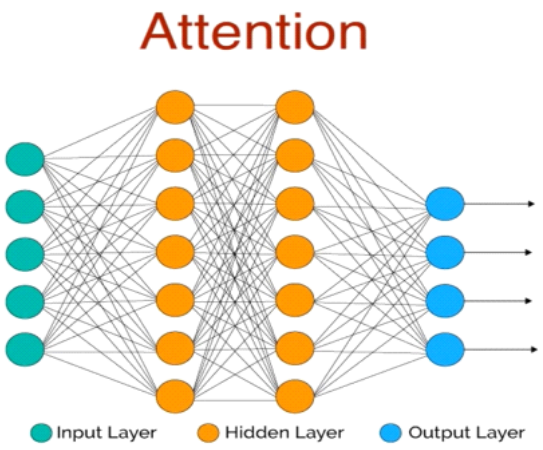
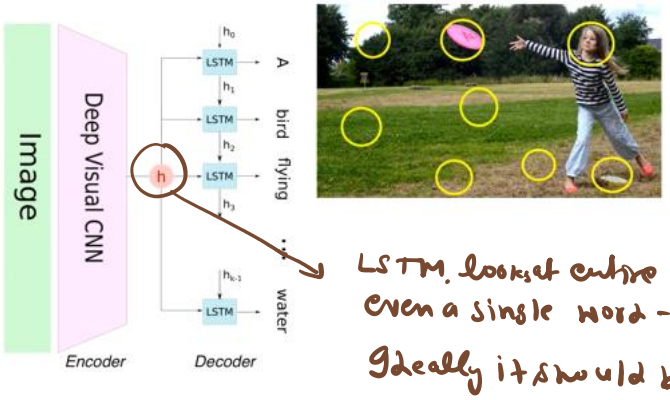
Attention in NN Basics

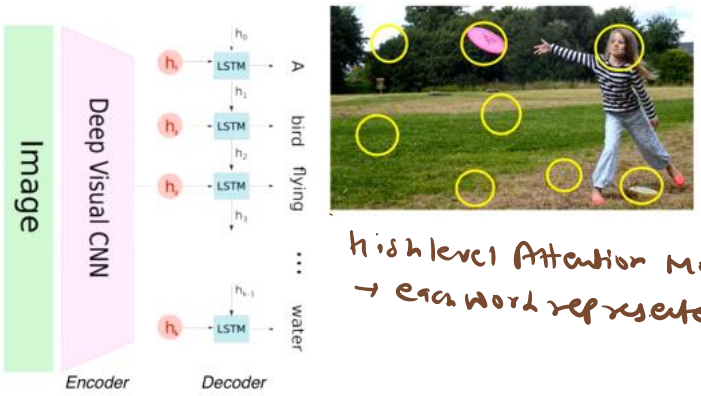
Saturday, October 23, 2021 11:43 AM





LSTM looks at entire image via this layer to caption even a single word - which is inefficient Method

Ideally it should be very specific

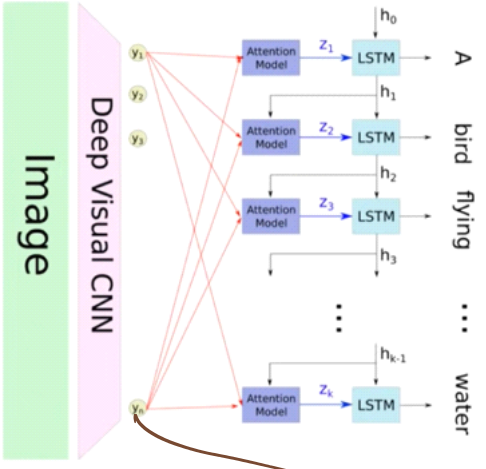


high level Attention Model

→ each word represented by a different vector independent

A Girl throws a Frisbee in the park.

Q - How does it exactly decide the regions to consider??



$$E(X) = \sum_n p(X = X_n)\, X_n$$

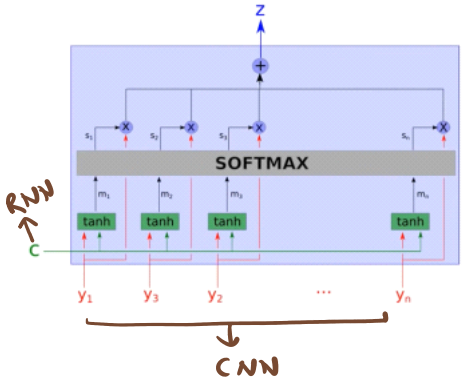
Ans- by using Attention unit considers all sub regions and context for all inputs & output a weighted Arithmatic mean of these region

