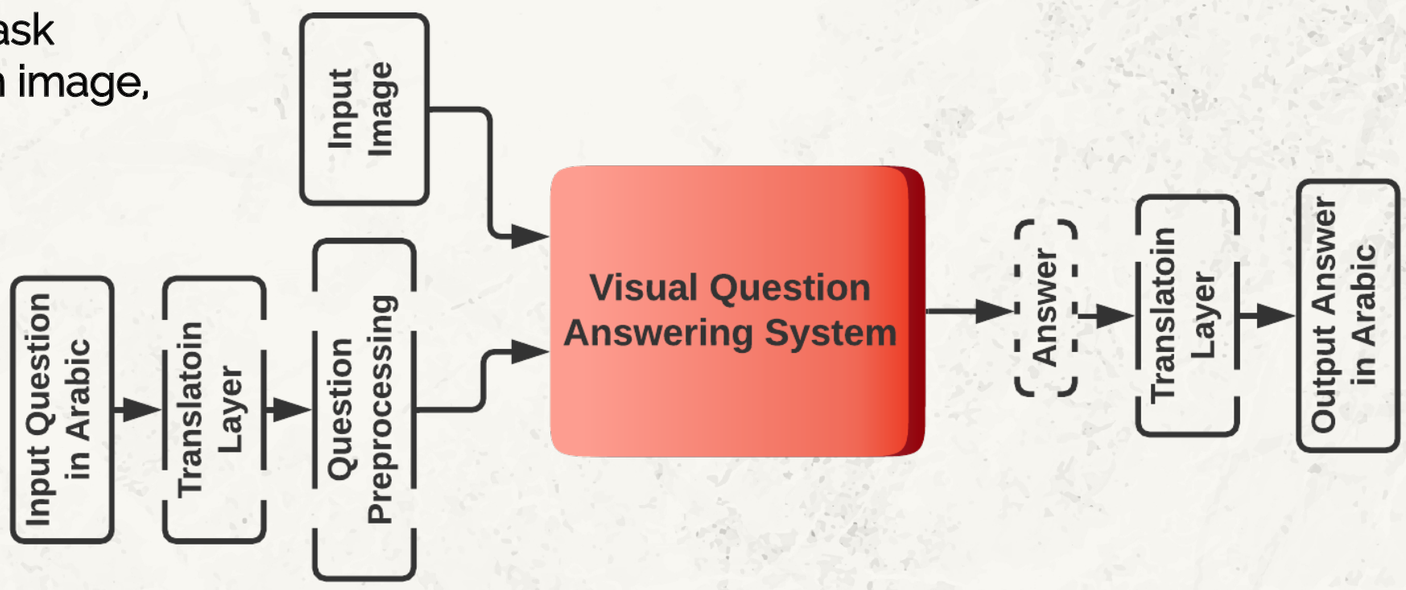
Arabic Visual Question Answering

Mariam Safieldin

Supervisor: Professor Shady ElBassuoni

Visual Question Answering (VQA) is a computer vision task where a system is given a text-based question about an image, and it must infer the answer.



Dataset

VQA v2.
Real MSCOCO images.
Questions has 5 categories:
Y/N - Object - Number - Color - Location.

Translation Layer

deep-translator python library.
Added extension to translate English numbers into Arabic.

Question Preprocessing

Special Chars Removal - Expanding Contractions - 1K Top Frequesnt Answers - Filter Questions - Tokenize - Padding |200|

Investigated Models

A - CNNs + LSTMs Encoder

Image Featurization:
CNN of 7 layers.
(64 x2 - Max2DPooling
128 x2 - Max2DPooling
256 x3 - Max2DPooling
Flatten)

Question Featurization:
3 LSTMs initialized with Glove word embeddings of dimension (300,).

Baseline Model

-Image Featurization: VGG16 pretrained model.
-Question Featurization: 1 LSTM initialized with Glove word embeddings of dimension (50,).

Joint Comprehension Concatenation

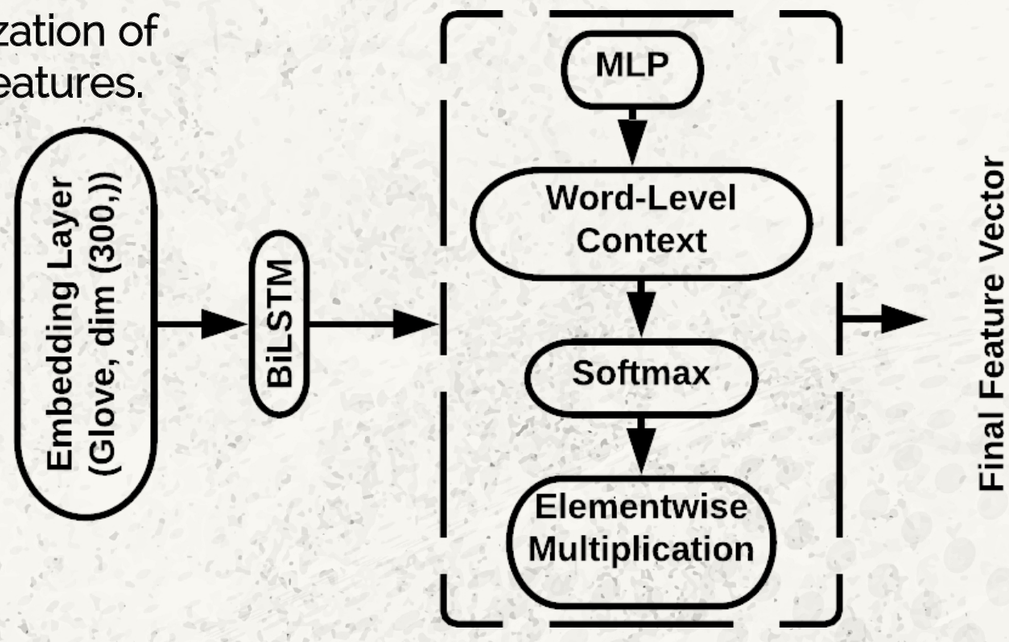
Answering Approach

Dense layer with softmax as an activation function (1000 units).

B - VGG16 + BiLSTM + Attention Mechanism Encoder

Image Featurization:
VGG16 pretrained model.
L2 Normalization of extracted features.

Question Featurization:



Experimental Results

Train Data Size: 100% - 100% - 80/45%
Validation Data Size: 100% - 100% - 20/45%

Model	Accuracy (%)
VGG16 + 1 LSTM	48.36
CNNs + 3 LSTMs	52.51
VGG16 + Bi-LSTM + Attention Mechanism	**59.68**

Conclusion

An Arabic visual question answering system was implemented. First model employs CNNs and LSTMs for encoding the image and question. Second model employs VGG16 pretrained model and a BiLSTM with an attention mechanism for encoding the image and question. The resulting features are concatenated and fed to a softmax layer to output an answer that is passed to the translation layer to generate the final answer in Arabic.

References

10.1109/ICCV.2017.285
arxiv:1511.02274
10.18653/v1/D16-1044
arXiv:1409.1556
10.3115/v1/D14-1162
arXiv:1506.00333
arXiv:1506.07285
arXiv:1610.01465v4
10.1007/s10462-020-09832-7
10.1016/j.image.2019.115648
arXiv:1705.03865v2
arXiv:1903.00366
arXiv:1906.10770
arxiv:1707.07998
arXiv:1505.00468
arxiv:1605.02697
arXiv:1612.07411
10.18653/v1/N16-1174
arXiv:1405.0312
pypi.org/project/deep-translator