Setting the Scene: Understanding Text in Images for VQA

Sweta Kotha, Meredith Riggs, Naoki Tsuda, & Sean Zhang
(Team 5)

Research Problem & Dataset

Scene Text Visual Question Answering

Q: What is the price of the bananas per kg?

A: $11.98

Q: What is the exit number on the street sign?

A: 2

A: Exit 2

Q: Where is this train going?

A: To New York

A: New York

Images from Reference 1

Datasets

TextVQA (2019)

Training and Validation Images: 25,119

Training and Validation Questions: 39,602

Test Images: 3,353

Test Questions: 5,734

Image Sources: OpenImages (v3)

ST-VQA (2020)

Training and Validation Images: 19,027

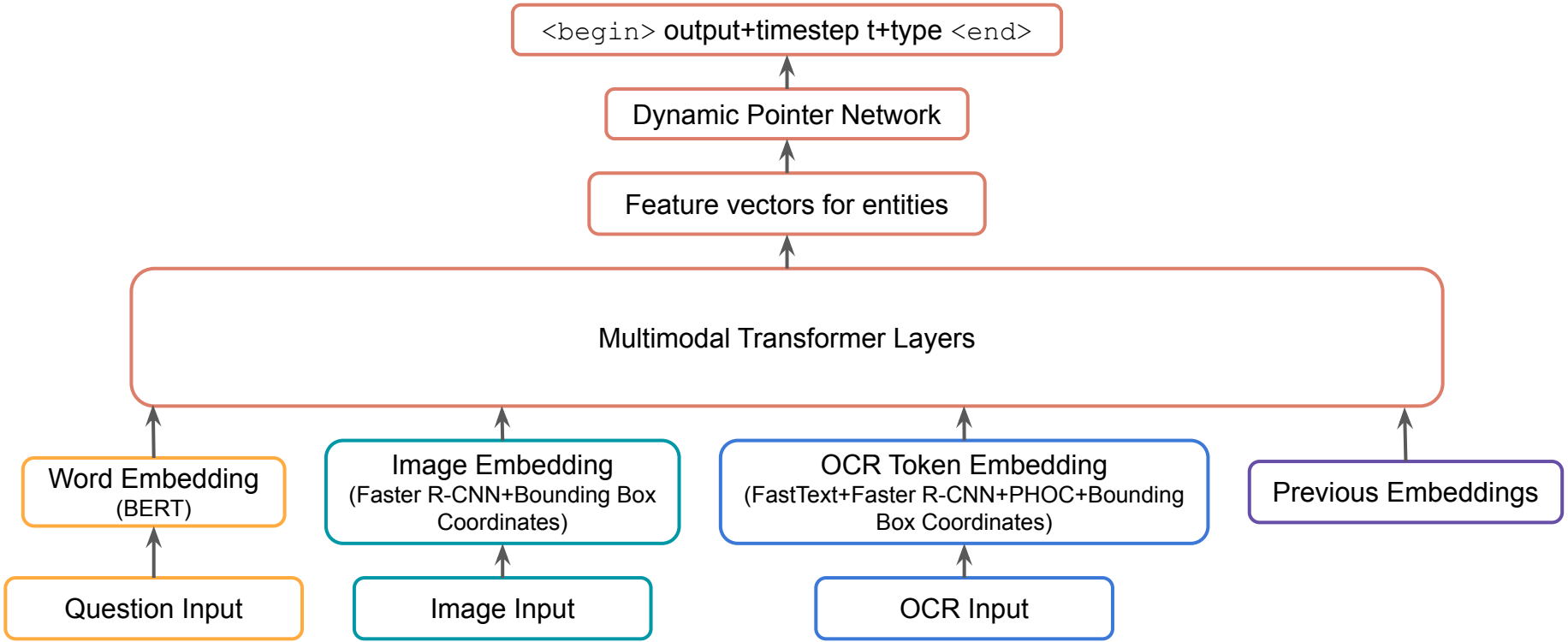
Training and Validation Questions: 26,308

Test Images: 2,993

Test Questions: 4,163

Image Sources: ICDAR2013, ICDAR2015, ImageNet, VizWiz, IIIT Scene Text Retrieval, Visual Genome, COCO-Text

Multimodal Multi-Copy Mesh Model (M4C)



