Visual Question Answering using Deep Neural Networks

ENGR-E-533: Deep Learning Systems
Indiana University Bloomington
Fall-2022

Team Members: Keerthana Sugasi (ksugasi@iu.edu)
Madhavan Kalkunte Ramachandra (madhkr@iu.edu)
Zachary James Petroff (zpetroff@iu.edu)

INDIANA UNIVERSITY BLOOMINGTON

Agenda

- Introduction
- Related Work
- Datasets
- Proposed Work
- Results
- Conclusion & Future Work

Introduction

What is VQA?

Given an image and a natural language question about the image, Visual Question Answering (VQA) is the task of providing an accurate natural language answer.

Types of questions?

- Truth based questions - Yes/No Answers.
- Word based questions - One word answers.

Applications of VQA

- Assisting the visually impaired.
- Translating text from images.
- Image retrieval systems.

Related Work

Types of approaches to solve VQA:

- **Non Neural Approaches**: The non-neural network approach generally takes a probabilistic approach to answer the questions.
    - *Related works*: *Multi-World QA, Answer Type Prediction*

- **Joint Embedding Approaches:** The concept of jointly embedding text and images allows one to learn the representation in a common feature space.
    - *Related works*: *Neural-Image-QA, Multimodal QA(mQA)*

- **Attention Mechanisms**: Attention mechanisms use local image features and assign different importance to different regions to improve the performance of models.
    - *Related works*: *Word-Guided model, Co-Attention model*

- **Compositional Models**: Compositional models propose modular network architecture for the VQA task, which involves connecting independently developed models that solve specific portions of the problem to generate a final output.
    - *Related works: Neural Module Networks*

- **Models using external knowledge bases:** Models that rely on additional information, both textual and visual, to generate a better understanding of the inputs and hence to answer.
    - *Related works*: *ahab*

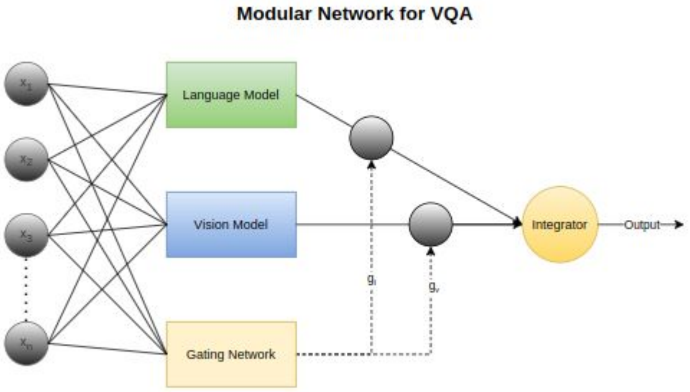
Datasets

Datasets used for VQA Tasks:

- **DAQUAR:** First and benchmarked dataset for VQA task. Contains mostly indoor images with semantic segmentation.
  - #images: 1449, #questions: 12468
- **Visual7W:** Images are derived from MS-COCO dataset and the questions are focussed for image captioning purposes.
  - #images:47300, #questions: 327939
- **Visual Madlibs:** Images are sourced from MS-COCO dataset and questions are majorly fill-in-the-blanks and Multiple choice.
  - #images: 10738, #questions: 360001
- **FM-IQA:** Images drawn from MS-COCO dataset, questions are either in english or chinese. Answers were generated using Baidu crowdsourcing server.
  - #images: 158392, #questions: 316193
- **COCO-QA:** Images sourced from MS-COCO dataset and contains one question per image
  - #images: 123287, #questions: 123287
- **VQA:** Contains open ended questions about images sourced from MS-COCO dataset.
  - #images: 265016 images, #questions: 1431087

