COMPUTER SCIENCE
& ENGINEERING
UNIVERSITY OF MICHIGAN

DiagNet: Bridging Text and Image

Xinyi Zheng[1] Shengyi Qian[1] Yi Wen[1] Mark Jin[1] Xiaoyang Shen[1]

[1] *Computer Science and Engineering, College of Engineering*

EECS 498-012, W19
Deep Learning
Project ID: 1

Introduction

Visual **Q**uestion **A**nswering (**VQA**) is to answer open-ended natural language questions related to images. Modern VQA tasks require reading and reasoning of both images and texts. We propose **DiagNet**, a model that could effectively combine multiple evidence of texts and images. Experimental results show that **DiagNet** is competitive in multiple VQA datasets. In TextVQA, we also empower **DiagNet** with a novel multi-task training strategy, combining question type evidence in a hybrid fusion. Empirical results show that this strategy helps **DiagNet** further improve performance in TextVQA.

VQA/TextVQA Datasets

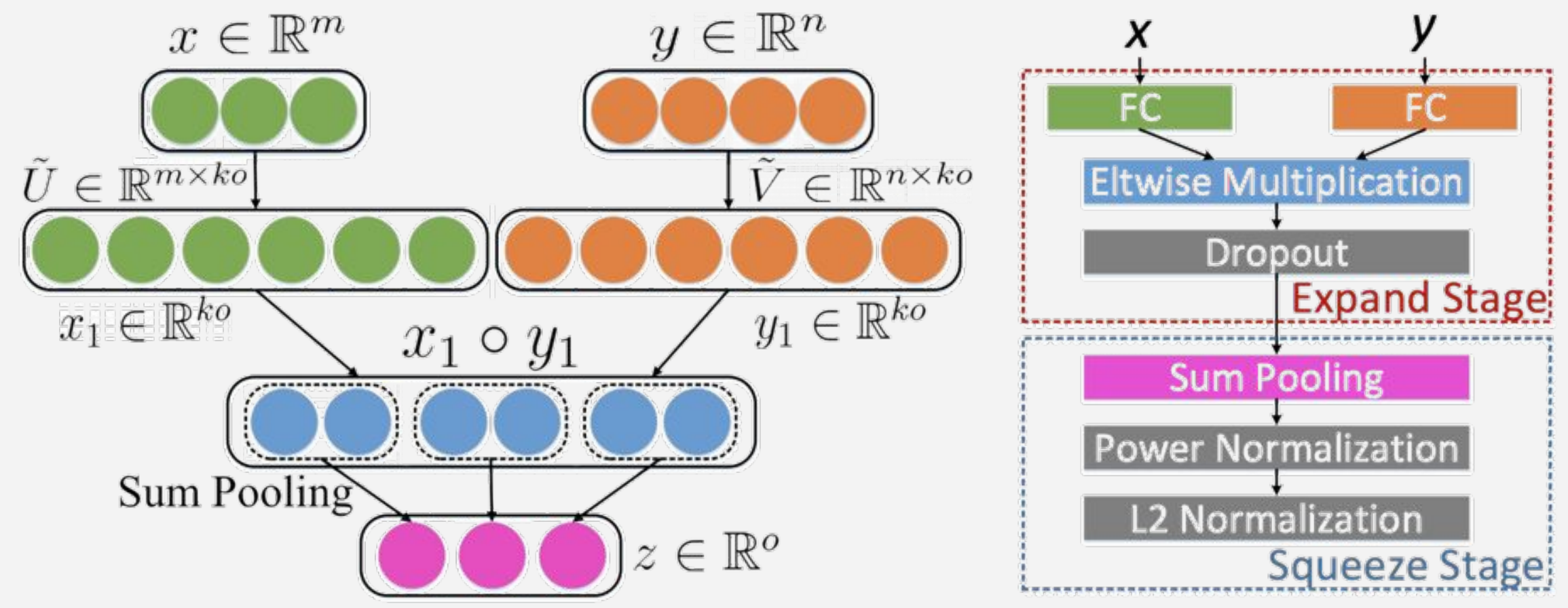
What is the mustache made of?

What is the license plate?

