Tampere University

Veera Kallio

VIDEO-BASED SCENE CLASSIFICATION USING PRETRAINED NEURAL NETWORKS

Bachelor's Thesis
Faculty of Information Technology and Communication Sciences
Examiner: Prof. Tuomas Virtanen
August 2021

ABSTRACT

Veera Kallio: Video-based Scene Classification Using Pretrained Neural Networks
Bachelor's Thesis
Tampere University
Electrical Engineering
August 2021

Neural network is a widely used machine learning method where a computer analyzes data and learns from it. There are many neural networks for data classification, and using existing models makes the classification faster and easier. This thesis investigates two well-known pretrained neural networks and their performance in video scene classification. The objective of this study is to find out how different models succeed in the scene classification task and whether there are differences between them, and to analyze which visual features are effective in classification.

The theoretical part of the study explains the principles of neural networks and in particular the convolutional neural networks, as well as introduces the neural networks used in the project, VGGNet and ResNet (Residual Neural Network). In addition, previous studies of the classification topic are discussed. It was observed that even though there is a lot of research on classification, video scene classification to event-type scenes specifically has been studied quite little and comparisons of different neural network models for this purpose were not found. In the experimental part, VGGNet and ResNet are used to classify video material that contains urban scenes from big European cities. Some of the material is used in training of the models and some in testing, so that the final performance can be evaluated. In addition, suitable values for the parameters of the models are searched in order to get the most optimal classification result.

The study shows that both VGGNet and ResNet perform well in the video scene classification, and there are no big differences between them. ResNet succeeded a little better in the task with 82.2 % accuracy, while the accuracy of VGGNet was 79.6 %. The results are consistent with previous studies. The most difficulties the neural networks had with telling apart very similar scenes, such as inside of tram and bus. Because of different features, the accuracies of different scene classes could vary even 40 percentage points. It can be concluded on the basis of the thesis, that pretrained neural networks can be utilized in video scene classification, and it is likely that also the results of other models than those used in this study would not differ from each other very much. However, the selection of an appropriate model slightly affects the accuracy of the classification, as well as the speed, and additionally the accuracy can be improved with parameter optimization.

Keywords: machine learning, convolutional neural network, video scene classification

TIIVISTELMÄ

Veera Kallio: Videoon perustuva maiseman luokittelu esiopetetuilla neuroverkoilla
Kandidaatintyö
Tampereen yliopisto
Sähkötekniikan tutkinto-ohjelma
Elokuu 2021

Neuroverkko on paljon käytetty koneoppimismenetelmä, jonka avulla tietokone voi analysoida dataa ja oppia siitä. Datan luokitteluun tarkoitettuja neuroverkkoja on olemassa paljon, ja valmiiden mallien käyttäminen nopeuttaa ja helpottaa luokittelua. Tässä työssä tutkitaan kahta tunnettua esiopetettua neuroverkkoa ja tarkastellaan niiden suorituskykyä videomaisemien luokittelussa. Työn tavoite on selvittää, miten eri mallit selviävät maisemanluokittelutehtävästä ja onko niiden välillä eroja, sekä mitkä visuaaliset piirteet auttavat luokittelussa.

Työn teoriaosuudessa selitetään neuroverkkojen ja erityisesti konvoluutioneuroverkkojen toimintaperiaate sekä esitellään tutkimuksessa käytetyt neuroverkot, VGGNet ja ResNet (engl. Residual Neural Network). Lisäksi kerrotaan luokitteluaiheesta tehdystä aiemmasta tutkimuksesta, jota onkin runsaasti, vaikkakin työssä havaittiin, että juuri videomaiseman luokittelua on tehty melko vähän, eikä tähän tarkoitukseen tehtyjä vertailuja eri neuroverkkomalleista löytynyt. Tutkimusosassa VGGNet:iä ja ResNet:iä käytetään luokittelemaan videoaineistoa, joka sisältää urbaaneja maisemia eurooppalaisista suurkaupungeista. Osaa aineistosta käytetään mallien opettamiseen ja osaa testaamiseen, jotta lopullista suorituskykyä voidaan arvioida. Lisäksi mallien parametreille etsitään sopivat arvot optimaalisimman luokittelutuloksen saamiseksi.

