Visual Question Answering

Amlan Kar, Nishant Rai, Shivam Malhotra

Problem Statement

We aim to build a deep learning system capable of answering open ended questions on real world images. We use the VQA dataset[1] released by Virginia-Tech for our experiments that contains images from the MS-COCO dataset annotated with open ended questions and top answers

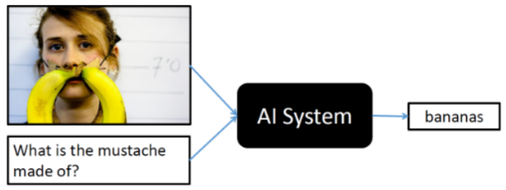


Fig.1: Problem Statement

Theory

- **Convolutional Neural Network**
  A Convolutional Neural Network is an artificial neural network which work by sliding windows through it's input looking for local features. These have shown to work extremely well for Image recognition tasks.

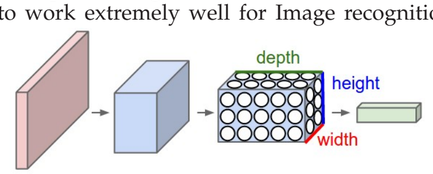


Fig.2: CNN Illustration [a]

- **Semantic Word Embeddings:**
  Recent advances such as word2vec, GloVe[2] and skip-thoughts[3] map words or sentences to high dimensional real valued vectors such that syntactic relations between the words are preserved. These have been shown to have strong semantic similarity properties as well.

- **Recurrent Neural Network**
  A recurrent neural network (RNN) is a class of artificial neural network where connections between units form a directed cycle. These have been shown to work extremely well in modelling temporal data.

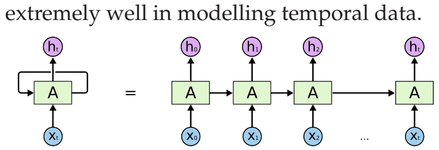
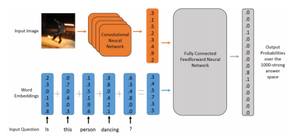


Fig.3: RNN Illustration [b]

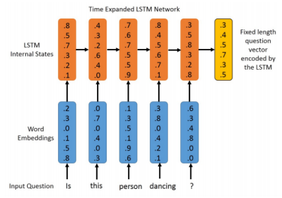
[a] Image taken from www.cs23 1n.github.io
[b] Image taken from www.colah.github.io

Baselines

- **CNN + BOW**[a]



- **CNN + LSTM**[a]



[a] Images taken from www.avisingh599.github.io

Incorporating Attention

- **Spatial Attention** Spatial Attention methods are a class of methods that primary work by selecting a subset of feature vectors (or learning a probability vector) from one of the earlier conv layers of a feature extractor model. These have been shown to work well for Caption Generation[4] and VQA[5]

- **Semantic Attention** We implement and present semantic attention models. Basically, we learn a question embedding and a transformation matrix into the image dimension and use this as a semantic attention weight vector on the semantic image features using MLPs and RNNs.

Results

All our models attained their optimal validation performance at roughly around 60 epochs of training. The evaluation metric and the results are described below:

The **evaluation metric** used for the VQA Challenge consists of evaluating an answer(processed) against 10 human answers(based on sampling from distribution). The accuracy of an answer is calculated according to:

$$\text{Acc}(ans) = \min\left\{\frac{\#\text{humans that said } ans}{3}, 1\right\}$$

Method	Accuracy
CNN+BoW	48.46
CNN+LSTM	51.63
ABC-CNN[5]	48.38
CNN+BowAtt	47.32
CNN+LSTMAtt	48.27
CNN+LSTM2	**55.17**

Table 1: Accuracy scores on VQA-test-dev split

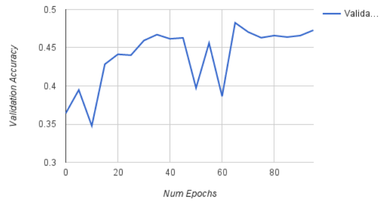
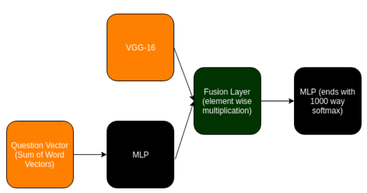


Figure 4. Accuracy vs Num Epochs (CNN+LSTMAtt)

Models Proposed

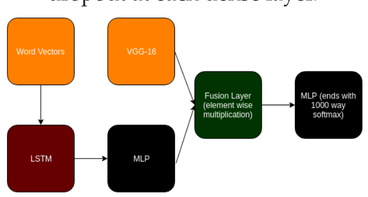
CNN+BOWAtt

- 3-Layer MLP to learn the matrix transfer embedding from the word vector space to the Image space. Trained with 50% dropout at each dense layer



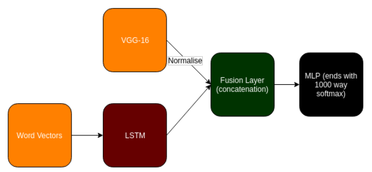
CNN+LSTMAtt

- 2-Layer LSTM used to create question embedding space from variable sized questions. 3-Layer MLP to learn the space embedding transfer. Trained with 50% dropout at each dense layer.



CNN+LSTM2

- Fusion layer only concatenates normalized CNN output with the 2-Layer LSTM question vector embedding. Again, trained with 50% dropout at each dense layer.



Observations

1. While the CNN+LSTMAtt model doesn't give competitive accuracies, it gets many 'tough' questions right which the CNN+LSTM2 model gets wrong. This suggests the possibility of using an ensemble.

2. The attention model failing might be due to the usage of the fused features alongwith the vanilla VGG and Question features. This redundancy and its subsequent effect on the extremely higher number of parameters in the model might be the reason for its poor test performance.

3. Attention doesn't seem to be very important to the Virginia-Tech VQA dataset with our preliminary results and results from the ABC-CNN[5] paper.

Future Work

1. Using Normalized CNN Features in all models[1]

2. Using a Knowledge base for real world reasoning

3. A mixture model of spatial and semantic attention

4. Perhaps using random 224x224 splits of an image to find the ConvNet features and averaging over results (as done by Krizhevsky et al. 2012) [6]

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433, 2015.

[2] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[3] Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Skip-thought vectors. *arXiv preprint arXiv:1506.06726*, 2015.

[4] Kelvin Xu, Jimmy Ba, Ryan Kiros, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*, 2015.

[5] Kan Chen, Jiang Wang, Liang-Chieh Chen, Haoyuan Gao, Wei Xu, and Ram Nevatia. Abc-cnn: An attention based convolutional neural network for visual question answering. *CoRR*, abs/1511.05960, 2015.

[6] Qi Wu, Peng Wang, Chunhua Shen, Anton van den Hengel, and Anthony Dick. Ask me anything: Free-form visual question answering based on knowledge from external sources. *arXiv preprint arXiv:1511.06973*, 2015.