Multi-Modal common semantic space for Image-Phrase Grounding

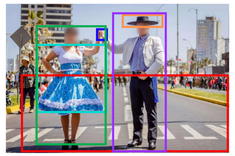
Sonam Goenka, Hassan Akbari

Advanced Computer Vision Course, Columbia University

INTRODUCTION

Objective: Learn a common semantic space shared between textual and visual modalities using unsupervised learning and use it to solve various problems such as Image Phrase Grounding and Visual Question Answering.

Image Phrase Grounding is the task of localizing within an image a given natural language input phrase.



A young lady in blue skirt and a man with a black hat are holding hands in the middle of a road.

Visual Question Answering is the task of answering natural language questions about an image.

- What sport is being played?
- Are there any humans?
- How many players are in the image?
- Is it raining?

APPROACH

We started with the base model developed by Akbari et al. which uses Multi-Modal attention mechanisms to learn a common semantic space using unsupervised learning. We tried to understand, improve and extend the model in the following ways

1. **Analyze relative contribution of CNN layers:** We analyzed which layers of the CNN (and their corresponding feature maps) correlate with different categories of words or phrases.

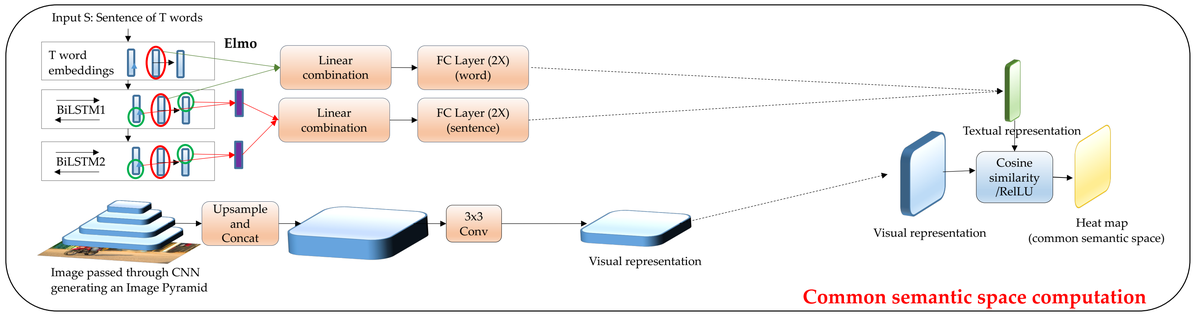
2. **Generate better visual representation:** In order to extract more localized information from the CNN layers, we experimented with larger convolution layers while creating the visual representation. We also experimented with changing how we combined features from the different CNN layers by not selecting one layer but instead took an Image Pyramid approach and up-sampled and concatenated the features

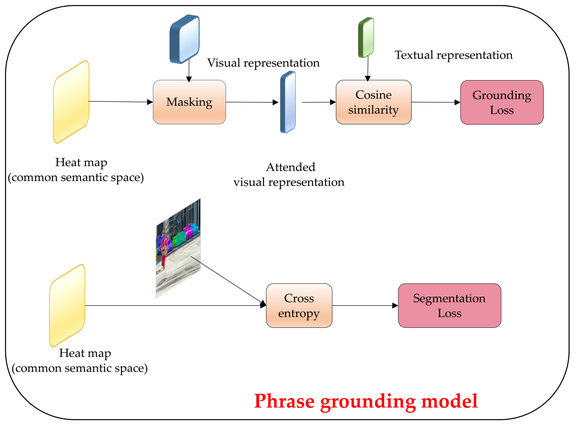
3. **Leverage supervised image segmentation data:** We utilized the instant segmentation data available in the MSCOCO dataset to add an additional supervised loss to the model. For this, we generated a heatmap for each category present in the image which we treated as probability predictions for the segmentation mask.

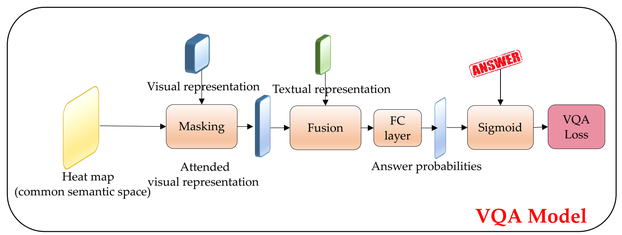
Furthermore, we extended the model to perform the task of Visual Question answering by

- Fusing the attended visual and textual representations and passing through multiple dense layers
- Treated the output of the final FC layer as a distribution over likely answers.
- Added our own Bi-LSTM layer to obtain contextualized sentence embeddings instead of using the ELMo ones.

MODEL ARCHITECTURE







EXPERIMENTS AND RESULTS

Our visual model was VGG network, we trained on Visual Genome Data and tested on Visual Genome, Flickr30 and Refer it dataset.

Level selection analysis: It seems the last layer was selected most of the time for sentence as it contains the most semantic information. For common verbs, the middle layer was selected. Second layer was hardly selected

Better Visual Extraction: Using 3x3 convolutions to extract more fine-grained information and using pyramid architecture to combine the information significantly improves the result

Segmentation Branch: Using supervision from segmentation task improves model to perform better on grounding task

Image Grounding Results

Pointing Game Hit Accuracy	Visual Genome	Flickr30k	ReferIt
Baseline model	**48.76**	**60.08**	**60.01**
Larger convolution (3 x 3 conv)	50.4	67.27	62.07
Pyramid Architecture	54.1	67.90	62.3
Segmentation Loss	**55.3**	**68.4**	**64.3**

EXPERIMENTS AND RESULTS

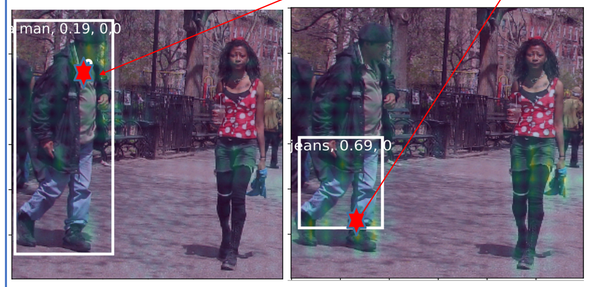
We trained our model on MSCOCO dataset and report results on the test-dev dataset as part of the VQA2.0 challenge. The current architecture did not scale that well to VQA

VQA Results

	Overall	Other	Number	Yes/No
State of the art	66.72	56.77	46.65	83.70
Our Model	54.59	44.89	36.5	70.76

QUALITATIVE RESULTS

A woman in a polka-dot shirt and black boots is walking down the street and being approached by a man in a black jacket and jeans



CONCLUSIONS

- Extracting better visual features can improve grounding models
- Leveraging supervision from task like segmentation can help model in understanding concepts better
- For task like VQA we need to further improve the model by using better attention mechanism