Feature Composition Approaches

By Abhilasha

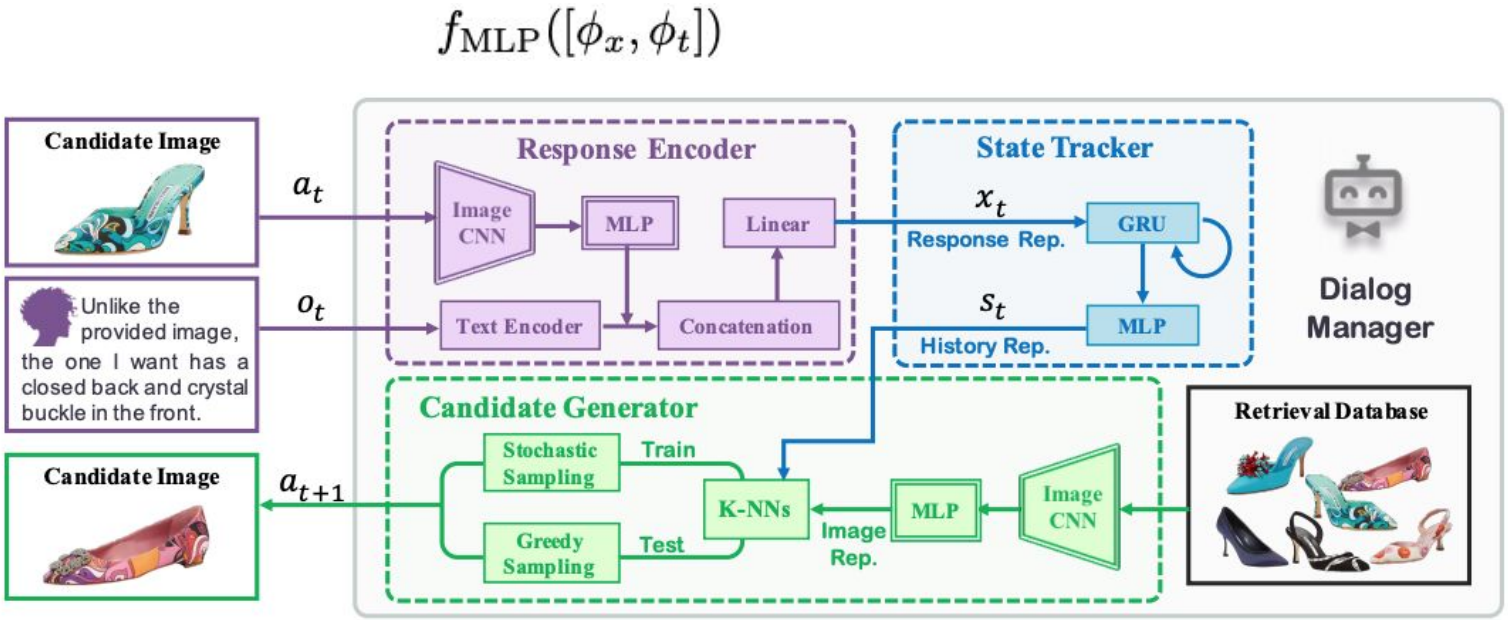
Notation

Goal : To learn embedding space for image + text

Image : φx

Text : φt

1. Concatenation



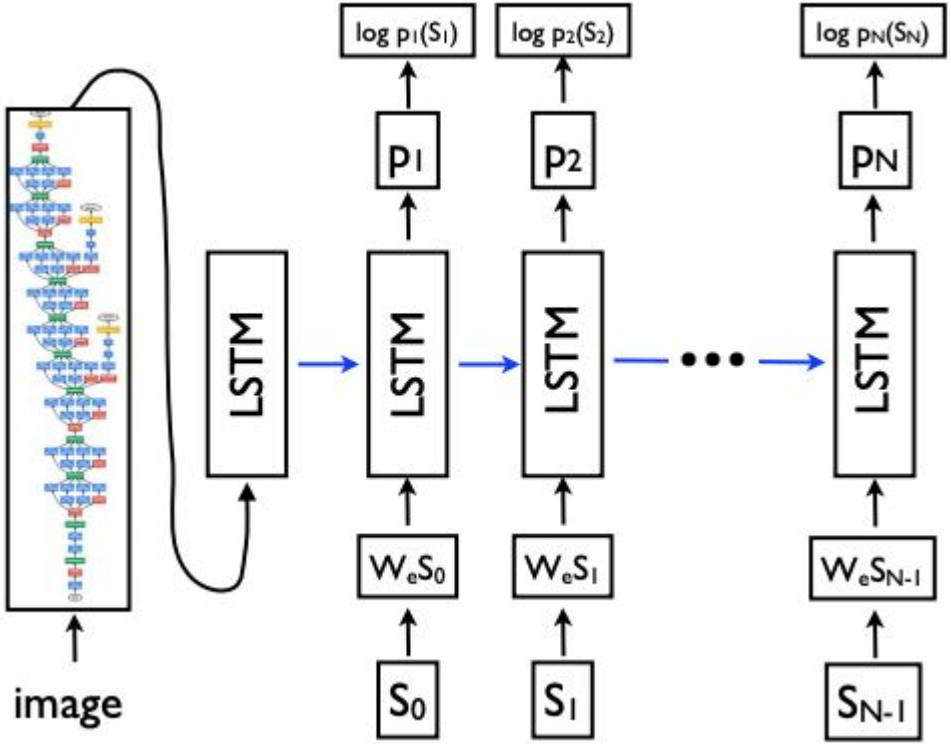
Source :[Dialog-based Interactive Image Retrieval] IBM research AI
https://arxiv.org/pdf/1805.00145.pdf

3. Attribute as Operator

embeds each text as a transformation matrix, Tt, and applies Tt to φx to create φxt.

Source :
http://openaccess.thecvf.com/content_ECCV_2018/papers/Tushar_Nagarajan_Attributes_as_Operators_ECCV_2018_paper.pdf

2. Show and Tell : A Neural Image Caption generato



we train a LSTM to encode both image and text by inputting the image feature first, following by words in the text; the final state of this LSTM is used as representation φxt.