VQA: 200,000 images from MS-COCO. 3 questions with 10 answers. Yes/No, Numeric, Other Questions

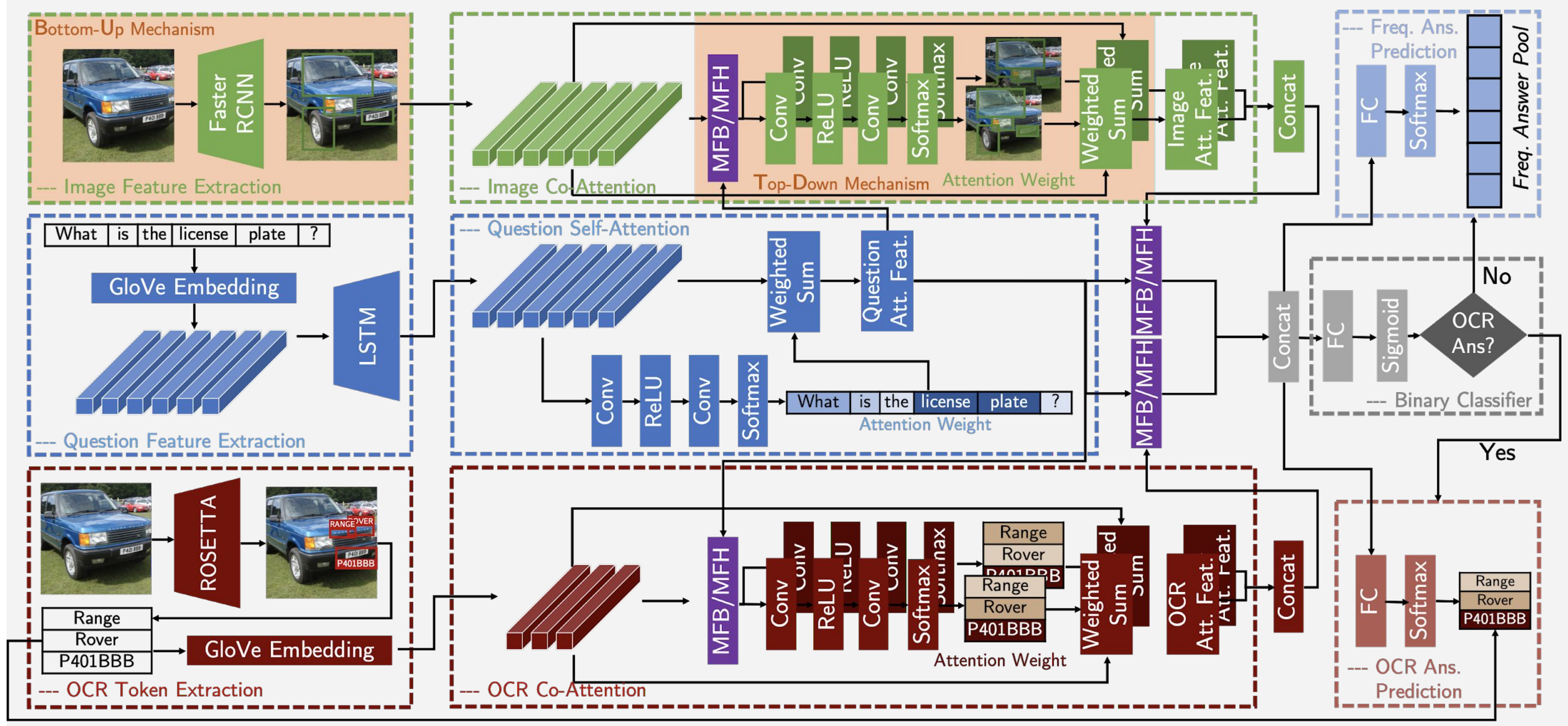
TextVQA: 28,000 images from Open Images. 1 question with 10 answers. Additional Extracted OCR tokens. ~⅓ answers come from OCR.

MFB & MFH



Multi-modal Factorized Bilinear Pooling (**MFB**) and Generalized Multi-modal Factorized High-order Pooling (**MFH**) integrate information from different modals in a high-dim space. MFH, used by default for DiagNet, cascades MFB blocks and achieve higher-order fusion.

DiagNet - Architecture



Faster RCNN is used for image feature extraction. GloVe embedding and LSTM is used for question feature extraction. ROSETTA is used for OCR tokens extraction. The question vector will perform self-attention while the OCR tokens and image features will perform co-attention with question. The concatenated attended vector will go through a binary classifier to determine the source of answer (frequent answer pool or OCR tokens).

Results - VQA

VQA

Model	Dataset	Accuracy
DiagNet without OCR	VQA v1.0 Val	65.01%
HieCoAttenVQA	VQA v1.0 Val	57.00%
MFH (replicated)	VQA v1.0 Val	56.47%
MFB (replicated)	VQA v1.0 Val	55.97%

State-of-the-art Methods on VQA v2.0 for Reference

Model	Dataset	Accuracy
ACRV-MSR	VQA v2.0 Test-standard	69.00%
MFH	VQA v2.0 Test-standard	68.16%
BUTD	VQA v2.0 Test-standard	70.34%
Pythia (I+Q)	VQA v2.0 Test-standard	70.01%

What is the boy doing?

surfing	surfing
Ground Truth	Prediction

How many donuts have m&ms on them?

