Improving Image Captioning and Visual Question Answering on Visual Genome dataset using Depth Map

Saumya Mehta(mehtasau), Madhav Jariwala(makejari), Krutik Oza(kaoza)

Abstract

Models trained to perform various computer vision tasks often struggle to perform cognitive tasks that require them to learn interactions and relationships between objects in an image. That is why we are using Visual Genome dataset which has dense annotations of objects, attributes, and relationships within each image. We hypothesize that adding depth maps to image features can help improve understanding of visual relationships which can then be leveraged to various dependent tasks such as Image Captioning and Visual Question Answering. Incorporating depth maps alongside other image features can help extract non-spatial relationships like 'holding' alongside spatial features like 'standing behind'.

INTRODUCTION

Scene graph generation is a task of mapping objects and their relationship of an image into a scene graph. For example, a person and a chair are objects in an image, but the relation between those two objects, the person 'sitting' on a chair is crucial to understand for any computer vision system. Datasets like visual genome[5], coco[6] provide us with information like class labels, bounding boxes and RGB features which provide meaningful information to capture spatial information. We hypothesize that adding depth information alongside these image features can help bring out non spatial relations between objects in images like *"behind", "holding"*. Our goal is to study the quality of image captions generated by using relations obtained after adding depth information.

Visual genome dataset does not provide us with depth information so, we generate pseudo depth maps from 2D visual genome images. This is done by training a RGB2Depth network using visual genome and NYU-Depth-V2[7] images. This network learns a mapping from RGB images in visual genome dataset to their corresponding depth maps. The corresponding class labels, location vectors(bounding boxes), RGB features and Depth features are then passed into a relation detection network which generates relations between objects in the image. A relation in relation detection network consists of a subject, a predicate and an object.

We use Macro Recall@K as our evaluation metric. We use Macro Recall instead of Recall@K because visual genome dataset has a high imbalance and some relations are very under represented. Using Macro Recall@K helps bring out the contribution of those under represented relations in the dataset.

Our contributions to this project are:

1.) Study the effects of adding depth maps on visual relation detection and observe generated relations.
2.) Add depth information to visual genome dataset and use the new dataset to train our relation detection network
3.) Apply the generated scene graph information to image captioning

RELATED WORK

a.) Depth Maps

Several works have leveraged depth maps to improve object detection[2],[3], but there have been a few attempts at utilizing depth maps to improve relation detection. Yang et el[1] uses hand crafted depth map features to improve visual relation detection while Sharifzadeh et el [4] uses their own Depth-VRD dataset to generate new relation detection which is our base paper.

b.) Scene Graph generation

Visual Relation Detection[8] kicked off Scene Graph generation as a concept. In Visual Relation detection, Word2Vec representations of subject, object and predicate along with predicate's image region(joint bounding box of subject and object) were used to train a model. Incorporating Knowledge graphs further improved the model's performance. Recent advances include Iterative Message Passing[9], Neural Motifs[11], Graph R-CNN[12] and attention graphs[13] to incorporate context within each prediction.

DATASET

a.) Visual Genome

Models trained to perform various computer vision tasks often struggle to perform cognitive tasks that require them to learn interactions and relationships between objects in an image. For example, Model Andrej Karpathy, Li Fei-Fei[14] describes one of the MS-COCO images as "two men are standing next to an elephant". But it is missing a deeper understanding of the image of what each object is doing and what is the relationship between different objects.

Author added three key elements which were needed in the existing dataset. a grounding of visual concepts to language (Kiros et al., 2014), a more complete set of descriptions and QAs for each image (Johnson et al., 2015), and a formalized representation of the components of an image (Hayes, 1978). And thus introduced the Visual Genome Dataset. Visual Genome dataset sets itself apart from other existing datasets as it not only focuses on objects like other datasets, in Visual Genome dataset, relationships and attributes are also part of every image in the database.

