

Indoor Scene Recognition using ResNet-18

Hafiz Zeeshan Ali^a, Summiya Kabir^b, Ghufuran Ullah^b

^a hali.mscs18seecs@seecs.edu.pk

^b skabir.mscs19seecs@seecs.edu.pk

^b ghufuran22bs@gmail.com

^{a, b} School of Electrical Engineering and Computer Science, NUST, Islamabad, Pakistan

^b School of Computer Science and Technology, NCWU, Zhengzhou, China

Abstract

Deep learning allows computational models consisting of multiple neural layers of processing to learn data representation at multiple abstraction levels. Such approaches have greatly strengthened state-of-the-art speech recognition, visual object recognition, scene recognition, NLPs, object detection, and many other fields such as drug discovery, distant surgeries and genomics. Scene Recognition is an area of visual recognition where we design and automate our system to recognize and identify the scene of the image. Automatic Scene Recognition or Scene Analysis is one of the hot topics in Deep Learning. If we look at the contributions of Deep Learning in this decade, we had come to know that Scene Recognition has been an obvious concern for scientists as it has great significance in security and surveillance too. Object recognition and recognition of indoor scenes plays a significant role in the cognition of service robots in the field. The development of deep learning has made fine-tuning of CNN (Convolutional Neural Network) on target datasets a common way of solving classification problems. Nonetheless, this approach cannot achieve adequate results easily for indoor scene classification due to over fitting when the datasets for scene preparation are inadequate. In order to compare techniques that participate to better accuracies, we have applied two different techniques to achieve our results i.e. Fine Tuning and concept of Freezing Layers. Within this project, a system of classification of the indoor scene is proposed to solve this issue. We are using ResNet-18 which contains 18 deep layers for classification. Furthermore, we are using transfer learning and performing classification on scenes based images.

Keywords: ResNet-18, Transfer Learning, Convolutional Neural Networks, Object recognition, Deep Learning;

1. INTRODUCTION

Scene detection is one of Computer Vision's products, implementation of Deep Learning, and requires a sense of object recognition to be expanded. While the availability of broader datasets such as ImageNet has improved the success of Convolutional Neural Networks (CNNs) in learning high-level features, scene recognition output has so far not achieved much quality. Scene Recognition is an area of visual recognition where we design and automate our system to recognize and identify the scene of the image. Automatic Scene Recognition or Scene Analysis is one of the hot topics in Deep Learning. If we look at the contributions of Deep Learning in this decade we'd come to know that Scene Recognition has been an obvious concern for scientists. Our ultimate goal is to accomplish results up to the feats of the human brain [1]. Even though there are few large datasets available e.g. ImageNet has increased Convolutional Neural Networks (CNNs)'s progress in learning high-level features, scene recognition performance has not achieved much efficiency so far. With thousands of labeled images we have made an effort, using Indoor scene image dataset from Kaggle repository, designed to represent real world indoor places and scenes. We are to implement re-engineered

Convolutional Neural Networks (CNNs) strategies to perform scene recognition that hits maximum accuracy and observe a novel measure of density and diversity to demonstrate the usefulness of these quantitative measures to estimate biases in the dataset and to compare different data sets. As datasets like ImageNet are not competitive enough for tasks like Scene Recognition, our idea is to implement a scene-centric database called Indoor Scenes by Kaggle. We plan to show state-of-the-art results on all existing scene metrics, utilizing our deep tools. Scope includes improved accuracy performance in Indoor Scenes dataset with Convolutional Neural Networks (CNNs) combination features. We are to implement re-engineered Convolutional Neural Networks (CNNs) strategies to perform scene recognition that hits maximum accuracy and observe a novel measure of density and diversity to demonstrate the usefulness of these quantitative measures to estimate biases in the dataset and to compare different data sets. As datasets like ImageNet are not competitive enough for tasks like Scene Recognition, our idea is to implement a scene-centric database called Indoor Scenes by Kaggle. We plan to show state-of-the-art results on all existing scene metrics, utilizing our deep tools. Scope includes improved accuracy performance in Indoor Scenes dataset with Convolutional Neural Networks (CNNs) combination features. As deep Convolutional Neural Networks are designed to benefit from a huge amount of data and learn from it, we are trying to implement scene recognition and our goal is to achieve results with efficient accuracy. A key aspect of Scene Recognition is the identification of the places where the objects are seated (e.g. bed, chair, sofa, light bulb, fan ...). We aim to change the idea of using the category of places by providing a more exhaustive list of the indoor scenes in the picture and a description of their spatial relationships, a category of indoor scenes will provide the appropriate level of abstraction to avoid such a long and complex description [2].

