Machine Learning Engineer Nanodegree

Capstone Project Proposal

Yiu Shun Wilson, LAM

2nd November, 2020

Indoor Scene Recognition

Domain Background

Indoor scene recognition has a wide variety of applications: in robotics, for example, it plays a major role in environment cognition and navigation[1]; in assistive technology recognizing indoor scene and environment helps develop applications and solutions for individuals with visual impairment to safely navigate indoor[2], and potentially, from the perspective of a speech-language pathologist, forms the ground of developing applications that are able to automate the recommendation / selection of dynamic, environment- and context-sensitive communication content for augmentative and alternative communication (AAC) (e.g. on smartphones or tablets) for end-users accordingly such that they can save time flipping / navigating over categories or pages on their AAC device, while the buttons / cards on the screen are semantically relevant to their physical and/or social context, and hence communication can become more efficient and effective.

However, indoor scene recognition remains a challenging open problem in the field of computer vision and artificial intelligence. Compared to open scene recognition, indoor scenes present vast variations, rich contents in the input and overlapping features among output classes; some indoor scenes are well defined by global spatial properties while some by local objects they contain. Optimizing indoor scene recognition models therefore yields importance to both machine learning/computer vision and real-world applications.

Problem Statement

Indoor scene recognition is a multinomial classification problem. Given an image as input and *N* categories of scene, the goal of an indoor scene classifier is to predict which category the image belongs to as output.

Datasets and Inputs

This project will initially and primarily use the MIT Indoor 67 dataset, which was developed in 2009 in MIT [3] and subsequently used by studies on indoor scene recognition [e.g. 1,2,4] and QSTP in-class competition on Kaggle. It contains 67 label categories and 15620 images, with each category containing around 100 images; the original, standard train-test split was 80:20 for each category. Details of dataset retrieval and description can be found in README.md.

Additional objective and input: Depending on the initial results, if we could reach or exceed the performance of the benchmark model which will be outlined below, an attempt could also be made to the SUN397 dataset, which is considered a more challenging dataset.

Solution Statement

Following the best practice in most image classification tasks, this project will fine-tune pre-trained Convolutional Neural Networks (CNN) with data augmentation. In particular, various EfficientNet [5] model architecture will be used to train our classifier, given its architectural scalability and reported state-of-the-art performance [2, 5]. Data augmentation such as random rotation, brightness adjustments will also be applied to training data.

Benchmark Models

While it is known that many published studies use a lot more complex CNN architecture on the MIT 67 Indoor dataset (e.g. [4]), the current benchmark will be a relatively simple CNN fine-tuning a pre-trained ResNeXt-101_32x16d, data augmentation, learning rate annealing, and early stopping from 2019 QSTP in-class Kaggle competition (source code). Since the reported accuracy (84.4%) of this model was calculated on the entire dataset, the benchmark model will be replicated by its source code and tested against our test set.

Evaluation Metric

The models will be evaluated on the Top-1 accuracy on a pre-defined test set. Specifically, the number of matches between the most probable predicted class from a softmax layer to the ground truth label will be computed and divided the by the number of test set samples.

Project Design

We will be using open-source pre-trained EfficientNets in the Pytorch framework and perform data pre-processing, training, and evaluation on AWS. The workflow will be as follows: (1) resizing images according to the defined EfficientNet architecture (e.g. 224 x 224 for B0), (2) following the Indoor 67 standard protocol, performing a 80:20 train-test split, and a further 80:20 train-validation split, (3) applying the same data augmentative transformations of our benchmark to *training* images, (4) implementing transfer learning by fine-tuning pre-trained EfficientNet models (B0 – B5), (5) to avoid overfitting, choosing the best model with the highest training and validation accuracy and least training and validation loss, and (6) performing the final test set evaluation.

References

1. Liu, S. & Tian, G. (2019). An Indoor Scene Classification Method for Service Robot Based on CNN Feature. *Journal of Robotics, Vol. 2019.*

2. Afif, M., Ayachi, R., Said, Y., & Atri, M. (2020). Deep Learning Based Application for Indoor Scene Recognition. *Neural Processing Letters, 51.*

3. Quattoni, A. & Torralba, A. (2009). Recognizing Indoor Scenes. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, 2009, pp. 413-420

4. Seong, H., Hyun, J., & Kim, E. (2019). FOSNet: an end-to-end trainable deep neural network for scene recognition. *IEEE Access*, 8, pp. 82066-82077.

5. Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *Proceedings of the 36 th International Conference on Machine Learning*, Long Beach, California, PMLR 97.