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

State-of-the-Art: Multi-way Classifiers on the top-frequent answers [2,4,7,18,28]

Drawback: Can not handle OOV answers

Multiple Choice

Visual7W **qaVG**

State-of-the-Art: Binary 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]

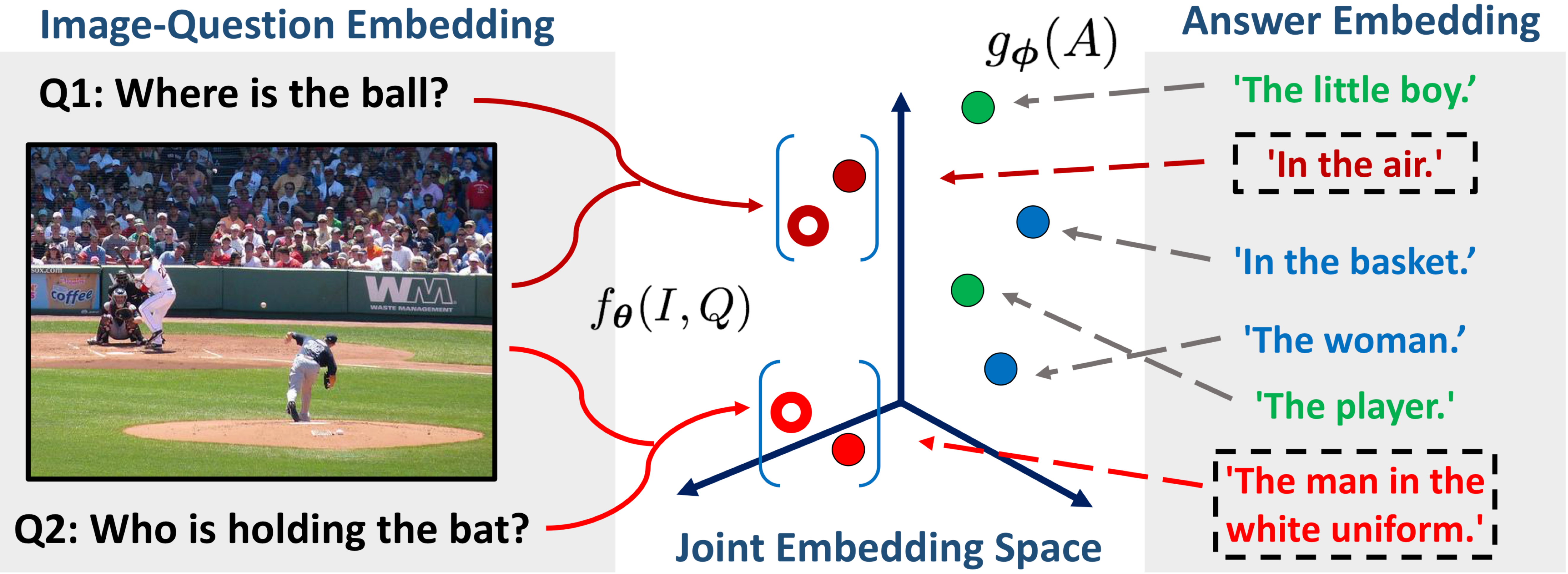
Research Question: **How to excel different settings simultaneously? How to transfer across settings?**

Our Contributions:

1. **A probabilistic framework with efficient training over large-scale answer vocabulary.**
2. **An efficient factorization model that unifies across Visual QA settings and datasets.**
3. **Extensive studies on and across multiple Visual QA benchmarks.**

Our Approach:

Factorize Visual QA Model as Embedding Learning



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 $g_\phi(A)$

Probabilistic Model of Compatibility (PMC)

$$p(a|i_n,q_n) = \frac{\exp(f_\theta(i_n,q_n)^\top g_\phi(a))}{\sum_{a'\in\mathcal{A}} \exp(f_\theta(i_n,q_n)^\top g_\phi(a'))} \quad (1)$$

$$\ell = -\sum_n^N \sum_{a\in\mathcal{C}_n} \sum_{d\in\mathcal{A}} \alpha(a,d) \log P(d|i_n,q_n), \quad (2)$$