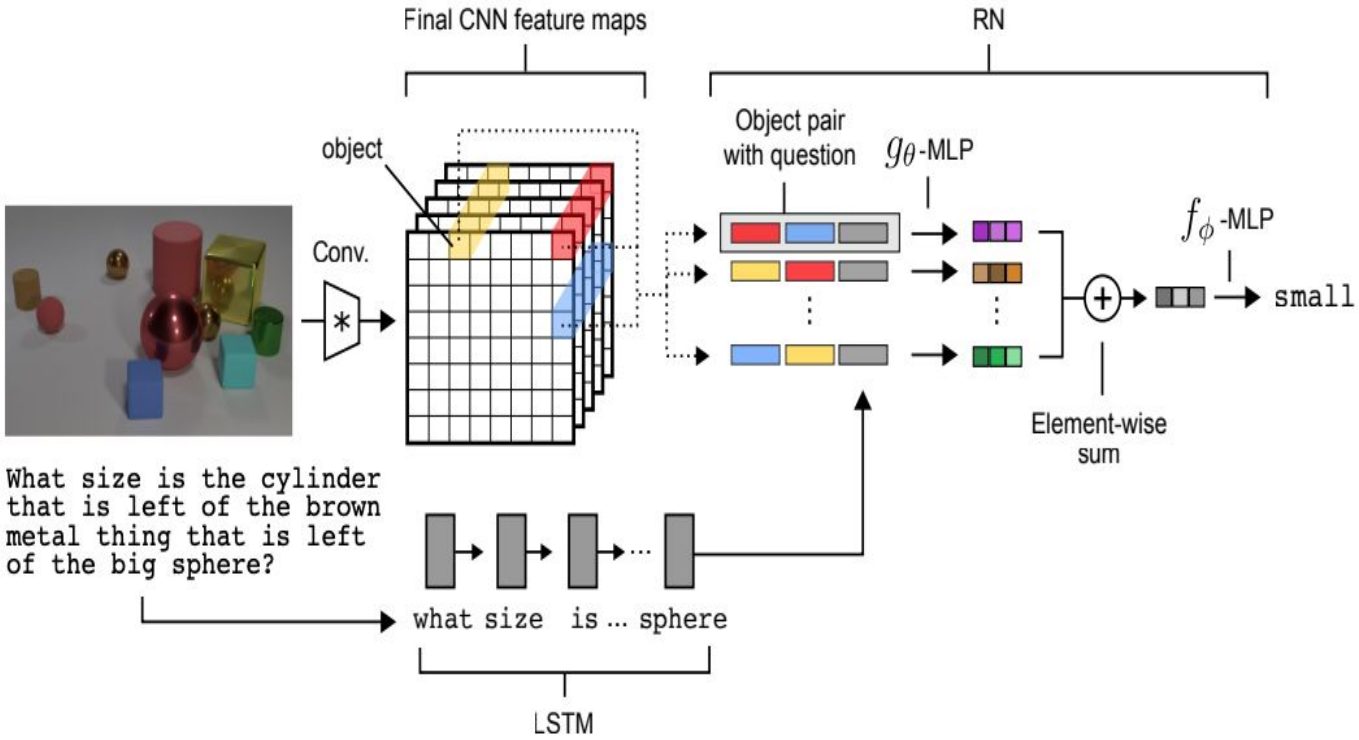
4. Parameter hashing

It is a technique used for the VQA task. In our implementation, the encoded text feature φt is hashed into a transformation matrix Tt, which can be applied to image feature; it is used to replace a fc layer in the image CNN, which now outputs a representation φxt that takes into account both image and text feature.

5. Relationship

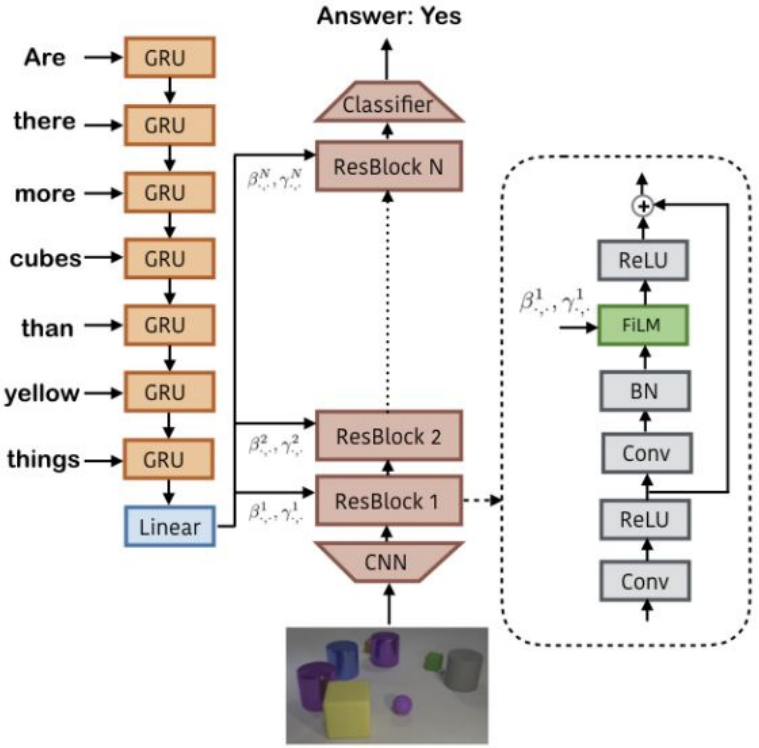
https://papers.nips.cc/paper/7082-a-simple-neural-network-module-for-relational-reasoning.pdf

It is a method to capture relational reasoning in the VQA task. It first uses CNN to extract a 2d feature map from image, then create a set of relationship features, each is a concatenation of the text feature φt and 2 local features in the 2d feature map; this set of features is passed through a MLP and the result is averaged to get a single feature φxt.



https://arxiv.org/pdf/1709.07871.pdf

6. FiLM: Visual Reasoning with a General Conditioning Layer



https://arxiv.org/pdf/1812.07119.pdf

TIRG : Composing Text and Image for Image Retrieval - An Empirical Odyssey

$$\phi_{xt}^{rg} = w_g f_{\text{gate}}(\phi_x, \phi_t) + w_r f_{\text{res}}(\phi_x, \phi_t),$$

$$f_{\text{gate}}(\phi_x, \phi_t) = \sigma(W_{g2} * \text{RELU}(W_{g1} * [\phi_x, \phi_t]) \odot \phi_x$$

$$f_{\text{res}}(\phi_x, \phi_t) = W_{r2} * \text{RELU}(W_{r1} * ([\phi_x, \phi_t])),$$