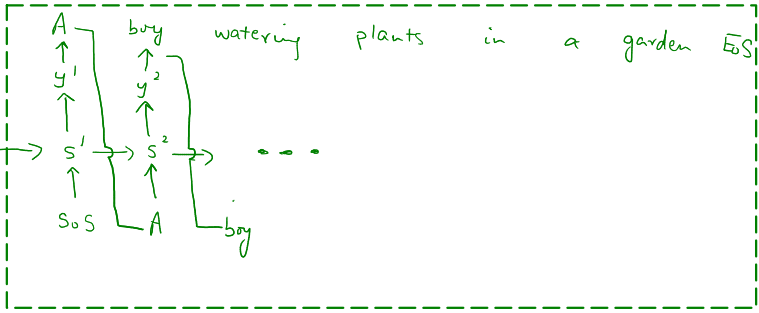
Image Captioning

MLP
CNN
⋮

Every filter's result can be considered as some encoding of the image.

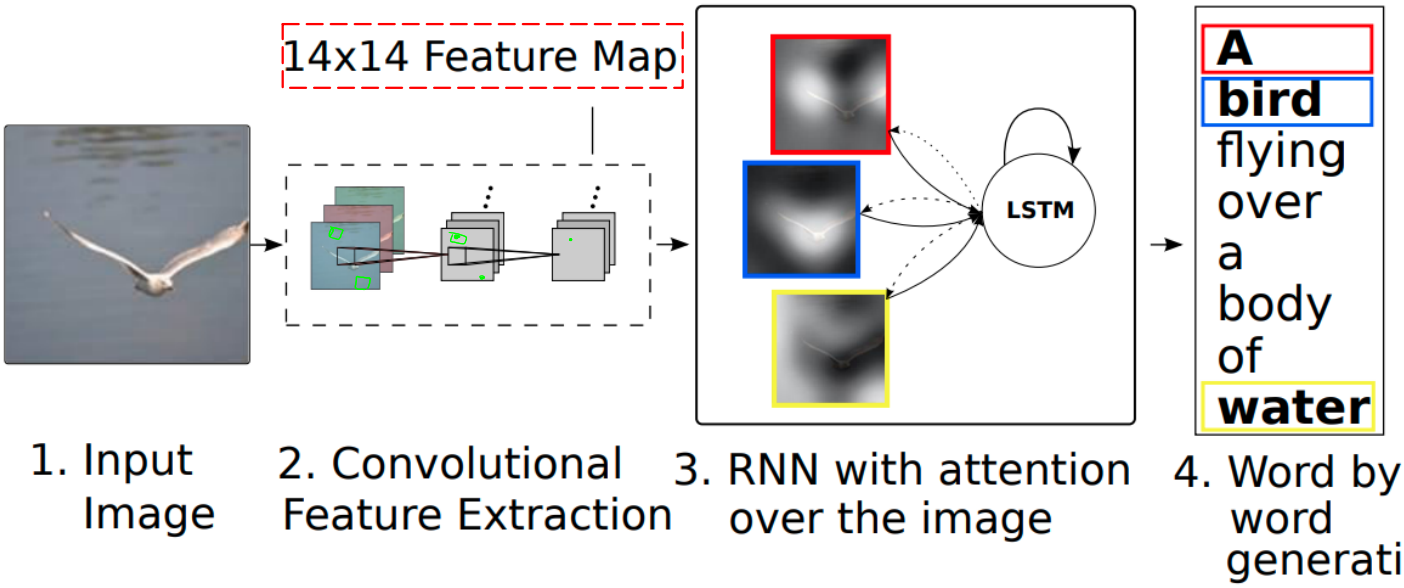
Encoding






One - to - many

Attention-based decoder for image captioning.




- $S^0$ is a 14x14 array produced after a series of convolutions and subsamplings.
- Each pixel in $S^0$ represents a portion of the input image.
- Attention weight $w_i^t$ represents the importance of portion $i$ in producing output word at time $t$.