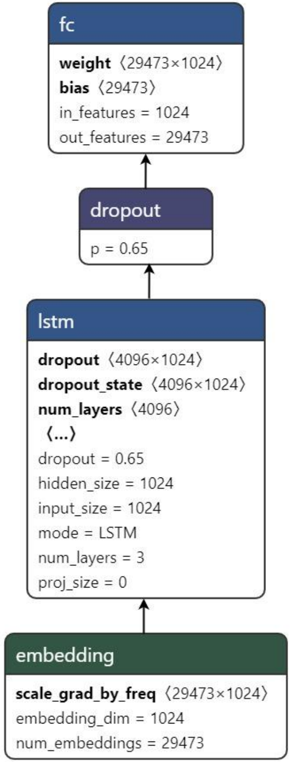
Proposed Work

- **Modular Networks:** A modular neural network is made up of several smaller neural network models that solve portions of a problem, with an integrator model to combine the outputs.
- VQA implemented as a Modular ANN that has independent models to learn the natural language and image aspects of VQA.
- Independent Language and Vision models are built as external knowledge-based models.
- **Language model** extract lingual information using RNN network and trained on WikiText dataset.
- **Vision model** extract image features using Autoencoders made up of Conv and Deconv layers and trained on COCO dataset.
- **The Modular network** combines the output of language and vision models and feeds to a DNN to make the final prediction.

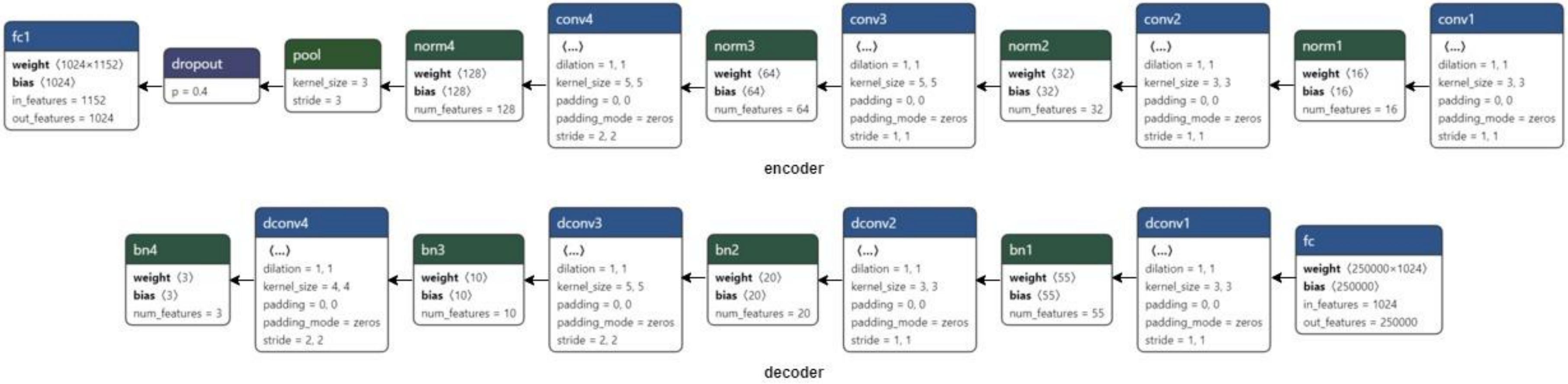


Language Model

- Embedding layer: Converts input into a 1024-dimensional feature vector.
- 3 LSTM layers: Each layer has a input and hidden state size of 1024 and a dropout of 0.65.
- Fully Connected layer: The fully connected layer converts input into a 29473 dimensional vector of probabilities. 29473 is the size of the network's vocabulary and each probability corresponds to a word in the network's vocabulary.

