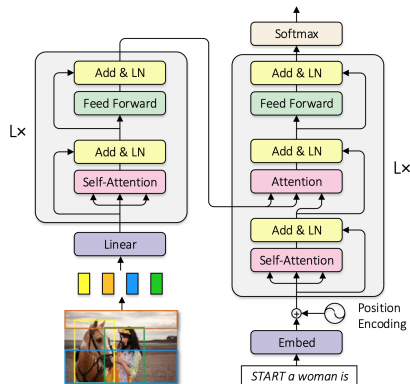


# Normalized and Geometry-Aware Self-Attention Network for Image Captioning

July 15, 2020

# Self-Attention Networks for Image Captioning

- ▶ Take a usual Transformer model for machine translation;
- ▶ Use a pretrained Faster-RCNN model to extract objects and pass objects' features as an input to the encoder.
- ▶ Do not use positional embeddings for encoder



# Normalization for an attention mechanism (N-SAN)

- ▶ Attention weights are calculated as:

$$\begin{aligned} S &= \text{Softmax}(QK^\top) \\ &= \text{Softmax}((XW_Q) \cdot (W_K^\top X^\top)) \end{aligned} \tag{1}$$

- ▶ The paper shows that it is beneficial to apply Instance Normalization to matrix  $Q$ :

$$\begin{aligned} \hat{x}_{btc} &= \frac{x_{btc} - \mu_{bc}}{\sqrt{\sigma_{bc}^2 + \epsilon}} \\ \mu_{bc} &= \frac{1}{T} \sum_{t=1}^T x_{btc}, \sigma_{bc}^2 = \frac{1}{T} \sum_{t=1}^T (x_{btc} - \mu_{bc})^2 \end{aligned} \tag{2}$$

- ▶ i.e. we normalize each sample independently across time dimension;

# Incorporating geometry information (G-SAN)

- ▶ What if we mix positional information into attention calculation?

$$S = \text{Softmax} (QK^{\top} + \phi(Q', K', G)) \quad (3)$$

- ▶ Matrix  $G$  carries some non-trivial information about objects geometry.
- ▶ Here  $\phi$  is a matrix of  $\phi_{ij}(Q'_i, K'_j, G_{ij})$
- ▶  $\phi_{ij}$  is a one-layer NN on top of combinations of  $Q'_i, K'_j$  and  $G_{ij}$ .
- ▶ Authors consider different variants to combine three ingredients together.
- ▶ We compute  $G_{ij} = \text{ReLU} \left( \text{FC} \left( \mathbf{f}_{ij}^g \right) \right)$  from  $\mathbf{f}_{ij}^g$  which is a 4-dimensional vector of:

$$\mathbf{f}_{ij}^g = \left( \log \left( \frac{|x_i - x_j|}{w_i} \right), \log \left( \frac{|y_i - y_j|}{h_i} \right), \log \left( \frac{w_i}{w_j} \right), \log \left( \frac{h_i}{h_j} \right) \right)^T \quad (4)$$

## Some results and considerations

- ▶ Authors test an NG-SAN model on 2 tasks: image captioning, video question answering;
- ▶ Authors additionally test an N-SAN model on 2 tasks: video captioning, machine translation.
- ▶ Almost all the scores are improved somewhat marginally: +0.1-0.3 absolute points (0.2-0.7% of relative improvement). But seems like it is a lot for image captioning.
- ▶ Authors do not provide stds of the runs which would be very helpful.