Region under Discussion in Visual Dialog: Ablations on Input Embeddings to QCS+RuD

Indu Panigrahi and Raymond Liu
{indup, rl27}@princeton.edu

Motivation & Goal

Applications in research: Visual dialog (VD) is a way to evaluate a model's understanding of an image.

Real-world deployment: VD systems can form the base for smart assistants that help visually-impaired people approach visual challenges by answering questions about their surroundings.

Understanding what information is important to a VD system can help improve the reliability of the system. This process involves examining the effect of the inputs on the model. Our project reproduces the Question-Category-Spatial (QCS) model with a Region under Discussion (RuD)[1]. Furthermore, we ablate different portions of the input embedding to evaluate the effect of those portions on the model's performance.

Related Work

VD: A conversation comprised of multiple questions that depend on dialog history (i.e., each question and answer depends on the past questions and answers in the conversation)[2].

GuessWhat?! (GW): Game and dataset used to train and evaluate VD models[3].

QCS Model: Takes an image and a natural language question as input and returns an answer to the posed question[3].

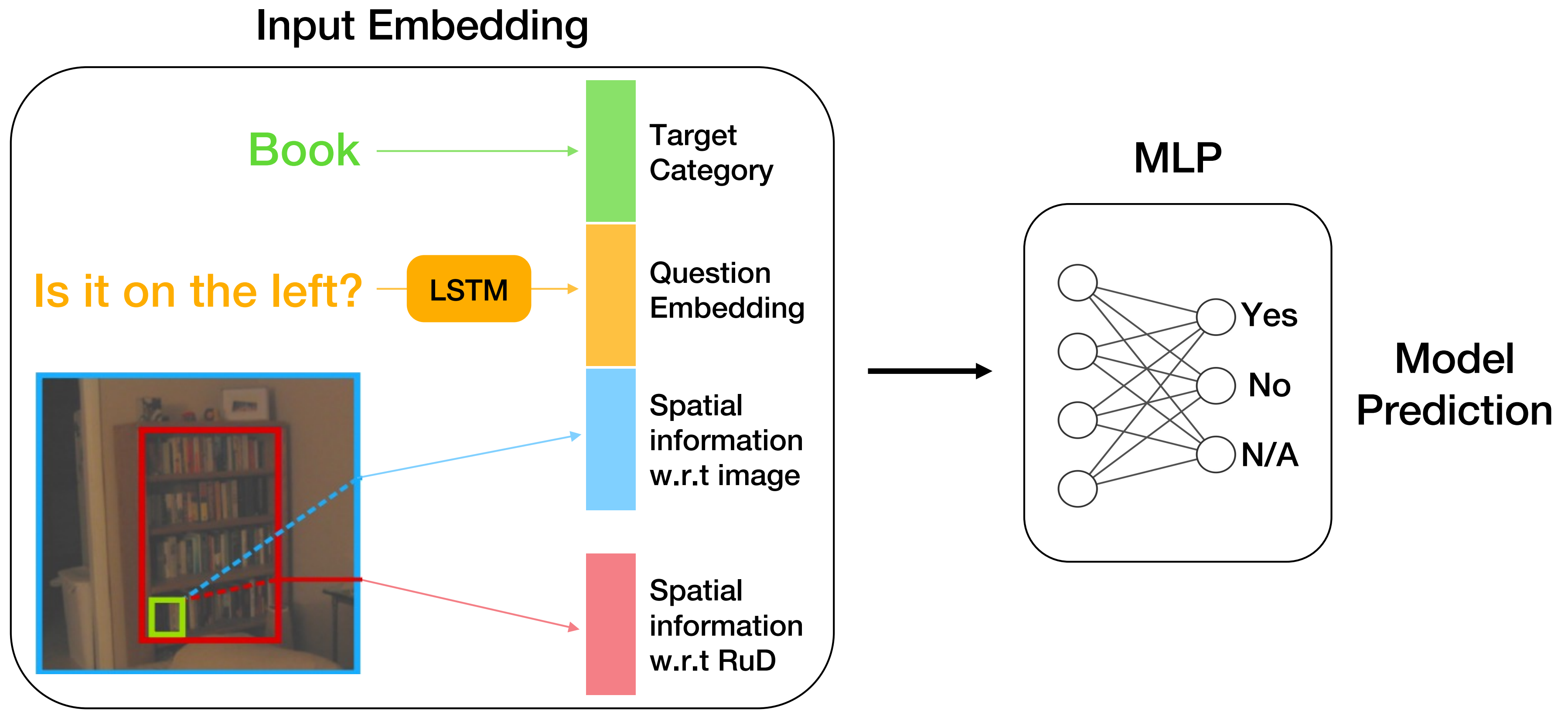
Interpretability of word embeddings: Word intrusion tests[4], modifying the sparsity of input embeddings[5]