2. LITERATURE REVIEW

A technique of Convolutional Neural Network named Multilevel Ensemble Network (MLEN) was proposed by [7] to enhance the accuracy in the identification of “small object-supported scenes”. Separate predictions were produced by applying features from various levels of the net [4]. Therefore, to render the final estimate, ensemble learning was done inside the net. Also, “Feature Transfer Path” was added and implemented a feature fusion technique to completely control low-level and high-level features but for further improvement in the accuracy, they had designed a class-weight loss function to resolve the issue of non-uniform class distribution. In the listed research [1], RGB-D Scene Recognition was implemented where depth and RGB features were merged by projecting them into a shared space and leaning a multilayer classifier further, which is jointly configured in an end-to-end network. This paper [1] proposed discriminatory patch representations in which neural networks were used and also proposed a hybrid model in which a semantic manifold was constructed on multi-scale CNNs. They devised rich background models that use Markov random fields to incorporate different dimensions, spatial connections, and different functions. Interactions significantly had improved accuracy. This paper was the extension of the limitations of the previous work, used Semantic manifold on the multi-scale CNNs model, and accuracy was improved by nearly more than 95% for outdoor scenes. It is elaborated in [8][9] that apart from exposure to a dense and rich variety of natural images, its hierarchical organization in layers of increasing processing complexity, an architecture that inspired Convolutional Neural Networks or CNNs, is an important property of the primate brain. Together with recent large databases, these architectures have achieved astonishing performance on classification tasks of objects. However, the baseline performance attained in scene classification tasks by these networks is below the output spectrum dependent on hand-designed interfaces and sophisticated classifiers.

3. METHODOLOGY

There are a lot of open-source deep learning frameworks such as Teano, Cafe, and MXNet, which have promoted the development of Deep Learning oriented projects.

3.1 Why Transfer Learning?

Transfer learning is a machine learning technique in which a model built for the first task is then reused as a starting point for a model created for the second task, for example, if we train a model to classify birds and pets, then use the same model changed just a little bit in the last layer to classify bees and docks. This is a common approach to deep learning that enables the fast development of new models and that is necessary as the preparation of a brand new model will take a lot of time, it can take many days and even weeks to complete. So if we use a pre-trained model then we typically only change the last layer and then do not need to retrain the entire model. Transfer learning, though, will produce fairly decent outcomes in success and this is why it is so common today. It saves time and focuses on performance [3].

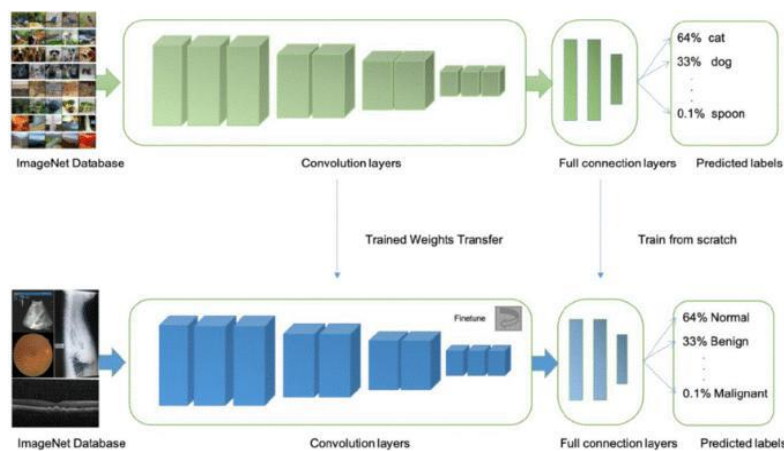


Fig.1 Transfer Learning

3.2 Proposed Model Flow

First, the CNN extraction feature model needs to be built based on pre-trained models using a flattened layer instead of a Softmax layer. Transfer learning is when a model is reused on a second task for a model developed for one task. The most common example given is when a model is trained on ImageNet with a second task being fine-tuned. ResNet, short for Residual Networks, is a classic neural network used as a backbone for several computer vision tasks. This model won the ImageNet Challenge in 2015. The fundamental advancement of ResNet was that it enabled us to effectively train extremely deep neural networks with 150+ layers. Workflow is shown in Fig. 2.

