Visual Question Answering with Various Feature Combinations

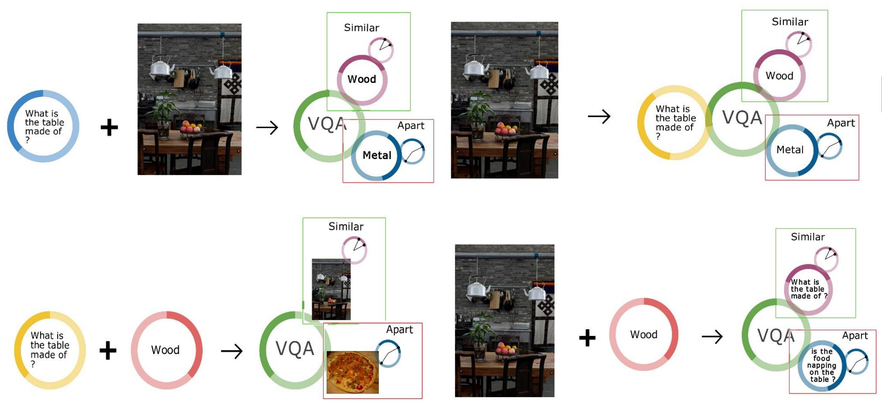
Jinwoo Choi and Siddharth Narayanan

Abstract

In this work, we present a Visual Question Answering task. We measure a similarity between two modalities through a multi-modal embedding and is based on deep multi-modal similarity model (DMSM) proposed by Fang et al [1]. Using DMSM, we conduct experiments with different experimental settings comprising of natural images, questions and answers.

