Final Report

Deep Learning In Computer Vision

Liang Hu lh3057, Weiran Wang ww2584, Yongfei Tan yt2775

1. INTRODUCTION (Weiran)

More and more domestic garbage waste is generated with the increasing of people's quality of life. Most of the wastes end up in landfills, and the waste does not go to the correct category. Some of modern waste contains chemical substances and is harmful for the earth. Dropping waste without classification causes issues that involve increase in landfills, eutrophication, consumption of toxic waste by animals, and etc. Recycling is the process of converting waste materials into new products. Recycling can prevent the waste of potentially useful materials. Classification of waste as organic or recyclable can help people quickly recognize types of waste. It will be helpful to not only save potential useful resources but also reduce toxic waste ending up in landfills.

We trained classification models predicting the class of waste in organic or recycled using 5 different pre-trained CNN models AlexNet, VGG16, GoogleNet, ResNet18, DenseNet121. What's more, we trained models using Vision Transformer to make comparisons between CNN models and transformer models.

2. OBJECTIVE (Weiran)

First of all, we intend to import data of waste images in two classes (recycle and organic) from Kaggle. Then, we classify images of waste into two categories. In general, detecting the picture, and then we identify them by analyzing components in the binary image captured and choose the corresponding on their relative pixel size. Our goal is to reach at least 90% accuracy on classification models. We aim to train these pictures on multiple models, and make a comparison between CNN model and Transformer model.

3. RELATED WORK (Liang Hu):

The paper “Advances in deep learning approaches for image tagging”, is to categorize and evaluate different image tagging approaches based on deep learning techniques like Alexnet, Vgg16, GoogleNet, and Resnet. We also use similar comparison techniques to analyze the efficiency and accuracy of our waste classification dataset. In addition, the paper also discussed the relevant problems and applications to image tagging, including data collection, evaluation metrics, and existing commercial systems. We also include some evaluation metrics when we generate our own accuracy.

4. DATASET (Liang Hu):

Waste classification into two categories organic and recycle. First, the organic category and recycling category are the most common categories for waste. Two categories of classification problems This makes it easier to collect and annotate a large dataset, and it also makes it easier to train and evaluate deep learning models.

Additionally, classifying waste into just two categories is a good way to start to compare and contrast different CNN models. Properly sorting waste into organic and recycled categories can help reduce the amount of waste that ends up in landfills and improve the efficiency of recycling programs.

Furthermore, classifying waste into just two categories is still a challenging problem for deep learning algorithms. It requires the ability to recognize and classify a wide variety of objects, many of which may be similar in appearance This means that developing a deep learning model that can accurately classify waste into just two categories requires training using different optimization neural networks

Meanwhile, our dataset was divided into an 85% training set and a 15% testing set. In specific, the training set images are 22564, and the testing set images are 2513. The two main categories of classification are organic and recycle.

5. METHOD

5.1 AlexNet (Yongfei Tan):

We used AlexNet as the starting point for our comparing and contrasting experiments. AlexNet achieved revolutionary performance on the ImageNet challenge in 2012. The Alexnet has eight layers with learnable parameters in total. The model consists of five convolutional layers, followed by 3 fully connected layers. It uses ReLU activation and two dropout layers. We tried adding a fully connected layer followed by a final sigmoid layer to replace the last fully connected layer and 1000 category softmax layer, with binary cross entropy loss function. We also tried modifying the out_feature variable of the last fully connected layer and used the cross entropy loss function, and finally achieved 87% accuracy for this model.

5.2 Vgg16 (Liang Hu):

VGG-16 is a deep-learning model. It is a convolutional neural network (CNN) that was trained on the ImageNet dataset, a large dataset of labeled images. The VGG-16 model is made up of 16 layers, including 13 convolutional layers and 3 fully-connected layers.
The VGG-16 model has relatively good performance on our datasets. We replace the last fc layer with a new fc layer with binary cross entropy loss to improve the accuracy. Finally, we achieve an accuracy of around 93%.

5.3 GoogLeNet (Weiran):

The GoogLeNet model was presented at the ImageNet recognition challenge in 2014. It was used to solve image classification problems and object detection problems. GoogLeNet is different from previous models such as AlexNet. It uses different kinds of methods such as 1x1 convolution and average pooling to create deeper architecture. GoogLeNet solved problems that most large classic classification models faced. The most innovative module in GoogLeNet is utilization of the Inception module. The Inception module is running multiple operations with multiple filter sizes in parallel, hence there are no trade-offs in the model. Another method used in GoogLeNet is the 1x1 convolutional layer. The model runs a 1x1 convolution filter to the input before passing them to the parallel operations. This operation reduces the number of feature maps in the input stack.

We had an experiment with GoogLeNet and want to see if this innovative module affects the result. There are 27 layers in the GoogLeNet architecture, including 9 inception modules and 5 max pooling layers. We modified the output class to 2 and the final accuracy of this model is around 85%.

