A Deep Learning Approach for Identification of Confusion in Unstructured Crowdsourced Annotations

Rachel Gardner, Maya Varma, Clare Zhu

Department of Computer Science, Stanford University

Introduction

- Large datasets are often labeled by paid crowdworkers, which is a lengthy and expensive process
- Pressing need for accurate automated evaluation of crowdsourced annotations
- **Goal:** Perform a binary classification of confusion in crowdsourced data labels and identify the correct answer from unstructured response text.
  - **Visual Question Response (VQR) Task:** Identify confusion given an image, a question referring to the image, and a crowdworker response
  - **Question Response (QR) Task:** Identify confusion based on question and response text with no image features.

Dataset

- 50,628 image-question-response trios, obtained from users on Instagram
  - Questions asked by a bot that analyzes image features
  - Dataset includes ground truth answers
- Generated binary labels to represent confusion, assigning 0 if the user response contains the true answer and 1 otherwise
- Identified spans (index range) in response containing the ground truth answer

Question: ["what", "color", "is", "the", "ball"]

Response: ["it", "looks", "kind", "of", "red"]

Ground Truth Answer: ["red"]

Label: 0

Answer Span: (4,4)

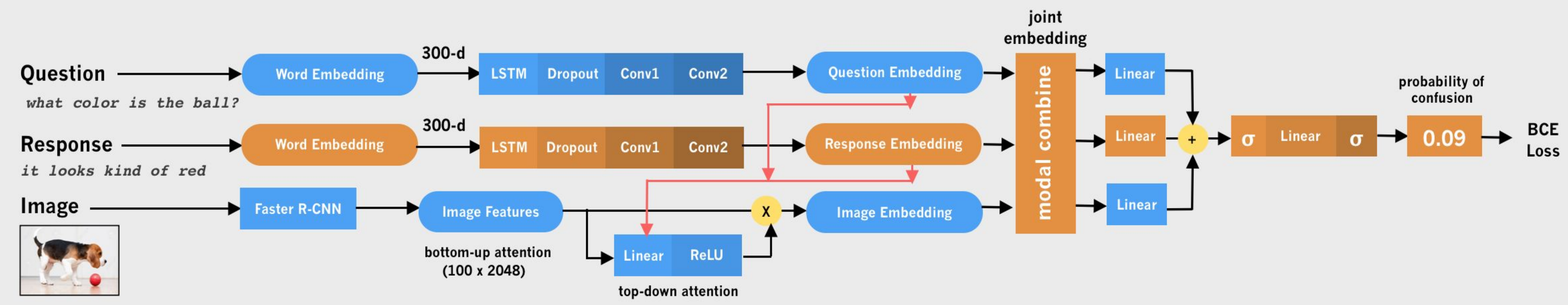
Image:

Dataset: The dataset includes images, questions, responses, ground truth answers, labels, and answer spans.