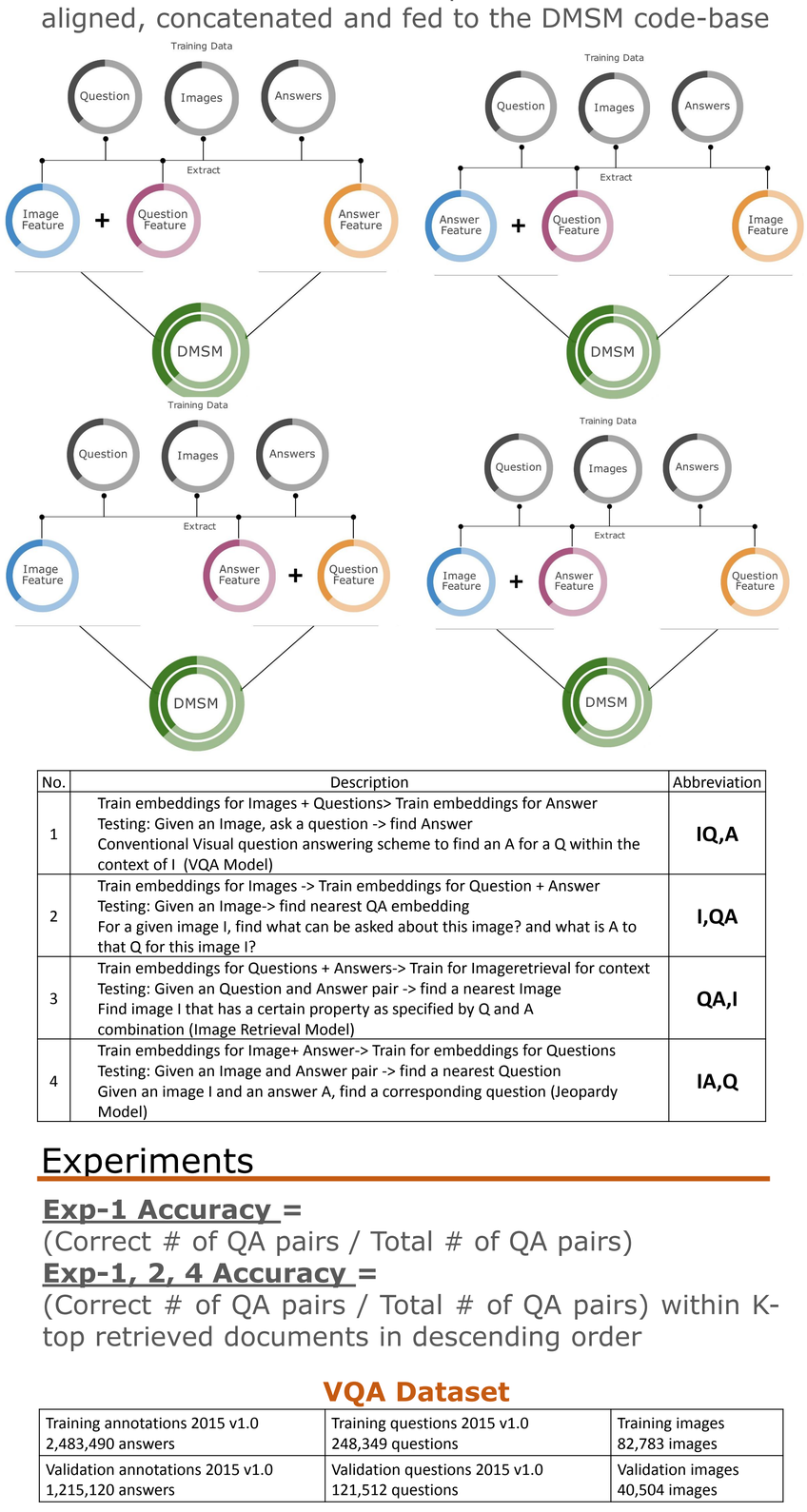
Introduction

The goal of our project is to enable machines to understand various combinations of images along with their corresponding questions and answers and respond to them appropriately. We look at all forms of inputs and outputs as opposed to a conventional DMSM model that learns two neural networks by mapping images and text fragments.



Approach

The feature sets of the required combination are aligned, concatenated and fed to the DMSM code-base



No.	Description	Abbreviation
1	Train embeddings for Images + Questions> Train embeddings for Answer Testing: Given an Image, ask a question -> find Answer Conventional Visual question answering scheme to find an A for a Q within the context of I (VQA Model)	**IQ,A**
2	Train embeddings for Images -> Train embeddings for Question + Answer Testing: Given an Image-> find nearest QA embedding For a given image I, find what can be asked about this image? and what is A to that Q for this image I?	**I,QA**
3	Train embeddings for Questions + Answers-> Train for Imageretrieval for context Testing: Given an Question and Answer pair -> find a nearest Image Find image I that has a certain property as specified by Q and A combination (Image Retrieval Model)	**QA,I**
4	Train embeddings for Image+ Answer-> Train for embeddings for Questions Testing: Given an Image and Answer pair -> find a nearest Question Given an image I and an answer A, find a corresponding question (Jeopardy Model)	**IA,Q**

Experiments

Exp-1 Accuracy =
(Correct # of QA pairs / Total # of QA pairs)
Exp-1, 2, 4 Accuracy =
(Correct # of QA pairs / Total # of QA pairs) within K-top retrieved documents in descending order