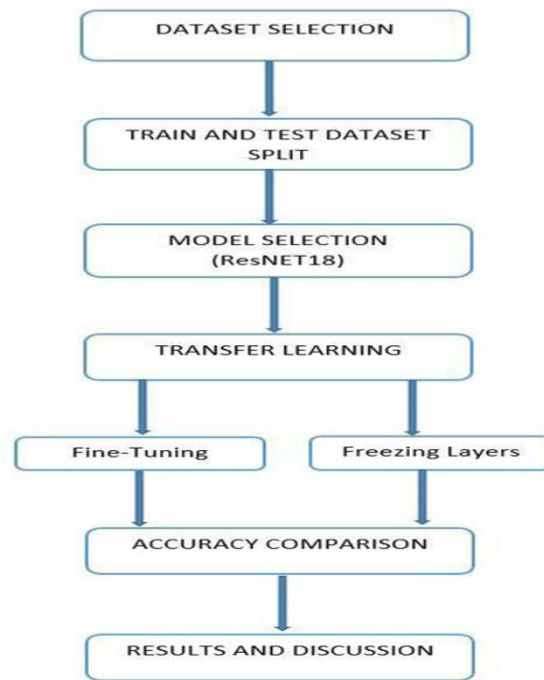


Fig.1 Proposed Research Model

3.3 Fine Tuning & Freezing Layers

Fine Tuning technique is flexible as model with the same design as a pre-trained model, which has demonstrated outstanding results in performing specific tasks to the one we are attempting to achieve and learn from scratch. We delete every layer's weights from the pre-trained model and retrain the whole model on our results. Here, technique has to be customized according to model requirements. Freezing a layer avoids the change in the weights. In transfer learning, this approach is sometimes used when the base model (trained on any other dataset) is made frozen and its layers cannot be updated further. In the case of images, same are run through the layers without updating weights.

4. EXPERIMENTATION

4.1 EXPERIMENT 1 (Fine Tuning):

We have utilized the concept of Transfer Learning in both experiments for our review of Indoor Scene Recognition. Transfer learning is described as enhancing learning in a new role by information transfer whereas most machine learning algorithms are built to solve single tasks, designing algorithms that promote

learning transfer is a topic of on-going interest in machine learning and learning from a similar task that has already been mastered [2][6]. It saves time and focuses on efficiency. We built a new layer and added the layer to the last one. To refine the model parameters, we used cross-entropy loss for loss and SGD optimizer, and the learning rate is set to 0.001. We have used the scheduler concept for the updating of the learning rate. We also chosen the phase size equivalent to 7 and gamma equals 0.1, which implies that our learning rate is compounded by 0.1 per 7 epochs, which implies that our learning rate is only modified to 10 per cent at every 7 epochs.

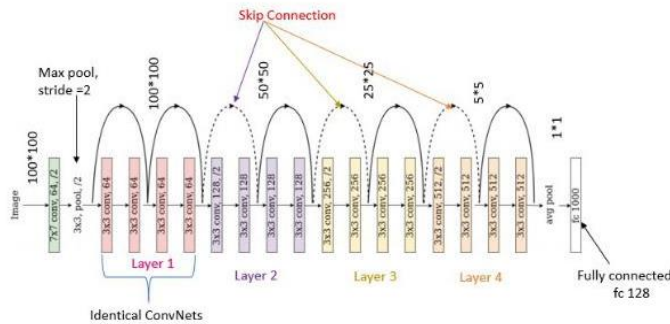


Fig.3 ResNet-18 Architecture (pluralsight, 2020)

By using a fine-tuning approach, what we did is based on ResNet18. ResNet-18 supports many types of programming languages and deep learning algorithms and offers a variety of pre-trained deep CNN models based on a variety of large-scale datasets. The deep CNN models chosen are all ResNet-18 with 18 deep layers of the network. The model used was pre-trained. We used “Indoor Dataset” Downloaded from Kaggle divided it into different categories of indoor scenes. We made a dataset for testing out of the available Kaggle dataset. We used only trained images and divided a portion into test images.

