Learning Answer Embeddings for Visual Question Answering

Hexiang Hu*, Wei-Lun (Harry) Chao*, Fei Sha
University of Southern California

Motivation:
Different Visual QA datasets has different evaluation criteria. (Open-end vs. Multiple Choice)

Open-end
- VQA 1&2
- Visual7W
- qaVQ

Multiple Choice
- State-of-the-Art: Multi-way Classifiers on a (I, Q, A) triplet. Output the one with highest probability [7, 13, 25]
- Drawback: Sensitive to the bias in the MC dataset [13]

State-of-the-Art:
- Binary classifiers on a (I, Q) triplet. Output the one with highest probability [7, 13, 25]
- Drawback: Cannot handle OOV answers

Research Question: How to excel different settings simultaneously? How to transfer across settings?

Our Approach:
Factorize Visual QA Model as Embedding Learning

Image-Question Embedding
- $g_{\theta}(A)$
- $f_{\theta}(I, Q)$
- $P(q_{i}|s_{i}, q_{n}, q_{o}) \propto \frac{\exp(g_{\theta|(I,Q)}(s_{i}, q_{n})) \exp(g_{\theta}(q_{o}))}{\sum_{n} \exp(g_{\theta|(I,Q)}(s_{i}, q_{n})) \exp(g_{\theta}(q_{o}))}$

Answer Embedding
- ‘The little boy.’
- ‘In the air.’
- ‘The basket.’
- ‘The woman.’
- ‘The player.’
- ‘The man in the white uniform.’

Joint Embedding Space

Q1: Where is the ball?

Most multimodal encoders (e.g., MLP, SAN, MCB) could be used for $f_{\theta}(I, Q)$.
A variety of text (sequence) encoders (e.g., BoW, LSTM) could be used for $g_{\theta}(A)$

Probabilistic Model of Compatibility (PMC)

Inference
- $a_{n} = \arg \max_{a_{n}} f_{\theta}(I, Q)^{T} g_{\theta}(a_{n})$, (8)

Experimental Results:
Performances with Different VQA Datasets

Transfer Learning across VQA Datasets

Settings, Vocab coverage

Transfer Results:
- Our factorized model with PMC outperforms all methods

Research Question:
How to excel different settings simultaneously? How to transfer across settings?

Ablation Study

Negative sampling
Inference efficiency
To learn or not to learn

Visualization of Answer Embeddings

Answers cluster with respect to syntax and semantics

Conclusion:
- Learning answer embeddings improves across multiple datasets.
- Our framework leverages SOTA embeddings of image & text.

Our Contributions:
1. A probabilistic framework with efficient training over large-scale answer vocabulary.
3. Extensive studies on and across multiple Visual QA benchmarks.