VQA Dataset

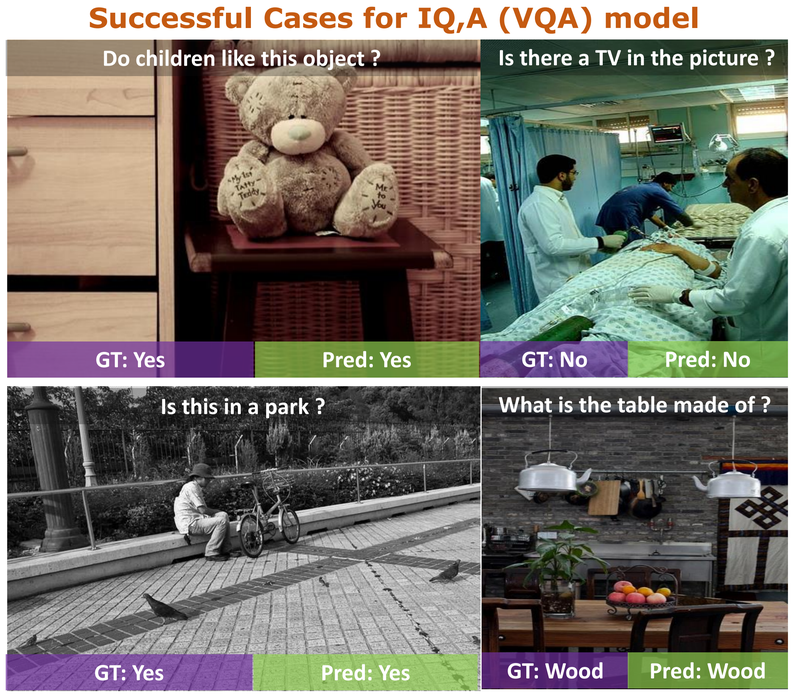
Training annotations 2015 v1.0 2,483,490 answers	Training questions 2015 v1.0 248,349 questions	Training images 82,783 images
Validation annotations 2015 v1.0 1,215,120 answers	Validation questions 2015 v1.0 121,512 questions	Validation images 40,504 images

Training parameter table

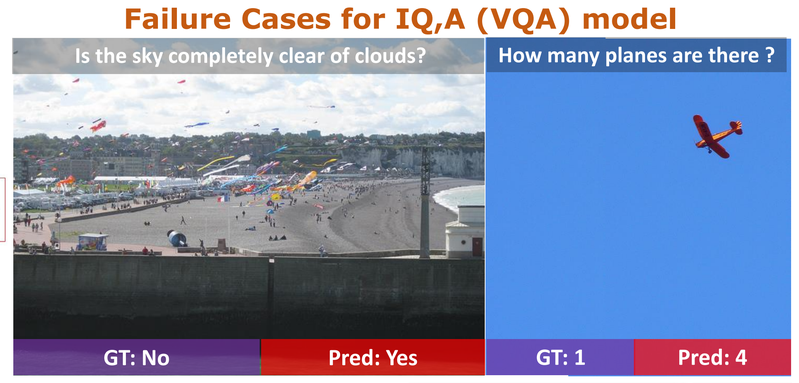
Batchsize: 1024	Target Architecture: 0,0
Maximum number of Iterations: 100	MIRROR_INIT: 0
Learning Rate: 0.02	MATH_LIB: CPU
Source Layer Dimensons: 1000,300	OBJECTIVE: MMI
Source Architectue: 0,0 #0: Fully Connected 1	SOURCE_ACTIVATION: tanh
Target Layer Dimensions: 1000,300	TARGET_ACTIVATION: tanh

Qualitative results

Successful Cases for IQ,A (VQA) model



Failure Cases for IQ,A (VQA) model



QA pair Retrieval Results for I,QA (QA Retrieval)