Approaches

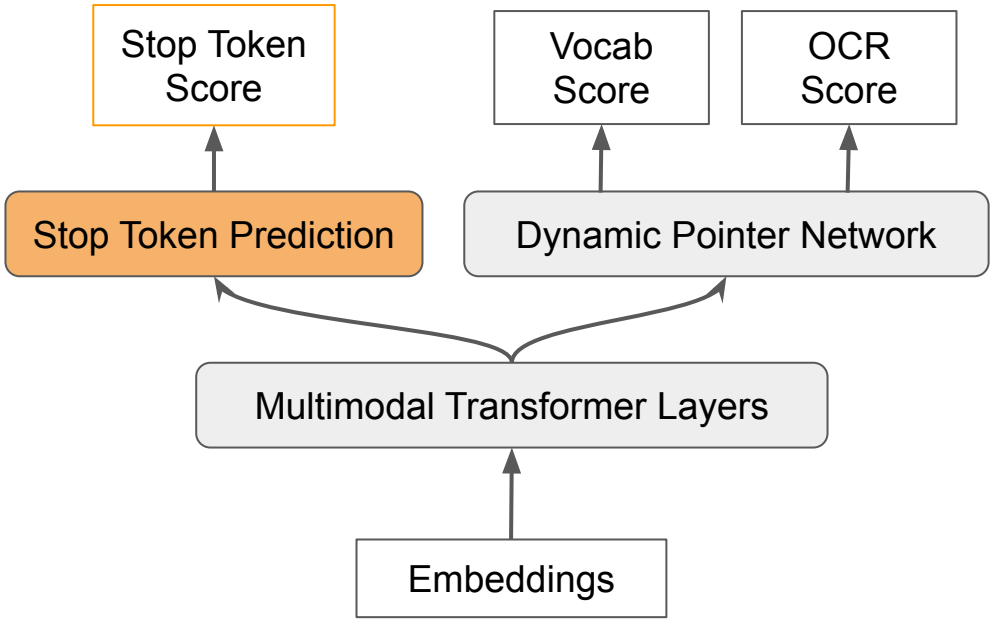
1. Stop Token Prediction
2. Hierarchical Transformer with Global Multimodal Transformer
3. Additional pre-training and fine-tuning on larger datasets

Stop Token Prediction

Linear layer + BCE

Input: Multimodal decoder representation

Output: Stop token score



Stop Token Prediction

- Overrides the dynamic pointer network when the stop token is predicted
- Positive scores are weighted to account for imbalance of samples

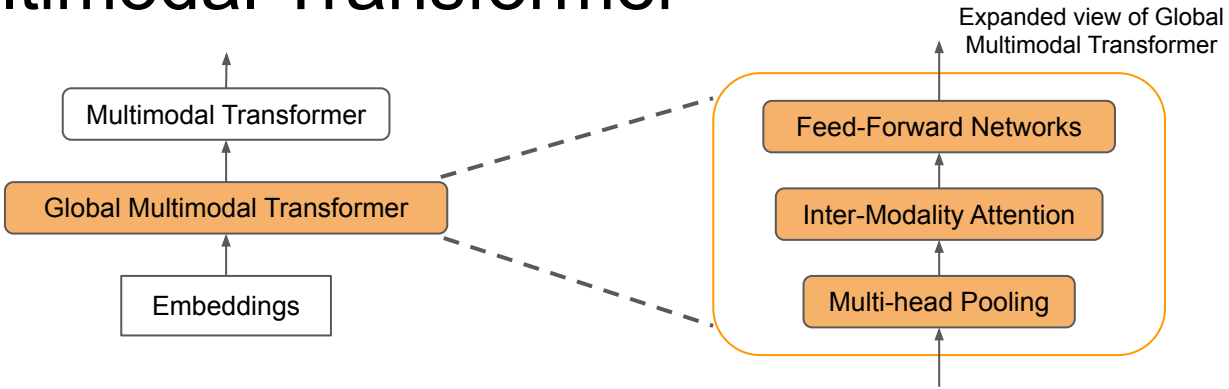
Stop Token Prediction Linear layer:

$$y_{stop} = W_{stop}\, x_{decoder} + b_{stop}$$

Loss function with Stop Token Prediction BCE:

$$Loss_{total} = Loss_{dynamic\ pointer} + Loss_{stop\ token\ prediction}$$

Global Multimodal Transformer[1]



Multi-Head Pooling

Creates fixed-length representation embedding with respect to weight distributions of objects in representations

Inter-Modality Attention

Each representations attend to other representations

Feed-Forward Networks

Context vectors from Inter-Modality Attention layer are fused with the embedding

[1]Liu, Y. and Lapata, M. Hierarchical transformers for multi-document summarization, 2019.

Global Multimodal Transformer

Multi-Head Pooling

For each head, compute the score and the value of the inputs to obtain the probability distribution:

$$\text{score}_{ij} = W_{\text{score}} x_{ij}$$

$$\text{value}_{ij} = W_{\text{value}} x_{ij}$$

$$\text{attention}_{ij} = \text{softmax}(\text{score}_{ij})$$

To compute the head vector for the representation, linearly transform and normalize the attention and the values:

$$\text{head}_i = \text{LN}(W_{\text{head}} \sum_j \text{attention}_{ij} \text{values}_{ij}))$$