1	1
Ground Truth	Prediction

What is flying overhead?

plane	clouds
Ground Truth	Prediction

Results - TextVQA

TextVQA (* indicates work in progress)

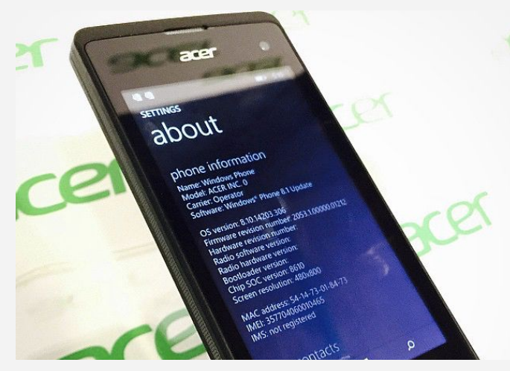
Model	Dataset	Accuracy
DiagNet without OCR	TextVQA v0.5 Val	11.25%
Pythia (I+Q)	TextVQA v0.5 Val	13.04%
DiagNet-OCR	TextVQA v0.5 Val	18.44%*
DiagNet	TextVQA v0.5 Val	*In Progress*
LoRRA + Pythia	TextVQA v0.5 Val	26.56%

What year are the photographs from?

1895	1895
Ground Truth	Prediction

What is this top word on this man's shirt?

fun	fun
Ground Truth	Prediction



What manufacturer made that phone?

acer	acer
Ground Truth	Prediction

What does the blue word on the plane say?

cargo	airlines
Ground Truth	Prediction

What brand of phones are shown?

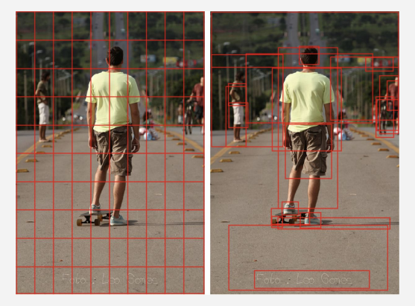
apple	ipad
Ground Truth	Prediction

What state is the plate from?

alaska	1946
Ground Truth	Prediction

BUTD & Attention

BUTD: Bottom Up Top Down [1]



Bottom **U**p: Propose Image Segmentation with Faster RCNN

Top **D**own: Downstream Co-Attention

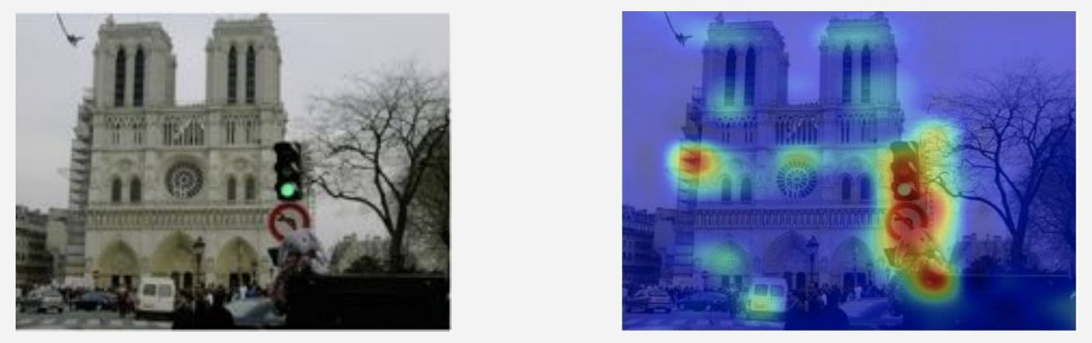
Self Attention

What color on the stop light is lit up?

Self-attend to each word of question

Co Attention[2]

What color on the stop light is lit up?



Attend to regions of image with question

Conclusions & Contributions

Our main contributions can be summarized as

- ❏ We combined insights by several VQA models and build our VQA base model (**DiagNet** without OCR). The performance is comparable to the state-of-the-art method on the VQA dataset.
- ❏ We extended VQA base model to **DiagNet** by adding a text detection branch and effectively combining features of both text and images.
- ❏ We proposed a multi-task training strategy on TextVQA, combining question type evidence in a hybrid fusion.

Reference

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, Lei Zhang, *Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering*

[2] Jiasen Lu, Jianwei Yang, Dhruv Batra, Devi Parikh, *Hierarchical Question-Image Co-Attention for Visual Question Answering*

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, & Devi Parikh. *VQA: Visual Question Answering.* In International Conference on Computer Vision (ICCV), 2015

[4] Zhou Yu, Jun Yu, Jianping Fan, & Dacheng Tao. *Multi-modal factorized bilinear pooling with co-attention learning for visual question answering.* CoRR, abs/1708.01471, 2017.

[5] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, Marcus Rohrbach. *Towards VQA Models that Can Read*. In CVPR 2019