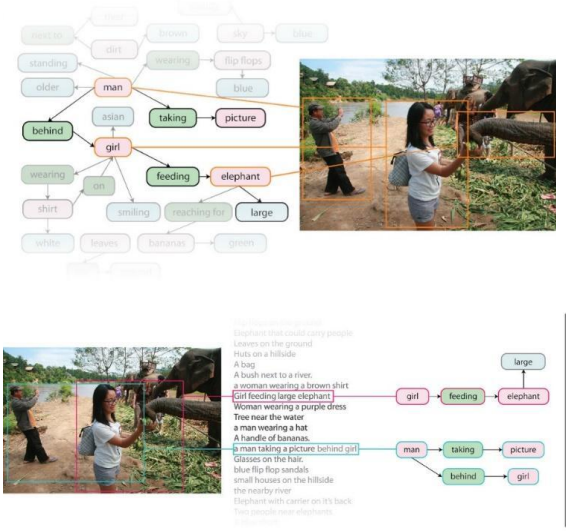

Fig 1 Relationships and attributes in visual genome dataset

Visual genome is the first dataset which can provide a structured formalized representation of an image.
Visual Genome dataset consists of seven components: Region Description, Objects, Attributes, Relationships, Region graphs, Scene graphs and Question Answer pair.

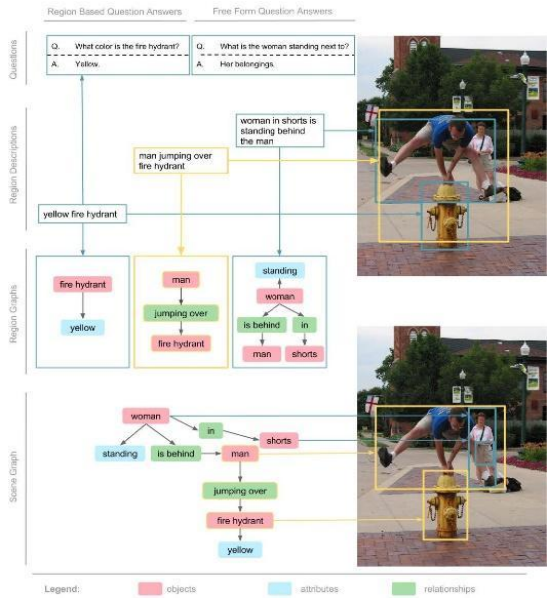

Fig 2 Visual Genome dataset components

There are 108k images in the VG Dataset. Each image has 35 objects, 26 attributes, and 21 relationships between objects on average. Objects can have zero or more attributes associated with it, which is colour, state etc. Due to canonicalization to its WordNet, it avoids giving multiple names to a single object, example: a man, a person, or a human. And it also helps connect information across the image which contributes to extracting relationships between objects.
A relationship is a link between two objects. Actions, spatial, comparative, or prepositional phrases can all be used to describe these relationships. A directed graph representing each region is created by combining the objects, attributes, and relationships from the regional descriptor. These relationships are directly from an object called subject to another object.

b.) NYU Depth V2

The dataset contains 1449 labelled images both in rgb and depth with 464 different scenes from 3 different cities and 407024 unlabelled images. The labelled dataset is synchronized and pre-processed to fill in missing depth labels. Labelled dataset is provided in the .mat format. The raw dataset is also available which contains raw images and depth value and accelerometer values from Kinect. The raw dataset is not pre-processed and is asynchronous.
For training the depth model based on ResNet-50, NYU Depth V2 is used as any other dataset like VG, and does not provide depth maps for images.

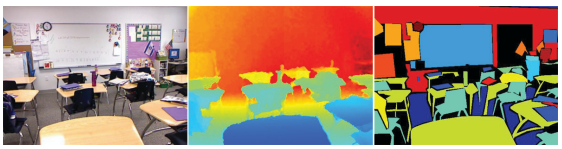

Fig 3 Output from the RGB camera (left), preprocessed depth (center) and a set of labels (right) for the image.

NYU Depth V2 contains RGB-D pairs. NYU Depth V2 and a CNN is used to map RGB images to its corresponding depth map.