Tutkimus osoittaa, että sekä VGGNet että ResNet suoriutuvat videomaiseman luokittelusta hyvin, eikä niiden välillä ole paljon eroa. ResNet onnistui tehtävässä hieman paremmin 82,2 % tarkkuudella VGGNet:in tarkkuuden ollessa 79,6 %. Tulokset ovat linjassa aiempien tutkimusten kanssa. Eniten vaikeuksia neuroverkoilla oli hyvin samankaltaisten maisemien, kuten raitiovaunun ja bussin sisätilojen, erottamisessa toisistaan. Erilaisten piirteiden takia eri maisemaluokkien tarkkuudet erosivat toisistaan jopa 40 prosenttiyksikköä. Tutkimuksen perusteella voidaan päätellä, että esiopetettuja neuroverkkoja voidaan hyödyntää videomaiseman luokitteluun, ja on todennäköistä, että myöskään muiden kuin tässä työssä käytettyjen mallien tulokset eivät eroaisi toisistaan kovin paljon. Sopivan mallin valinnalla voidaan kuitenkin jonkin verran vaikuttaa luokittelun tarkkuuteen sekä myös nopeuteen, ja lisäksi tarkkuutta voidaan parantaa parametrien optimoinnilla.

Avainsanat: koneoppiminen, konvoluutioneuroverkko, videomaiseman luokittelu

CONTENTS

LIST OF SYMBOLS AND ABBREVIATIONS

Adam	adaptive moment estimation
ANN	artificial neural network
CNN	convolutional neural network
DNN	deep neural network
FC	fully connected
GPU	graphical processing unit
NN	neural network
ReLU	rectified linear unit
ResNet	residual neural network
SGD	stochastic gradient descent
SVM	support vector machine
VGG/VGGNet	convolutional neural network model developed by Visual Geometry Group of University of Oxford

1. INTRODUCTION

When humans are watching a video, we can usually easily identify what type of scene it is located in. For that we use our visual senses and our prior understanding of the world and similar scenes. Slightly similarly, a computer is not able to tell anything about a video if it has not learned about different scenes and their characteristics beforehand. Because it would be very time consuming, almost impossible, to teach by hand which visual features are relevant to which scenes and how to look for them, machine learning is used to let the computer learn the features from a training data by itself.

In classification, the choice of classifier is important as it determines how the learning is performed, which can naturally affect the success of the classification a lot. Some models are more general and perform well in various classification tasks while some work better in more specific tasks, and in addition, the speed of different classifiers can vary. This study examines two convolutional neural network based architectures, VGGNet [1] and ResNet (residual neural network) [2], and their performance in video-based scene classification. The objective is to classify a specific dataset of urban city scenes [3] and to compare the classifiers based on the two models, as well as to compare the classification of different scenes.

The thesis is structured as follows. The background of the convolutional neural networks behind the study and other publications related to scene classification or pretrained convolutional neural networks are discussed in Chapter 2. The models used in the study and their background are explained in Chapter 3. Chapter 4 covers the dataset used, optimization of the models, and the results of the experiments. Finally, Chapter 5 summarizes the study and discusses the meaning of the results as well as possible future work.

2. BACKGROUND

Machine learning is one of the fastest growing fields of technology at the moment. It has wide and constantly expanding range of applications from medical field to marketing. Machine learning is a branch of artificial intelligence where applications learn from data and improve their performance over time. That is obtained by finding patterns and features in the data and making predictions based on them. [4] Convolutional neural network is a machine learning method that has gained a lot of popularity in recent years, and it is a very powerful tool in classification.