Global Multimodal Transformer

Inter-Modality Attention

Pass the head vector through self attention:

$$query_i = W_{query} head_i$$

$$key_i = W_{key} head_i$$

$$value_i = W_{value} x_{ij}$$

$$context_i = \sum_i softmax(query_i key_i)$$

Feed-Forward Networks

Sum the context vector from each head, and linearly transform the contexts:

$$context_i = W_{heads} context_i$$

Then, combine the summed context vectors with the input embeddings and normalize:

$$g_{ij} = W_{heads}(context_i + x_{ij})$$

$$x_{ij} = LN(g_{ij} + x_{ij})$$

Experimental Setup

Global Multimodal Transformer Layers: 1

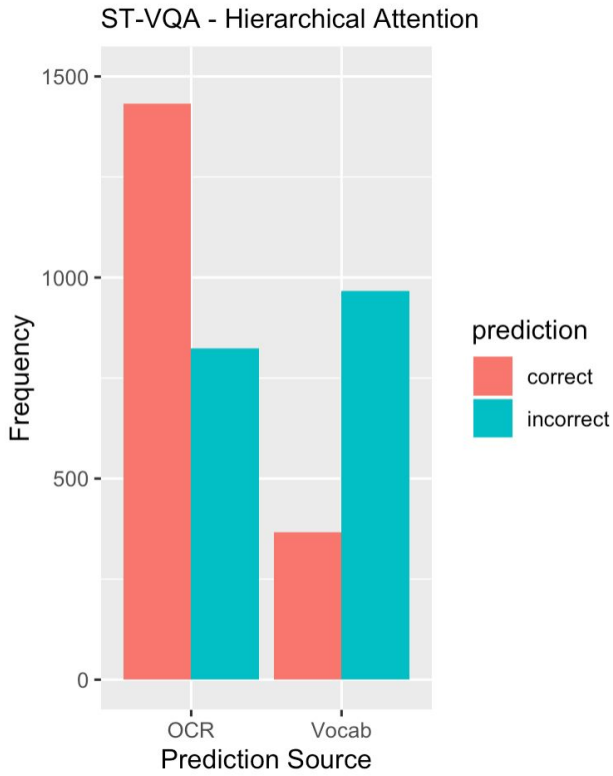
Multimodal Transformer Layers: 3

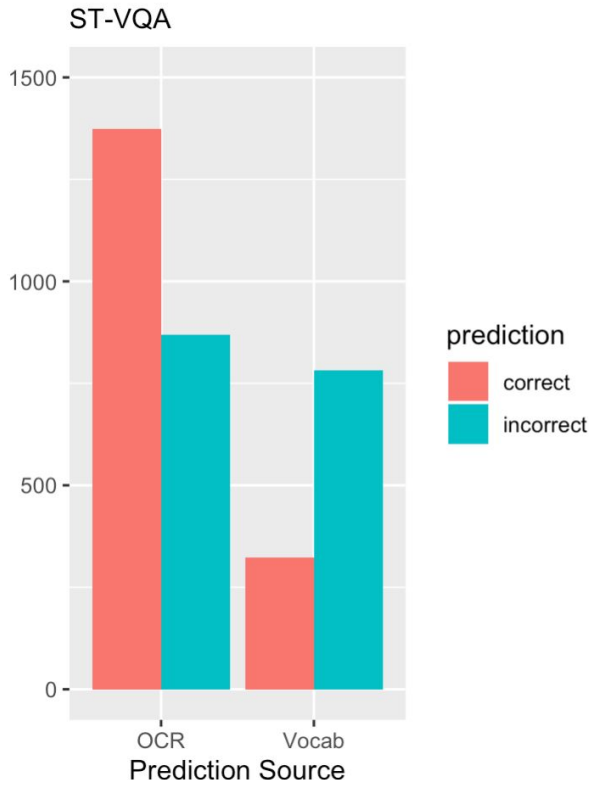
Batch size: 60~80

Learning rate: 1e-4 with warmup during the first epoch

Hierarchical Transformer

0.01 percent increase in accuracy and 0.01 increase in average ANLS





What are the visible letters on the front of the jersey?

M4C: la

M4C with Hierarchical Transformer: geils

Hierarchical Transformer

(Based on Nature of Tokens in Predicted Answers)

Previously...

- Correct: identifying colors, reading signs or labels
- Incorrect: counting objects

Now

- Improvements in counting objects or recognizing numbers
- Same mistakes on identifying colors (usually due to spatial understanding)

Question: What color is the rooster?
Answer: Black
Predicted: Red

References

[1] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusiñol, Ernest Valveny, C.V. Jawahar, Dimosthenis Karatzas, "Scene Text Visual Question Answering", ICCV 2019, 2019

[2] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, Marcus Rohrbach, "Towards VQA Models that can Read", CVPR 2019, 2019

[3] Lluís Gómez, Ali Furkan Biten, Rubèn Tito, Andrés Mafla, Marçal Rusiñol, Ernest Valveny, Dimosthenis Karatzas, "Multimodal grid features and cell pointers for Scene Text Visual Question Answering", 2020

References

[4] Yu Jiang, Vivek Natarajan, Xinlei Chen, Marcus Rohrbach, Dhruv Batra, Devi Parikh, “Pythia v0.1: the Winning Entry to the VQA Challenge 2018”, 2018