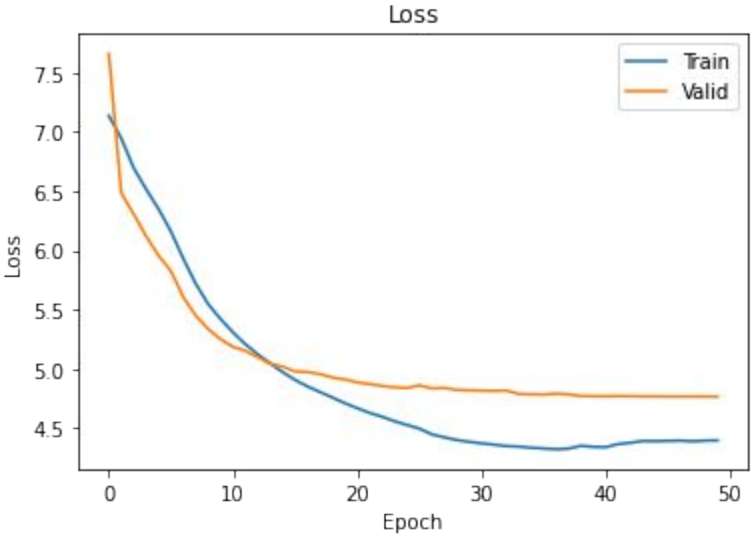


Vision Model

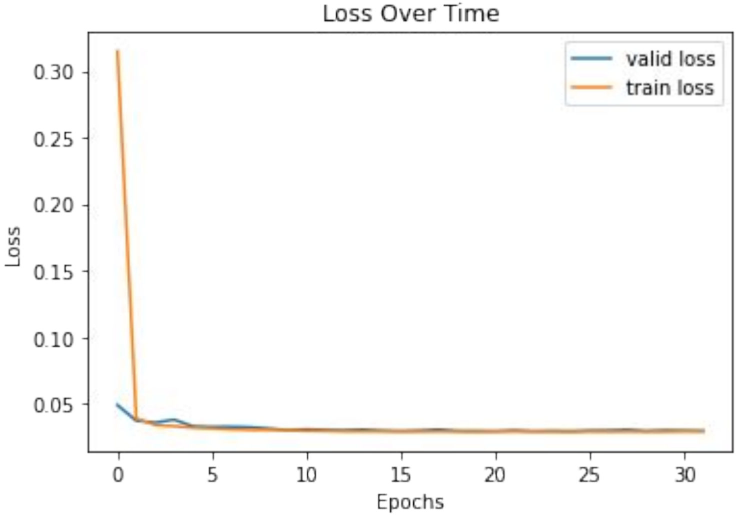


- Encoder: Made up of 4 convolutional layers followed by a fully connected layer
- Decoder: Made up of a fully connected layer followed by 4 DeConv layers.