Inference →

$$a^* = \arg\max_{a\in\mathcal{A}} f_\theta(i,q)^\top g_\phi(a), \quad (8)$$

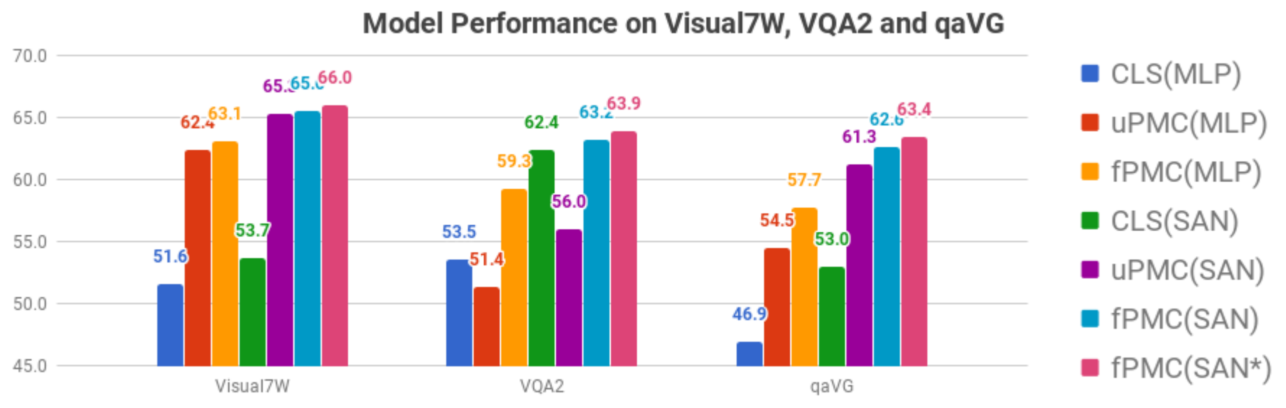
← **Learning**

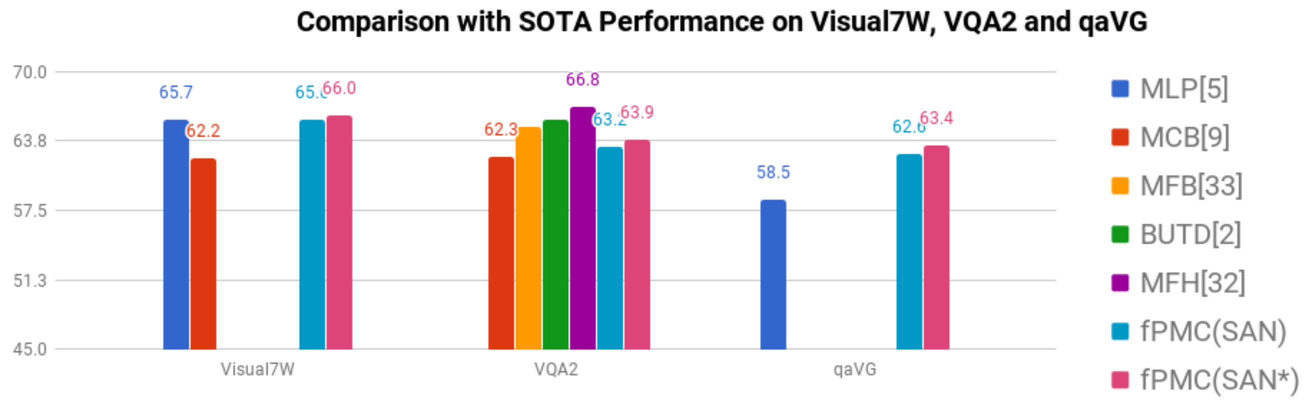
(a) Stochastic negative **sampling** for efficient training

(b) Weighting with $\alpha(a,d)$ to incorporate **semantics**

Experimental Results:

Performances with Different VQA Datasets





Transfer Learning across VQA Datasets

Settings. Vocab coverage

Table 6. The # of common answers across datasets (training set)

Dataset	Top-K most frequent answers 1K	3K	5K	10K	all	Total # of unique answers
VQA2, Visual7W	451	1,262	2,015	3,585	10K	137K
VQA2, qaVG	495	1,328	2,057	3,643	11K	149K
Visual7W, qaVG	657	1,890	3,070	5,683	27K	201K

Transfer Results.

Our **factorized model** with **PMC** outperform all methods

Table 5. Results of cross-dataset transfer using either classification-based models or our models (PMC) for Visual QA. (f_θ = SAN)

	Visual7W				VQA2				qaVG			
	CLS	uPMC	fPMC	fPMC⋆	CLS	uPMC	fPMC	fPMC⋆	CLS	fPMC	fPMC	fPMC⋆
Visual7W	53.7	65.3	65.6	**66.0** ↑	19.1	18.5	**19.8** ↑	19.1	42.8	52.2	**54.8** ↑	54.3
VQA2	45.8	56.8	60.2	**61.7** ↑	59.4	56.0	60.0	**60.9** ↑	37.6	51.5	54.8	**56.8** ↑
qaVG	58.9	66.0	68.4	**69.5** ↑	25.6	23.6	25.8	**26.4** ↑	53.0	61.2	62.6	**63.4** ↑

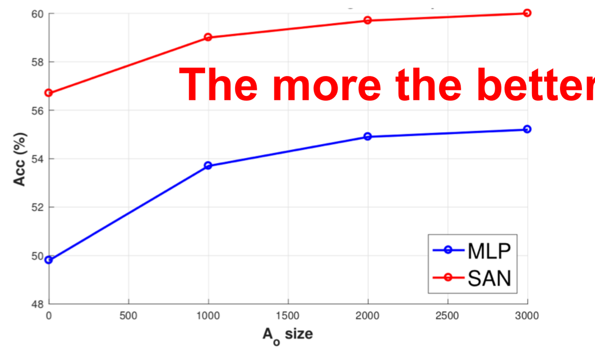
Detailed Results on Seen/Unseen Answers.

