Visual Natural Language Query Auto-Completion for Estimating Instance Probabilities

Samuel Sharpe, Jin Yan, Fan Wu, Iddo Drori

COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

CVPR LONG BEACH CALIFORNIA June 16-20, 2019

Introduction

Objectives of our new vision-language task:

1. Auto-complete natural language queries conditioned on image as context.
2. Estimate probabilities of instances conditioned on a natural language query ***independent*** of bounding box, segmentation, attention mechanism
3. Use 1 and 2 to select instances from a pre-segmented image.

Context Based Query Auto-Completion

- Query Auto-completion (QAC) recently based on neural language models with word or character embeddings [lastnam 2018].
- Use FactorCell LSTM [lastname 2018], utilize CNN features as context.

Traditional LSTM

Appending context c to character embedding w and hidden state h equivalent to augmenting bias.

$$h_t = \sigma([w_t, h_{t-1}, c]\mathbf{W} + b)$$
$$= \sigma([w_t, h_{t-1}]\mathbf{W} + \mathbf{V}c + b)$$
$$= \sigma([w_t, h_{t-1}]\mathbf{W} + b')$$

FactorCell LSTM

Add context dependent weight matrix A.

$$\mathbf{Z_R} \in \mathbb{R}^{r \times h \times m} \quad \mathbf{Z_L} \in \mathbb{R}^{m \times (e+h) \times r}$$
$$\mathbf{A} = (c \times_1 \mathbf{Z_L})(\mathbf{Z_R} \times_3 c)$$
$$\mathbf{W}' = \mathbf{W} + \mathbf{A}$$
$$h_t = \sigma([w_t, h_{t-1}]\mathbf{W}' + b)$$

Estimating Instance Probabilities

- Input completed query into fine-tuned BERT and output probabilities for each object class appearing in region using a sigmoid cross entropy loss.

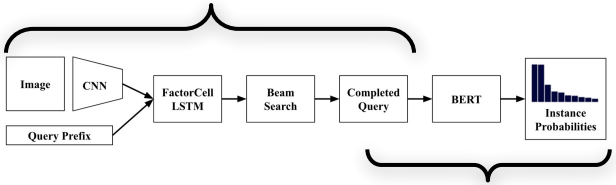
Datasets

- Subset of Visual Genome, and ReferIt datasets.
- Training query to instance probabilities network: use region descriptions in place of queries and associated objects as ground truth.

Dataset	Avg. Characters per Query
ReferIt	16.9 (SD = 12.3)
Visual Genome	26.2 (SD = 8.5)

Network Architecture

Image Query Auto-Completion Network



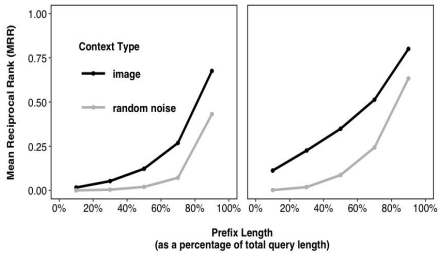
Results

Comparison of image query auto-completion perplexity using **image vs. noise** for each dataset. As expected image context results in a lower (better) perplexity.

Image Query Auto-Completion Perplexity

Context Type	Visual Genome	ReferIt
Corresponding image	2.38	2.66
Random noise	3.84	3.49

Image Query Auto-Completion MRR



Comparison of image query auto-completion MRR results for VG (left) and ReferIt (right) using the image vs. noise. Horizontal axis denotes varying prefix lengths as percentage of total query length and context. MRR improves when increasing query prefix length, and is better when using images.

Instance Probability Network Results: 0.7618 F1-Score on 2,909 classes

Example Results

Example results: (a) input query prefix and image; (b) estimated instance probabilities; (c) instance segmentation; (d) resulting selected instances and auto-completed query conditioned on query prefix and image.

Scan QR Code for Paper

Scan QR Code for GitHub Repo