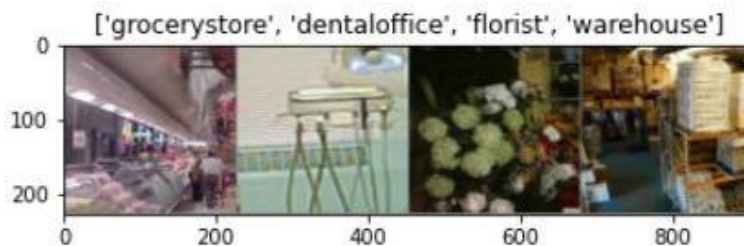


Fig.4 Grid of Batch Data

The purpose of using Transfer Learning was to fine tune the model and fine tune the weights of ResNet-18 to achieve the best accuracy possible.

4.2 EXPERIMENT 2 (Freezing Layer):

First, in the beginning, we have frozen all the layers of pre-trained model and just train the very last layer. So we have looped all the parameters, and set the gradient equals to false, because it would freeze all the layers at

the beginning and then we have built the new last layer. After building a new last layer, we set the gradient equals to true and then the last layer loss, optimizer and scheduler have been set accordingly.

For both experiments epoch number is equivalent to 50. GPU was accessible for the first experiment and it took approximately 1.5 hours to train but for the second experiment because GPU was not accessible then we used CPU so it took 9.8 hours to train. Both experiments were performed in Google Colab.

5. RESULTS AND DISCUSSIONS

We tried to train our model so that it may generalize well. As datasets like ImageNet are not competitive enough for tasks like scene recognition, our idea is to implement a scene-centric database called Kaggle's indoor scenes with over thousands of labeled pictures of scenes. We aimed to demonstrate state-of-the-art performance by applying our deep features to all current indoor benchmarks. The scope requires increased results in terms of precision of the data collection. To avoid errors on training we tried to generalize as a training set in Neural Networks has input and output nodes and desired hidden layers as, despite many successes, Convolutional Neural Networks still suffer from a significant weakness called Over-fitting. The training set error is normally pushed to a relatively low amount, but the error becomes very high as fresh data is sent to the network. We made sure that the testing scenarios were memorized by the network and taught to generalize to different runtime circumstances. By the end of our training and research work, we have managed to achieve an accuracy of 74% with training time 142m 2s. On the other hand freezing layers optimization resulted in 63% accuracy as weights were made frozen except the last layer as in the last layer we had updated layer according to our dataset sub categories.

6. CONCLUSION

Advanced Object Detection or Scenario Interpretation is one of the main topics in Deep Learning. We tried to experiment with Indoor Scene Recognition here, and the ResNet model we used was a pre-trained one. 'Indoor Scenes Dataset' was downloaded from 'kaggle' repository and images available in the dataset were only for training purposes that is why we split a part of them into the test images. The transfer learning method was used and the weights of ResNet18 were further fine-tuned according to our dataset. After experimenting with Fine Tuning and Freezing Layers we find out that fine tuning is much more efficient as due to instability in the dataset, the accuracy is not great in outcome of frozen layers. In the future, we will build more classes in our dataset that does not only contain the indoor scenes but also outdoor scenes. Accuracy will be further developed to suit the state-of-the-art tests.

References

- Jiang, Shuqiang. 2017. Multi-Scale Multi-Feature Con-text Modelling for Scene Recognition in the Semantic Manifold. IEEE.
- Hussain, M., Bird, J. J., & Faria, D. R. (2018). A Study on CNN Transfer Learning for Image Classification. In *UK Workshop on Computational Intelligence* (pp. 191-202). (Advances in Computational Intelligence Systems; Vol. 840). Springer. https://doi.org/10.1007/978-3-319-97982-3_16
- kaggle. n.d. ResNet-18 Architecture (<https://images.app.goo.gl/VucAo43hFR4uNud48>).
- LeCun, Yann. n.d. "Deep learning." n.d. link.springer. (<https://link.springer.com/chapter/10.1007/978-3-319-97982-3>)
- MIT. n.d. <http://places.csail.mit.edu/>. Sight, plural. 2020. ResNet-18 Architecture.

<https://images.app.goo.gl/VucAo43hFR4uNud48>.

Sun, Ning. n.d. Fusing Object Semantics and Deep Appearance Features for Scene Recognition. IEEE

Zhang, Longhao. 2019. Multi-Level Ensemble Network for Scene Recognition.

Zhou, Bolei. 2014. Learning Deep Features for Scene Recognition using Places Database. nips.com

Zhou, Bolei. 2017. Places: A 10 million Image Database for Scene Recognition.