Approach

Visual Question Response (VQR) Model

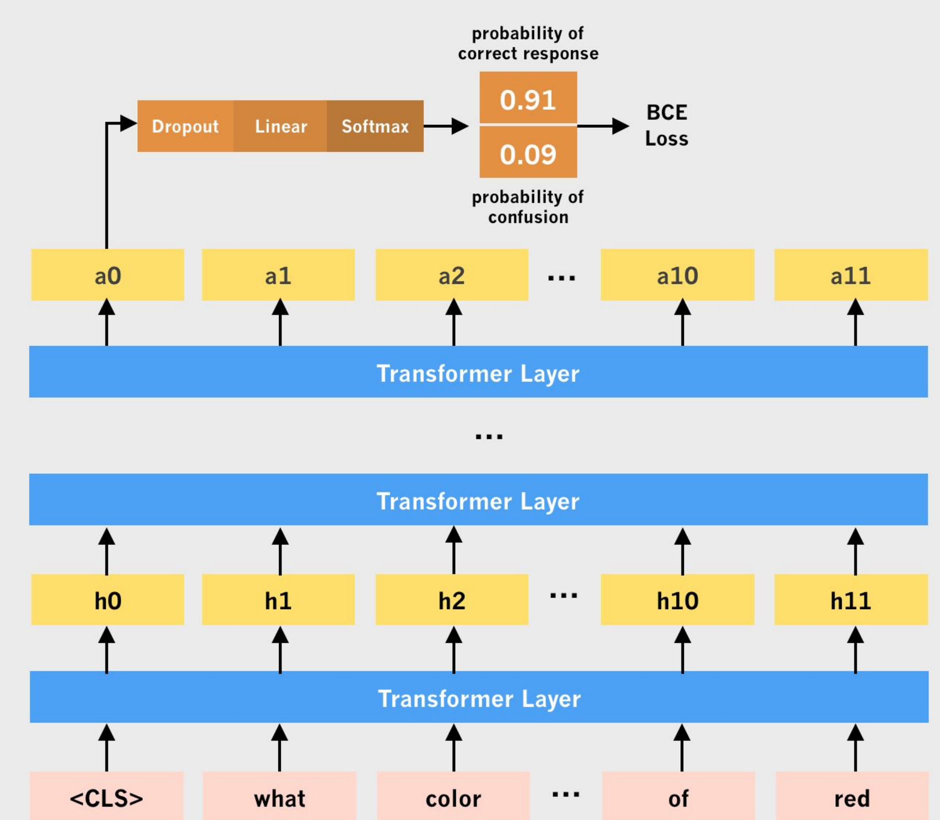
- Pythia is a network created by Facebook AI for the visual question answering task.
  - Standard architecture is not designed for binary classification or answer span detection. The model also works only with formatted text input, not natural language
  - Thus, significant customization was necessary
- A pretrained Faster R-CNN model is utilized to compute bottom-up attention over images. Features are then weighted with respect to the question and user response
- Question, response, and image embeddings are combined with a weighted Hadamard product, creating a joint embedding representing the entire input
- *Binary classification*: Joint embedding is passed through linear layers and sigmoid nonlinearities to generate a probability of confusion, ranging between 0 and 1.
- *Answer prediction:* Joint embedding is passed through two linear layers and softmax functions to identify the start and end index of the answer span within the response.



VQR Model Architecture for Binary Classification

Question Response (QR) Model

- Google AI's pretrained BERT base uncased model serves as an effective starting point for the QR task
- *Binary Classification:* Pooled outputs passed through a dropout layer and a fully connected layer, followed by softmax
- *Answer Prediction:* Encoded hidden states corresponding to the last attention block are passed through a single fully connected layer
  - Two output classes, representing the start and end index of the answer span within the response
- *Multi-Task:* Simultaneously performs binary classification and answer prediction for all examples
  - Prediction loss set to 0 when the model identifies confusion, since no associated span exists



QR Model Architecture for Binary Classification

Results and Analysis

- Binary Classification Task

Test Set Results	AUC-ROC
Baseline: Bag of Words	0.50
Baseline: GLoVe Embeddings	0.74
VQR Binary Classifier	0.79
QR Binary Classifier	0.84

- Answer Prediction Task

Test Set Results	F1 score
VQR Answer Prediction	0.46
QR Answer Prediction	0.77

- Multi-Task Model

Test Set Results	AUC-ROC	F1 score
QR Multi-Task	0.84	0.78

- QR model achieves higher performance than VQR model on both tasks
- Multi-task QR outperforms single-task QR

Conclusion

- VQR and QR models can effectively identify crowdworker confusion and extract answers.
- Multi-task QR model can perform both tasks
- High performance of QR model suggests that analysis of images may not be necessary in resource-constrained settings
- Custom tokenization methods enable effective handling of unstructured input

Selected References

[1] P. Anderson, et al. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on CVPR*, pages 6077–6086, 2018.

[2] Y. Jiang, et al. Pythia v0.1: the winning entry to the VQA challenge 2018. *arXiv:1807.09956*, 2018

[3] J. Devlin, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint *arXiv:1810.04805*, 2018.