Supervised classification is a machine learning task where a model of the distribution of class labels is built in terms of predictor features. This classifier is then used to predict the class label of the testing instances from their prediction features. [5] Scene classification is a core problem in understanding the visual content of a video or an image. It can be utilized in video categorization and retrieval, object detection, indoor positioning, etc. While it is relatively easy for a human to recognize the scene of a video, for a computer it is very difficult and nontrivial task. With machine learning and neural networks, it is possible to produce accurate and quick models which can analyze complex, bigger data. [6]

2.1 Convolutional neural networks (CNN)

Neural networks (NN) or artificial neural networks (ANN) are computing systems that are inspired by biological neurons in the human brain. They consist of layers of nodes called input, hidden and output layers. Each node, also called a perceptron, has a weight and the data is passed forward from one layer to the next one. If there are several hidden layers, the network is called deep neural network (DNN). Figure 2.1 illustrates the structure of a simple neural network where each node is connected to another.

Convolutional neural networks (CNN) are deep neural networks often used in computer vision. They have layers that utilize a mathematical operation called convolution, which is defined as [7]

$$S(i,j) = (K * I)(i,j) = \sum_m \sum_n I(i-m, j-n)K(m,n), \tag{2.1}$$

where I is a two-dimensional input image with indices i and j, and K is a kernel with

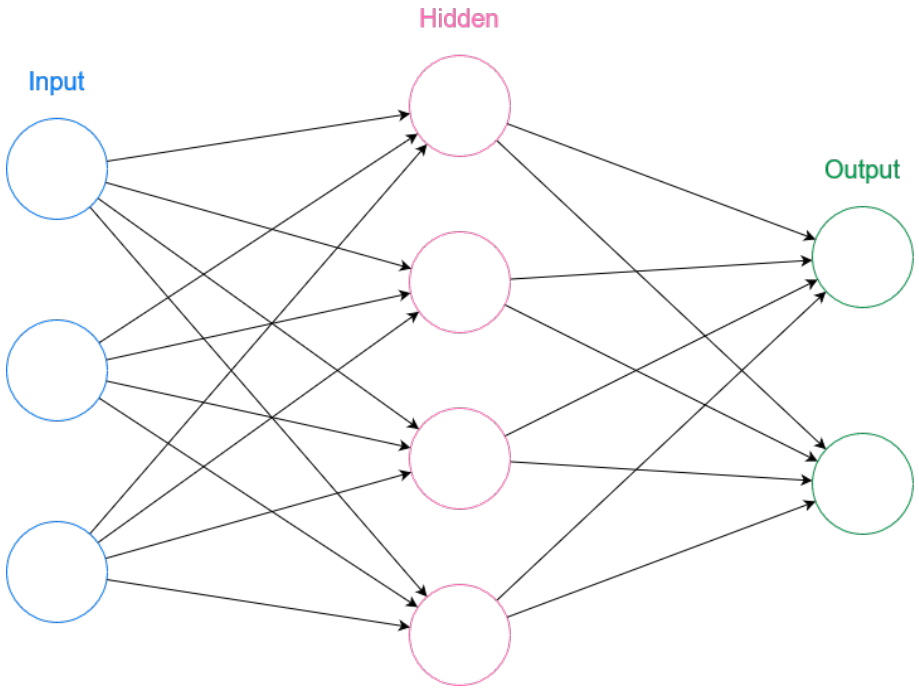

Figure 2.1. *An example of a simple neural network.*

indices m and n. The convolution operation is denoted with an asterisk and the output S is sometimes called a feature map. The kernel is a two-dimensional array of weights and it operates as a feature detector, sweeping across the image and finding certain features.

In addition to convolutional layers, CNNs have pooling layers to reduce dimensionality and complexity, and fully connected (FC) layers to perform the classification in the end based on the features extracted in the previous layers. After a convolution or pooling operation, an activation function, often a rectified linear unit (ReLU), is applied to create nonlinearity. FC layers usually use a softmax activation function, which gives a probability. [8]

In 1989, LeCun et al. designed a neural network based on backpropagation to recognize handwritten zip codes [9]. He continued his research and in 1998 LeCun et al. proposed the first formal CNN model, LeNet-5 [10]. At the time however, the development of CNN was limited by computing power and dataset scale constraints, but after the development of computer hardware and GPU supported CNNs, a number of CNN models and large-scale datasets have been introduced [11].