This mean is the Product of Prob. & actual values

The Prob come from the context

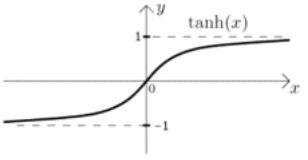
$m_i = Cy_i$ → CNN, * , ↓ RNN

Attention Unit



$$m_i = \tanh\left(y_i W_{y_i} + C W_c\right)$$

CNN, RNN

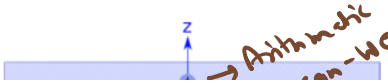


Alternatively $m_i = Cy_i$ can also be applied but tanh is more fine grained

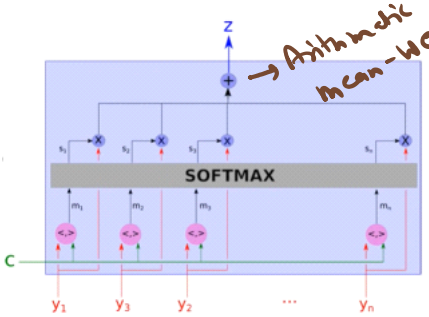
m_i - learnable parameter of Attention unit

tanh is used to convert very high values very close to 1 so that they have a small differences & vice versa

Attention Unit



$$m_i = cy_i$$



→ Arithmetic mean-weighted avg

$m_i = cy_i$

$s_i = \frac{e^{m_i}}{\sum_n e^{m_n}}$ $\quad s_i \in [0,1]$ $\quad \sum_n s_n = 1$ — Softmax Probability - determine the relevance

$z = \sum_n s_n y_n$ - output of a specific region given context

Types of Attention

1. Soft Attention: different parts, different subregions

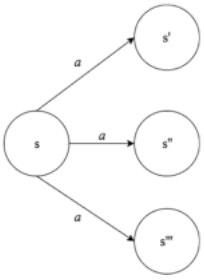
$$z = \sum_n s_n y_n$$

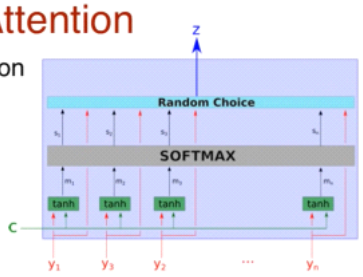
Soft Attention is *Deterministic*

Types of Attention

2. Hard Attention: only ONE subregion

Hard Attention is *Stochastic*

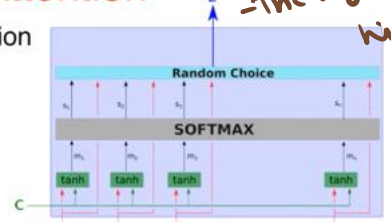




Types of Attention

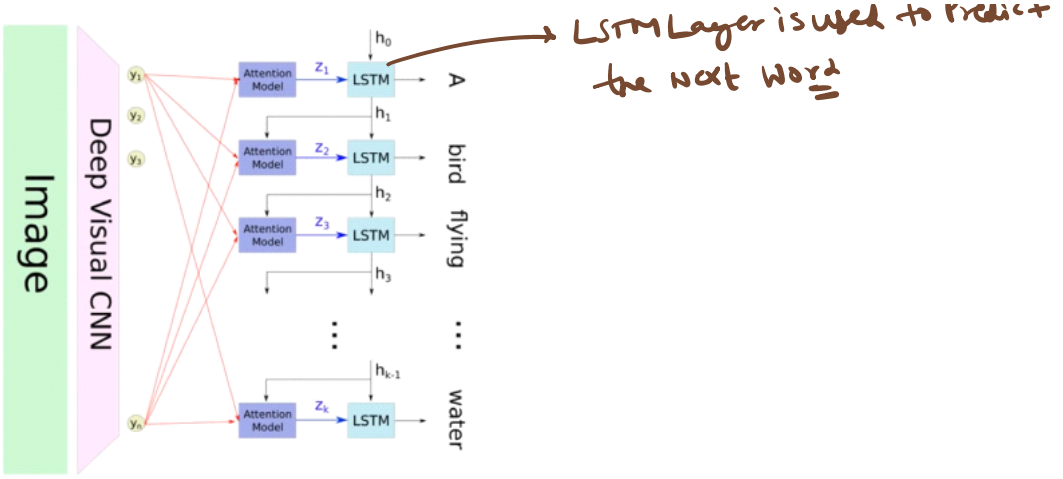
2. Hard Attention: only ONE subregion

Hard Attention is *Stochastic*



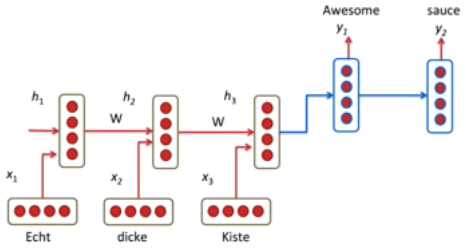
- The region which has highest Prob' has more chances to come

→ LSTM Layer is used to Predict the next word



Applications

1. Neural Machine Translation (NMT)



Applications

2. Microsoft's Attention GANs

Figure 1. Example results of the proposed AttnGAN. The first row gives the low-to-high resolution images generated by G_0, $G1$ and G_2 of the AttnGAN; the second and third row shows the top-5 most attended words by F_1^{attn} and F_2^{attn} of the AttnGAN, respectively. Here, images of G_0 and G_1 are bilinearly upsampled to have the same size as that of G_2 for better visualization.

Applications

3. Teaching Machines to Read & Comprehend