Table 2. Analysis of cross dataset performance over Seen/Unseen answers using either CLS or PMC for Visual QA

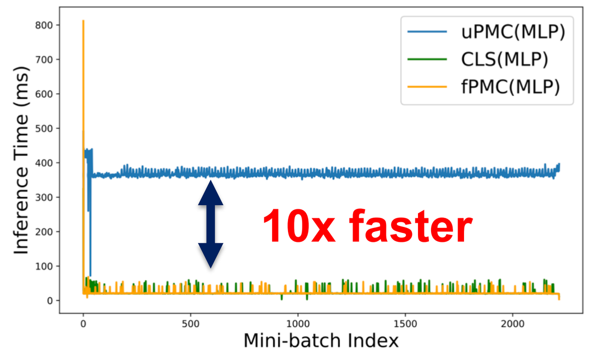
	Visual7W											
	CLS(SAN)			uPMC(SAN)			fPMC(SAN)			fPMC(SAN⋆)		
	S	U	All	S	U	All	S	U	All	S	U	All
VQA2	59.8	25.0	45.8	57.4	54.6	56.8	60.7	58.5	60.2	61.7	59.4	62.5
qaVG	63.4	25.0	58.9	66.7	45.3	66.0	69.1	47.7	68.4	70.2	46.9	69.5

Ablation Study

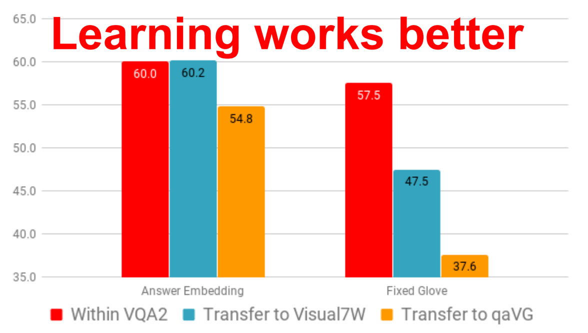
Negative sampling



Inference efficiency

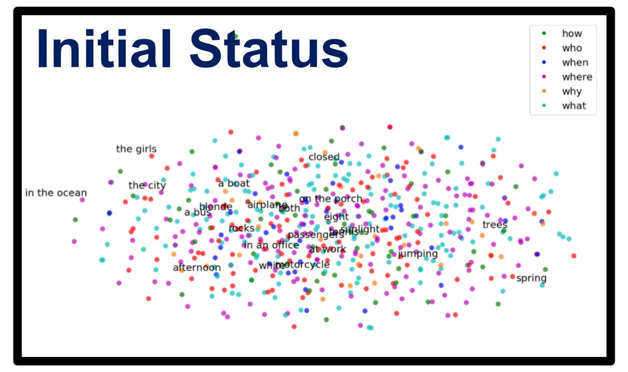


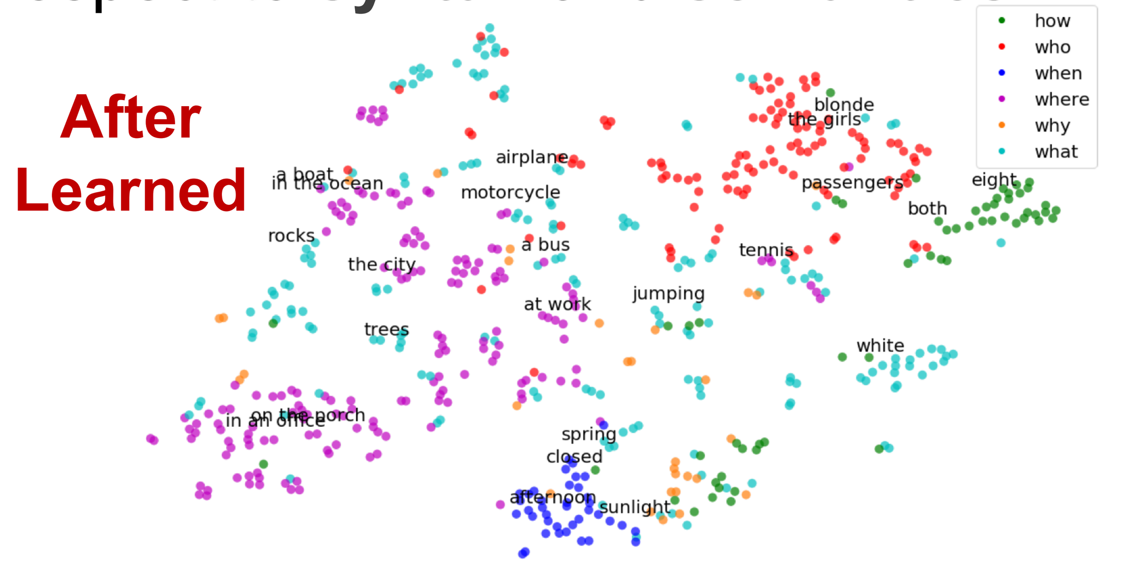
To learn or not to



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.