Project Proposal - ECE 176

Benjamin Liang
Mathematics
A16954232

Andrew Onozuka
Computer Engineering
A16760043

Abstract

The overall picture of this project proposal is the implementation of a convolutional neural network to determine the depth and surface normals from images in the NYUv2 RGB-D dataset. The goal of this project is explore new deep learning models, different feature extraction techniques, and different types of layers in a multi layered perceptron to achieve better results.

1 Problem Definition

With the every growing advancements of AI and ML in autonomous vehicles, the problem we hope to solve is the detection of depths of images and surface normal estimations which would be useful to autonomous vehicles to determine how far they are and detect objects. There have been many works around this topic with different implementations being used that vary from different deep learning architectures to the features being used to even the loss function within the deep network. We would like to explore different solutions to deepen our understanding of neural networks and how it can be applied to surface normal predictions and depth perception. Although, these experiments may not result in the best performance, we hope to better understand the model used in cutting edge research around this topic.

One of the topics we would like to get deeper into is the selection of features and how important of a role it plays in the final model. How does the ordering of different layers affect the performance? From previous papers and works on this topic, there have been some ideas of segmenting an image from its different components and then doing a depth estimation into a surface normal prediction. We want to explore ideas such as what would happen if we were to not segment an image and run the depth model as is. Would the performance be better or worse? These are just some the ideas that we would like to dive deeper into to better understand the importance of features in these models.

Our understanding of the problem is fairly limited as of now and we would hope to gain a deeper understanding of it as we continue this project. So far, we understand the overarching model and the features that cutting edge researchers decided to use. We understand that the dataset used for training the model is not a mapping of a matrix of pixel values to labels but rather a matrix of pixel values to another matrix of pixel values representing the surface normals of the objects in the image.

2 Tentative Method

The method we are planning to use is a traditional CNN which utilizes classification of images in one of the layers to then output a continuous output of pixel values representing the surface normal values. We hope to utilize segmentation and hypercolumn representation of pixel values over the many layers we hope to have in our model. Hypercolumn representation of a pixel represennts the concatenation of all layer outputs a some specific pixels which we hope to experiment more with to see if it affects the performance in any way and by how much. Our model will utilize a sequential model with multiple hidden layers.

3 Experiments

We are planning to use the NYUv2 RGB-D and other similar datasets as used in both [1] and [2]. The NYU depth dataset v2, or NYUv2 for short, is a widely used dataset for depth estimation and semantic segmentation tasks using computer vision. It consists of RGB images paired with corresponding depth maps, captured from a variety of indoor scenes. The scenes cover diverse indoor environments, including bedrooms, kitchens, living rooms, and offices, so that depth perception algorithms can be trained on from a large scale source.

As for the data format of the dataset, lots of structure is already provided, typically organized into separate folders for RGB images and depth maps. Each RGB image is accompanied by its corresponding depth map, enabling paired training for depth estimation models. Both RGB images and depth maps are commonly stored in standard image formats such as JPEG or PNG.

Some other relevant information: The dataset includes synchronized RGB-D data, enabling joint learning for tasks like depth estimation and semantic segmentation. Ground truth annotations for semantic segmentation are available for a subset of scenes, supporting supervised learning experiments. NYUv2 is commonly divided into training, validation, and test sets, enabling standardized evaluation of models.

The goal of our experiment is to combine insights from both papers to enhance dense prediction tasks. We want to develop a hybrid architecture incorporating elements from the Mask Transformer model and surface normal prediction techniques, to leverage the model's efficiency in processing dense data while integrating features inspired by surface normal prediction for improved accuracy and robustness. We will evaluate the performance of this hybrid approach on tasks such as semantic segmentation and depth estimation using datasets like NYUv2.

Our experiment will either look like using segmented images to train our neural network, or taking semantic segmentation and combining it with CNN instead of a PolyMax [2]. We will compare our results against the state of the art methods already shown on the same dataset - NYUv2.

References

[1] Aayush Bansal, Bryan Russell, and Abhinav Gupta. Marr revisited: 2d-3d alignment via surface normal prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5965–5974, 2016.

[2] Xuan Yang, Liangzhe Yuan, Kimberly Wilber, Astuti Sharma, Xiuye Gu, Siyuan Qiao, Stephanie Debats, Huisheng Wang, Hartwig Adam, Mikhail Sirotenko, et al. Polymax: General dense prediction with mask transformer. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1050–1061, 2024.