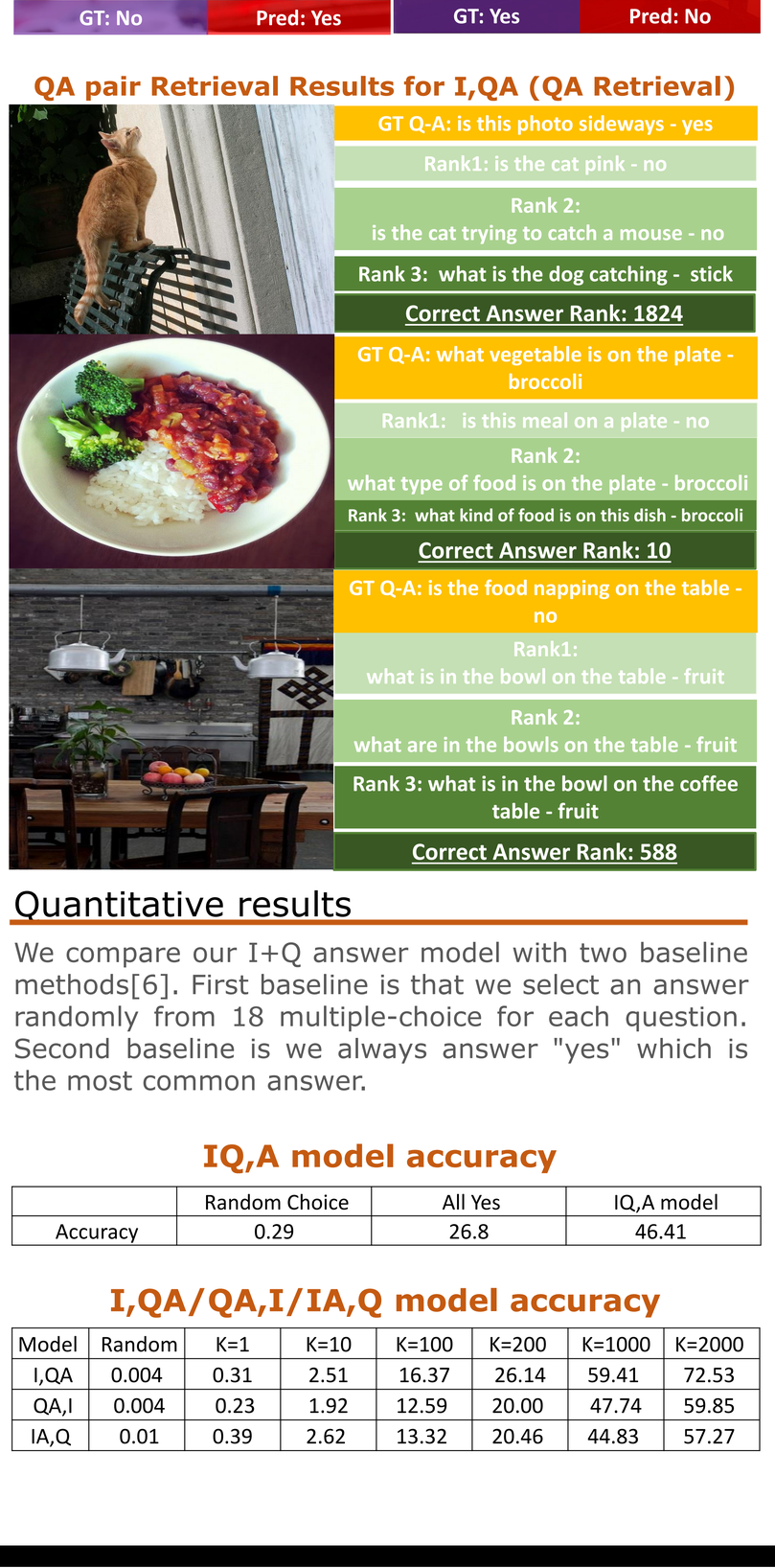
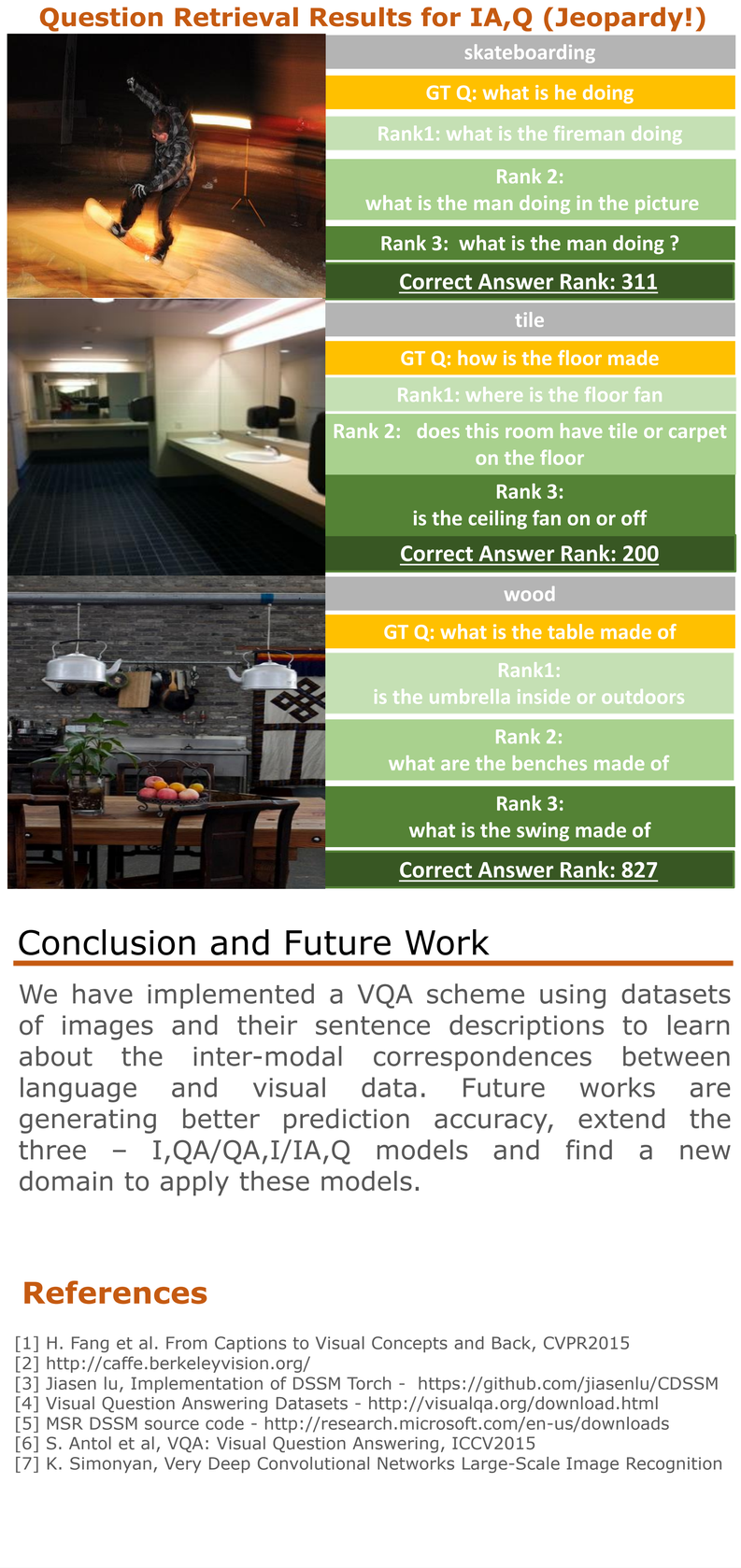