NETWORK ARCHITECTURE

a.) RGB to Depth Network

Further this architecture consists of a ResNet-50 architecture by [17] which generates depth maps for its corresponding RGB images. This model is trained on NYU Depth Dataset v2, and it is trained from scratch. As mentioned in [17]**,** the first part of the architecture is initialized by pretrained weights and the second part of the architecture is trained by the number of convolution and unpooling layers. Unpooling layer increases the size of the feature map, opposite to the pooling layer. It is followed by the dropout layer. Then the final convolution layer predicts the results

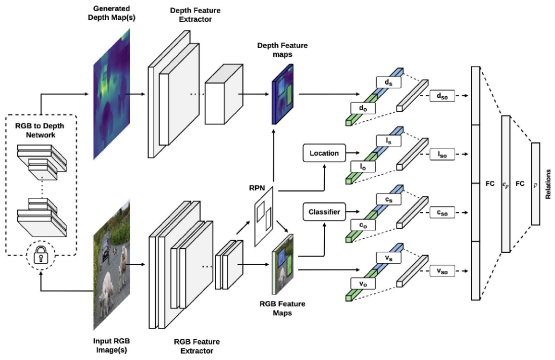

Fig 1 Network Architecture

b.) Feature Extraction Network:
Features of RGB images are extracted using VGG16[16] architecture and it is pretrained on ImageNet Dataset and fine-tuned on Visual Genome Dataset. The model is optimized using Stochastic Gradient Descent mentioned in Neural Motifs[11] with momentum, with batch size of 18 and learning rate of 1.8*10^-2.

ResNet18 Network is used to extract features from the depth maps. Feature extraction for depth images is trained from scratch purely on depth maps. Learning rate used to train this network is 10^-4 and batch size of 32 for 30 epoch. This network is trained from scratch as the pretrained model is trained on RGB images. We want to use this network for depth images so the pretrained information is considered not useful and so the network is trained from scratch.

c.)Relation Detection Network:
Relation detection model is a fully connected layer which takes in subject and object as input and generates a Bernoulli variable for each predicate class. Each relation (subject, predicate, object) returns 1 if the relation is present else 0.
Each feather is then passed into fully connected hidden layer which has 64 neurons for class probability with dropout weight of 0.1, 512 neurons for RGB feature map with dropout weight of 0.8, 4096 neurons for depth feature maps with dropout weight of 0.6 and 20 neurons for location features with dropout weight of 0.1. Output of this layer is then connected to the next hidden layer with 4096 neurons with 0.1 dropout and the last layer is classification layer. Batch size of 16, 30 epochs for training and the learning rate is 10^-5.

EVALUATION

We use two types of Metric to evaluate the performance of the mode. First is Recall@K which for simplicity purposes, we would refer to as Micro Recall@K. Second, we use Macro Recall@K. Macro Recall has its advantages over Micro Recall@K. Let's discuss them in detail.

a.) Micro Recall@K:

This metric is chosen for this dataset because even though the generalized trained model can detect the relationship accurately, the test set might not have similar labeled relationships between given subject and object. This can yield incorrect results which we would want to avoid. Micro Recall is the metric that shows proportion of the relevant prediction that was in the top-K predictions. There are some limitations of this metric and that is why we can instead use Macro recall.

b.) Macro Recall@K

Macro Recall@K is the metric that overcomes the limitation of the Recall@K or micro Recall@K. When we have a highly imbalanaced dataset, micro recall gives more weightage to classes with most occurrences, This can be misleading because it does not take into account the underrepresented

classes. That is why we use Macro Recall@K to alleviate this problem.

$$MACRO\ R@K = \sum_{(s,p,o) \in T_{p'}} \frac{MICRO\ R@K(p)}{|T_p|} \quad (1)$$

Here, the $T_p \subset T$ is set of all relations with predicate p for which micro recall@K is calculated.