5.4 ResNet-18(Liang Hu):

ResNet-18 is a convolutional neural network (CNN) trained on the ImageNet dataset for image classification. It is made up of 18 layers, including convolutional, pooling, and fully-connected layers. ResNet-18 can be used for 2 category classification by fine-tuning the model on a dataset of images labeled with the two categories of interest. This involves modifying the final fully-connected layers of the ResNet-18 model to output probabilities for the two categories and then training the modified model on the labeled dataset. We change the last fc layer into softmax function, instead of using Sigmoid function. Using softmax, we can output the probability that the image belongs to each of the two classes, which allows us to make a more informed decision about which class the image belongs to. Also, using a binary cross entropy loss function.

Binary cross-entropy loss is often used in two-category classification because it is a measure of the distance between the predicted probabilities and the true probabilities for the two classes. This loss function is designed to be used with models that output probabilities, such as those that use the sigmoid or softmax activation functions. Finally, we achieve the accuracy around 92%

5.5 DenseNet121 (Weiran)

We experimented with the classification model with DenseNet121 and want to see if the CNN model performs better when the depth of deep convolutional layers increases. As convolutional layers increase, there is an issue in classic CNN models that gradients vanish since the path from input to output is too big. DenseNet121 resolves the issue of ‘vanishing gradient’ by simplifying the connectivity pattern between layers. Each layer in DenseNet is connected directly with every other layer. It exploits the potential of the network through feature reuse. For each layer, the feature maps of all preceding layers are used as inputs, and its own feature-maps are used as inputs into all subsequent layers. DenseNet also has the advantage of strengthening

feature propagation, encouraging feature reuse, and substantially reducing the number of parameters.

There are a total of 121 layers in the DenseNet121 model including 120 convolutional layers and 1 fully connected layer. We modified the fully connected layer by adding a softmax at the end of the fc layer and changing the output class to 2. The final accuracy of the model is 96%.

5.6 Vision Transformer (Yongfei Tan):

We also experimented with Vision Transformer, because it emerged as a competitive alternative to CNN models. Vision transformer employs a Transformer-like architecture, and it outperforms the current state-of-the-art of CNN in terms of computational efficiency and accuracy. The model splits the images into a series of positional embedding patches, and then the resulting sequences are processed by the transformer encoder. We defined a ImageClassification and added a linear layer on top of a pre-trained ViT model. In order to set the attributes to the configuration of the model, we specified the number of output, which is 2, and set the label mapping. Then, we set the TrainingArguments to instantiate a Trainer. In the end, we got 90.5% accuracy for the model using Vision Transformer.

6. CONCLUSION (Yongfei Tan)

Model	Accuracy	Layers	Model Description
AlexNet	87%	8	5 convs 3 fc layers
VGG16	93%	16	13 convs 3 fc layers
GoogLeNet	85%	22	21 convs 1 fc layer

ResNet18	92%	18	2 convs 14 res blocks 1 fc layer
DenseNet121	96%	121	120 convs 4 avg pool

This is the result of our experiments using different models. We can see that numerous variations have developed over the years resulting in several CNN architectures, from AlexNet published in 2012 to DenseNet in 2017. The accuracy has been gradually increasing in most of the cases with time. DenseNet121 got the best performance compared to other models. Since each layer is connected to every other layer in the network, the DenseNet model improves parameter efficiency and flow of gradient, and achieves better performance. But on the other hand, it also requires more computational resources than other CNN architectures.

Besides the comparisons of different CNN architectures, we are also interested in other approaches, which leads us to the Vision Transformer. Vision transformer is different from CNN, and it employs a Transformer-like architecture. With the self-attention mechanism and transformer model, encoder-decoder structure, the model can integrate information across the entire image even in the lowest layers. But it also needs pre-training on large external datasets.

We also plan to discover more models using transformer architecture, such as Swin transformer, which is a new vision transformer published in 2021. It uses two key concepts, Hierarchical Feature Maps and Shifted Window Attention, and brings greater efficiency and has greater accuracy. We hope to get better performance with this model.

7. REFERENCES

1. Fu, Jianlong and Yong Rui. “Advances in deep learning approaches for image tagging.” *APSIPA Transactions on Signal and Information Processing* 6 (2017): n. pag.
2. Alzubaidi, L., Zhang, J., Humaidi, A.J. *et al.* Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *J Big Data* 8, 53 (2021).
3. Bressem, K.K., Adams, L.C., Erxleben, C. *et al.* Comparing different deep learning architectures for classification of chest radiographs. *Sci Rep* 10, 13590 (2020).

4. Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv preprint arXiv:2010.11929* (2020).
5. Liu, Ze, et al. "Swin transformer: Hierarchical vision transformer using shifted windows." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021.

Presentation Link & Github Link:

Presentation link: https://youtu.be/7tAQgbFIOnM

Github link: https://github.com/LarryHu0217/4995proj