Results - Language and Vision Models

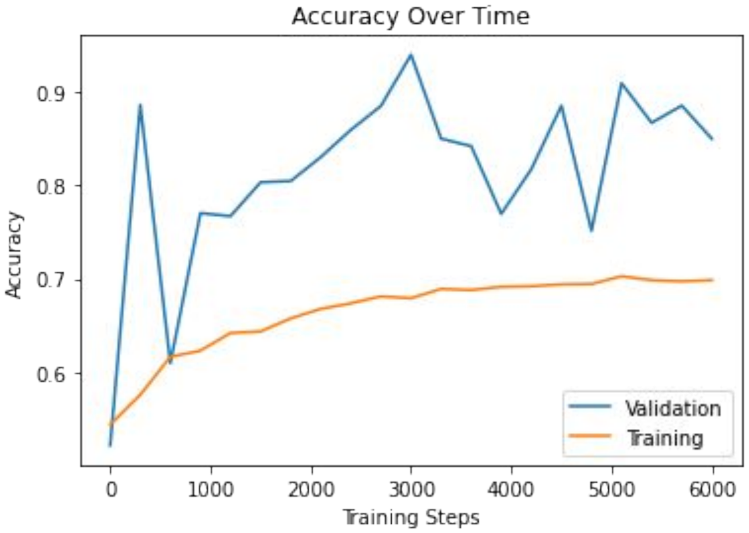


Language Model Loss

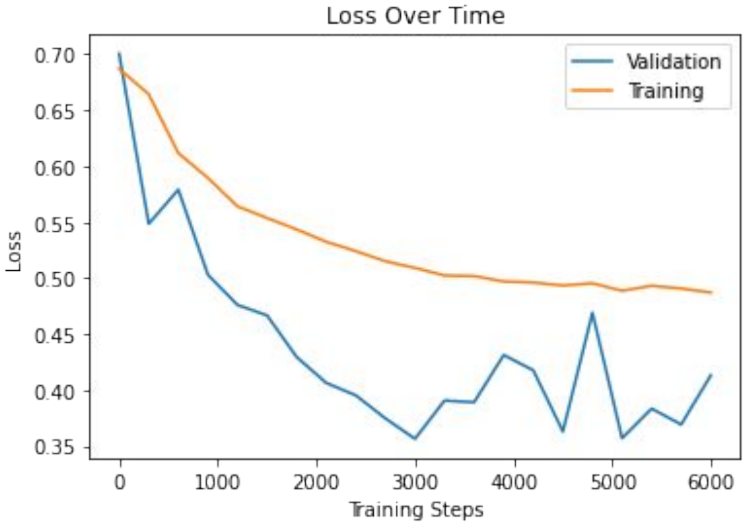


Vision Model Loss

Results - Combined Models



Combined Model Accuracy



Combined Model Loss

Conclusions and Future Work

- **Conclusions**
  - Modular network architecture was successful in using the independently learned language and vision features to effectively address VQA tasks.
  - With our architecture, we were able to achieve a test accuracy of ~78%

- **Future Work**
  - Different pertaining methods could be tested on VQA data.
  - Using TextualQA datasets to train language model may result in better models.
  - For vision model, different pertaining tasks such as GAN could have been used.

References

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, “VQA: visual question answering,” CoRR, vol. abs/1505.00468, 2015. [Online]. Available: http://arxiv.org/abs/1505.00468
[2] Q. Wu, D. Teney, P. Wang, C. Shen, A. Dick, and A. Van Den Hen-gel, “Visual question answering: A survey of methods and datasets,” Computer Vision and Image Understanding, vol. 163, pp. 21–40, 2017.
[3] A. K. Gupta, “Survey of visual question answering: Datasets and techniques,” arXiv preprint arXiv:1705.03865, 2017.
[4] M. Malinowski and M. Fritz, “A multi-world approach to question answering about real-world scenes based on uncertain input,” Advances in neural information processing systems, vol. 27, 2014.
[5] K. Kafle and C. Kanan, “Answer-type prediction for visual question answering,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 4976–4984.
[6] M. Malinowski, M. Rohrbach, and M. Fritz, “Ask your neurons: A neural-based approach to answering questions about images,” in Proceedings of the IEEE international conference on computer vision, 2015,pp. 1–9.
[7] M. Ren, R. Kiros, and R. Zemel, “Image question answering: A visual semantic embedding model and a new dataset,” Proc. Advances in Neural Inf. Process. Syst, vol. 1, no. 2, p. 5, 2015.
[8] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu, “Are you talking to a machine? dataset and methods for multilingual image question,” Advances in neural information processing systems, vol. 28, 2015.
[9] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei, “Visual7w: Grounded question answering in images,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 4995–5004.
[10] K. Chen, J. Wang, L.-C. Chen, H. Gao, W. Xu, and R. Nevatia, “Abc-cnn: An attention based convolutional neural network for visual question answering,” arXiv preprint arXiv:1511.05960, 2015.

References

[11] H. Xu and K. Saenko, “Ask, attend and answer: Exploring question-guided spatial attention for visual question answering,” in European conference on computer vision. Springer, 2016, pp. 451–466.
[12] J. Lu, J. Yang, D. Batra, and D. Parikh, “Hierarchical question-image co-attention for visual question answering,” Advances in neural information processing systems, vol. 29, 2016.
[13] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, “Neural module networks,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 39–48.
[14] C. Xiong, S. Merity, and R. Socher, “Dynamic memory networks for visual and textual question answering,” in International conference on machine learning. PMLR, 2016, pp. 2397–2406.
[15] “Understanding lstm networks,” https://colah.github.io/posts/2015-08-Understanding-LSTMs/, accessed: 2010-09-30.
[16] S. Merity, C. Xiong, J. Bradbury, and R. Socher, “Pointer sentinel mixture models,” 2016.
[17] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays,P. Perona, D. Ramanan, P. Doll’a r, and C. L. Zitnick, “Microsoft COCO: common objects in context,” CoRR, vol. abs/1405.0312, 2014.[Online]. Available: http://arxiv.org/abs/1405.0312
[18] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, “Making the v in vqa matter: Elevating the role of image understanding in visual question answering,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 6904–6913.

Thank you

INDIANA UNIVERSITY BLOOMINGTON