RESULTS

Model	R@20	R@50	R@100
Model-v	48.18	55.49	60.23
Model-(v ,d)	52.44	59.99	62.43
Depth-(l, c,v,d)	57.16	65.32	67.21

Table 1 Micro Recall@K scores

Model	Macro R@20	Macro R@50	Macro R@100
Model-v	7.28	8.91	11.31
Model-(v ,d)	8.74	11.30	12.40
Depth-(l, c,v,d)	15.9	17.51	19.13

Table 2 Macro Recall@K scores

Here, we can analyze that the model trained with all four features: class feature, location feature, depth feature and visual features. The performance starts decreasing when we use these features in combination or if we use each feature on its own. The model with only visual features and depth features does okay but not as good as the best model. This shows that even though depth features can improve the performance of the mode, it requires other important information about the image such as localization of the image and class features.

DISCUSSION

From our experiments, we observed that adding depth features on top of image features results in a better performance. Our best performing model was the lcvd model which passed bounding boxes, class labels, visual features(RGB features) and depth features. The results from macro Recall@K show that adding depth features can bring out the contribution of under-represented relations more so than just using visual relations which is in line with our hypothesis.

CONCLUSION

We deployed an RGB-Depth network trained on visual genome and NYU-Depth-V2 datasets to add depth information to visual genome dataset. We used a metric macro Recall@K to evaluate performance on a highly imbalanced dataset and we studied performance of our relation detection models with and without depth maps and presented an analytical summary of the same.

This project can be further extended by using the newly generated relations to tackle visual question answering tasks.

REFERENCES

[1] Hsuan-Kung Yang, An-Chieh Cheng* , Kuan-Wei Ho* , Tsu-Jui Fu, and Chun-Yi Lee. 2018. Visual Relationship Prediction via Label Clustering and Incorporation of Depth Information. In ECCV

[2] Liefeng Bo, Xiaofeng Ren & Dieter Fox. Unsupervised Feature Learning for RGB-D Based Object Recognition. In STAR

[3] Andreas Eitel, Jost Tobias Springenberg, Luciano Spinello, Martin Riedmiller, Wolfram Burgard. 2015. Multimodal Deep Learning for Robust RGB-D Object Recognition. In CVPR

[4] Sahand Sharifzadeh, Sina Moayed Baharlou, Max Berrendorf, Rajat Koner, Volker Tresp. In 2020. Improving Visual Relation Detection using Depth Maps.

[5] Ranjay Krishna, Yuke Zhu, Oliver Groth Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, Li Fei-Fei. 2016.Connecting Language and Vision Using Crowdsourced Dense Image Annotations. In

[6] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays,

Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, Piotr Dollár. 2015. Microsoft COCO: Common Objects in Context. In CVPR

[7] NYU Depth Dataset V2.[dataset]

[8] Cewu Lu, Ranjay Krishna, Michael Bernstein & Li Fei-Fei. 2016. Visual Relationship Detection with Language Priors. In ECCV

[9]Cewu Lu, Ranjay Krishna, Michael Bernstein & Li Fei-Fei. Visual Relationship Detection with Language Priors

[10]Danfei Xu, Yuke Zhu, Christopher B. Choy, Li Fei-Fei. Scene Graph Generation by Iterative Message Passing

[11] Rowan Zellers, Mark Yatskar, Sam Thomson, Yejin Choi. In 2017. Neural Motifs: Scene Graph Parsing with Global Context

[12]Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, Devi Parikh. Graph R-CNN for Scene Graph Generation

[13]Martin Andrews, Yew Ken Chia, Sam Witteveen. Scene Graph Parsing by Attention Graph

[14] Andrej Karpathy, Li Fei-Fei. 2014. Deep Visual-Semantic Alignments for Generating Image Descriptions. In CVPR

[15]Karen Simonyan, Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition

[16]Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, 2015. Deep Residual Learning for Image Recognition. In CVPR

[17]Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, Nassir Navab. Deeper Depth Prediction with Fully Convolutional Residual Networks

Github:https://github.iu.edu/makejari/DepthMaps