Image Retrieval for QA,I (Image Retrieval)

Question Retrieval Results for IA,Q (Jeopardy!)



Quantitative results

We compare our I+Q answer model with two baseline methods[6]. First baseline is that we select an answer randomly from 18 multiple-choice for each question. Second baseline is we always answer "yes" which is the most common answer.

IQ,A model accuracy

	Random Choice	All Yes	IQ,A model
Accuracy	0.29	26.8	46.41

I,QA/QA,I/IA,Q model accuracy

Model	Random	K=1	K=10	K=100	K=200	K=1000	K=2000
I,QA	0.004	0.31	2.51	16.37	26.14	59.41	72.53
QA,I	0.004	0.23	1.92	12.59	20.00	47.74	59.85
IA,Q	0.01	0.39	2.62	13.32	20.46	44.83	57.27

Conclusion and Future Work

We have implemented a VQA scheme using datasets of images and their sentence descriptions to learn about the inter-modal correspondences between language and visual data. Future works are generating better prediction accuracy, extend the three – I,QA/QA,I/IA,Q models and find a new domain to apply these models.

References

[1] H. Fang et al. From Captions to Visual Concepts and Back, CVPR2015
[2] http://caffe.berkeleyvision.org/
[3] Jiasen lu, Implementation of DSSM Torch - https://github.com/jiasenlu/CDSSM
[4] Visual Question Answering Datasets - http://visualqa.org/download.html
[5] MSR DSSM source code - http://research.microsoft.com/en-us/downloads
[6] S. Antol et al, VQA: Visual Question Answering, ICCV2015
[7] K. Simonyan, Very Deep Convolutional Networks Large-Scale Image Recognition