References & Acknowledgements

Advised by Professor Karthik Narasimhan for COS 484 *Natural Language Processing*
Special thanks to Mauricio Mazuecos for code clarifications.

[1] Mazuecos et al. 2021. “Region under Discussion for visual dialog". *EMNLP.*
[2] Das et al. 2017. "Visual dialog”. *CVPR*.
[3] de Vries et al. 2017. “Guesswhat?! visual object discovery through multi-modal dialogue”. *CVPR*.
[4] Chang et al. 2009. “Reading tea leaves: How humans interpret topic models”. *NEURIPS*.
[5] Şenel et al. 2018. "Semantic structure and interpretability of word embeddings”. *TASLP*.

QCS+RuD



The QCS model takes an input embedding composed of the target category of the object of interest, an LSTM embedding of the posed question, and spatial information about the object with respect to the entire image. The embedding is fed into a Multilayer Perceptron (MLP), and the model outputs an answer (Yes, No, or N/A).

QCS+RuD model is the same as QCS except it appends spatial information of the object with respect to a Region under Discussion (RuD) to the input embedding. The RuD uses dialog history to zone in on the region of the image that is relevant to the conversation.

Reproduced Paper Results: Baseline & Ablations (all)

Type of Question	Model				
	QCS	QCS+RuD	-super	-2nd	-neg
object	90.80	91.06	90.89	91.23	90.93
spatial	67.72	68.74	69.53	69.44	69.78
color	62.49	62.92	63.33	62.54	63.04
action	64.75	66.20	66.17	65.81	66.66
size	63.34	62.83	62.90	63.71	61.95
texture	72.36	70.59	71.37	71.03	71.59
shape	66.78	69.77	68.77	67.77	67.77
GW Accuracy	78.16	78.69	79.00	78.92	78.98

Ablations on Input Embedding

As the input embedding consists of several parts, we alternately ablate these parts to examine the effect of each on the model. We do not ablate the question embedding as the task is to engage in dialog by answering the questions.

-image refers to ablating the spatial information w.r.t. the image
-target refers to ablating the target category embedding
-RuD refers to ablating the Region under Discussion

Results & Analysis

Results for Ablations on Input Embedding (with history)

	QCS+RuD	-image	-target	-image -RuD	-image -target	-target -RuD
object	91.06	90.78	74.55	90.78	72.39	73.63
spatial	68.74	68.35	68.77	68.35	67.82	66.97
color	62.92	62.97	60.09	62.97	59.85	57.35
action	66.20	65.85	61.99	65.85	61.73	60.14
size	62.83	65.10	66.20	65.1	63.71	63.78
texture	70.59	73.25	63.37	73.25	62.49	61.71
shape	69.77	68.77	63.79	68.77	60.13	63.79

1. **RuD captures a significant amount of image information.**

   The performance of **QCS+RuD** and **–image** are similar. Thus, the RuD seems to provide enough visual information. Furthermore, **–target–RuD** performs worse than **–target**, indicating that the model can eventually deduce the target object from the RuD.

2. **“Target category” seems to help with questions about object type, texture, and shape.**

   All models that include the target category in the input embedding perform significantly better on questions about object type, texture, and shape than models that ablate the target category.

   For example, **–image–RuD** performs significantly better than **–image–target**, indicating that including the target category helps more than including the image spatial for these types of questions.

3. **Sometimes, the ablations remove too much information.**

   Removing information can be detrimental to performance. For example, **–image–RuD** and **–image–target** perform much worse than **QCS+RuD**.

Future Work

A potential improvement could be figuring out how to leverage visual features from computer vision models. The current model relies on COCO annotations to construct the RuD because the QCS paper found that visual features worsened the performance. However, as annotations are not necessarily available for all datasets, computing features would be useful. Additionally, we could try different question embeddings by, for example, using a transformer instead of an LSTM.