Text Based Information Retrieval
Visual Question Answering
Part 2

Johan Edstedt, Elias Regopoulos

I. Approach

As in the first part we begin by training an LSTM-autoencoder for the questions. This time we used a bidirectional autoencoder. We changed our approach to prune the questions from unnecessary information, e.g. "What is on the table in the image201 ?" -> "What is on the table ?", since this information can not generalize. Once we have trained this autoencoder, we combine the information with another bidirectional LSTM that encodes the CNN features by scanning over the image from top-to-bottom and vice versa. This combined information is then concatenated and passed through two dense layers with selu [1] activations and then to an output layer. In figure 1, the architecture of the network is visualized.

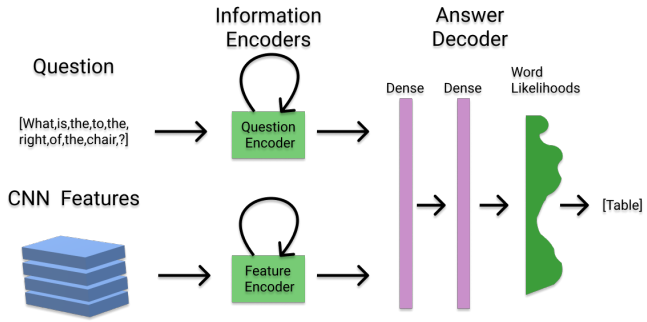

Fig. 1: Overview of our proposed answering architecture

II. Related Work

Our work is based on the original work by Malinkowski et al. [2] but differs in some key aspects. In their model both the question and visual information is fed to an LSTM simultaneously. We found that this gave us a far too complicated model with too many parameters. We instead used seperate LSTMs for the question and CNN features which makes the model much smaller and more feasible to train.

III. Network and training details

We used bidirectional relu LSTMs with a dropout rate of 0.2 for the question and 0.5 for the CNN features, and two fully connected selu [1] hidden layers with 0.5 dropout. For the LSTMs and FC layers, we used 256 units. We used stochastic gradient descent with 0.9 momentum, a learning rate of 0.01, and a batch size of 64.

IV. How to use the model

For instructions on how to use the models, please refer to the provided README.

V. Results

We are able to improve on the original work of Malinowski et al. [2] with about 2% increased accuracy as can be seen in table I. During training we observed that the visual answerer was very good at fitting to the training data, where we had 30% validation accuracy compared to 24% validation accuracy for the language only model. However, this did not generalize well to the test data where we see a more modest improvement of around 1% compared to the language only model, however with much lower test loss (3.88 compared to 4.6 categorical crossentropy). It seems like generalizing learned representations of images to never before seen images for visual question answering is a difficult task.

	Accuracy	WUPS @0.9	WUPS @0.0
Malinowski et al. [2]	19.43	25.28	62.00
Language only*	21.65	28.32	64.04
Visual Answerer*	22.03	29.45	65.69

TABLE I: Single word answering performance. *1-word answers compared to first GT word

REFERENCES

[1] Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-normalizing neural networks. *CoRR*, abs/1706.02515, 2017.

[2] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. Ask your neurons: A neural-based approach to answering questions about images. *CoRR*, abs/1505.01121, 2015.

APPENDIX

A. Question Autoencoder

1) Accurate answer: `what is on the left side of the blinds ?`

2) Inaccurate Answer:

- Question (original):
  `what is to the right corner of the night stand ?`
- Question (reconstructed):
  `what is to the right side of the night stand ?`

B. Question Answerer

1) Accurate answer:

- Question:
  `what is the sink colour ?`
- Answer:
  `white`

2) Inaccurate answer:

- Question (original):
  `what is on the left side of the cup ?`
- Answer (reconstructed):
  `cup`
- Answer (true):
  `tissue roll`

C. Visual Answerer

1) Accurate answer:

- Question:
  `what is the object on the side table ?`
- Answer:
  `lamp`

2) Inaccurate answer:

- Question (original):
  `what is to the right side of the bunk bed ?`
- Answer (reconstructed):
  `cabinet`
- Answer (true):
  `bed`