2.2 Related work

Scene classification is a common task in computer vision and different methods have been studied widely. Nowadays most of them are based on convolutional neural networks, although for instance support vector machines (SVM) can also be used. However, often scene classification is done for satellite images and not so much to event-type scenes like in this study. Different pretrained CNN models have also been compared in many papers

but apparently not for scene classification specifically.

In their article [3] *A Curated Dataset of Urban Scenes for Audio-Visual Scene Analysis* Wang et al. studied audio-visual scene classification for TAU Urban Audio-Visual Scenes 2021 dataset, which is used in this project. They were able to obtain significantly higher accuracies with their audio-visual method compared to using only one of the modalities. In their experiment, concatenating the audio and video embeddings in an early stage resulted in better performance.

Fu et al. [12] applied an attention mechanism after CNN for video scene classification in their study *A Novel Attention-based Neural Network for Video Scene Classification in Complex Background*. They were also able to increase the classification accuracy with this weight calculation method. In their article [13] *Scene Classification via a Gradient Boosting Random Convolutional Network Framework* Zhang et al. were first to propose a framework that combines several deep neural networks for scene classification. They applied their method to two satellite image datasets and again obtained state-of-the-art performance. In addition to low-level features, object detection can be used in scene classification. Li et al. [14] studied this in their paper *Objects as Attributes for Scene Classification*. Using object features alone or together with low-level features both gave great results.

In their article [15] *Optimizing Pretrained Convolutional Neural Networks for Tomato Leaf Disease Detection* Ahmad et al. studied the performance of pretrained VGG, ResNet, and an Inception V3 model for classification and identification of tomato leaf diseases. In their experiments, VGG had the worst and Inception V3 the best accuracies. They also managed to improve the accuracies significantly by using hyperparameter tuning, which is performed in this project as well. Szymak et al. [16] compared more different models, 15 in total, for underwater object classification in their study *The Effectiveness of Using a Pretrained Deep Learning Neural Networks for Object Classification in Underwater Video*. Also in their experiments the VGG performed the worst while ResNet was one of the best ones. The best model was a DenseNet-201 but it also had the longest training time, while ResNet-18, that was almost as accurate, was very fast.

3. METHODOLOGY

In the experiments two different state-of-the-art CNN models were used in order to determine how well they are able to classify a custom dataset of urban video scenes. The selected models are VGGNet and residual neural network (ResNet), and their pretrained implementations from Torchvision library of a Python machine learning framework PyTorch [17] are used. Using a pretrained model on a new problem instead of starting from scratch is called transfer learning. It can save a lot of time and improve the performance as it uses the knowledge that was learned from a related task. VGG and ResNet are introduced in the next two sections respectively, and after that the proposed model and implementation of the classification are explained.

3.1 VGGNet

Simonyan and Zisserman from Visual Geometry Group (VGG) of University of Oxford introduced VGGNet model or shortly VGG in 2014 in their article *Very Deep Convolutional Networks for Large-Scale Image Recognition* [1]. It was deeper than most leading CNN:s at the time and managed to outperform them, making it one of the state-of-the-art image classification models, and it is still widely used today even though better models have been introduced since.

VGG is a deep neural network having 11 to 19 weight layers depending on the configuration. Before VGG, one of the best models was AlexNet [18], which has 8 layers. VGG is very similar to it but is roughly twice as deep, which improves the accuracy. They managed to increase the number of layers without making the network too heavy by using a very small 3x3 filter size in the convolutions. One configuration also uses 1x1 filters, which adds nonlinearity. In addition to convolutional layers, the VGGNets have max-pooling layers and in the end three fully connected layers and a softmax activation function. The convolutional layers use ReLU activation functions, which the AlexNet had showed to be better than previously commonly used functions [18]. The configuration of VGG with 16 weight layers is shown in Figure 3.1. Other versions have 11, 13 or 19 layers. Later, batch normalization has also been applied to VGGNets to further improve the performance and to make them faster. This study uses a 16-layer VGG with batch normalization.

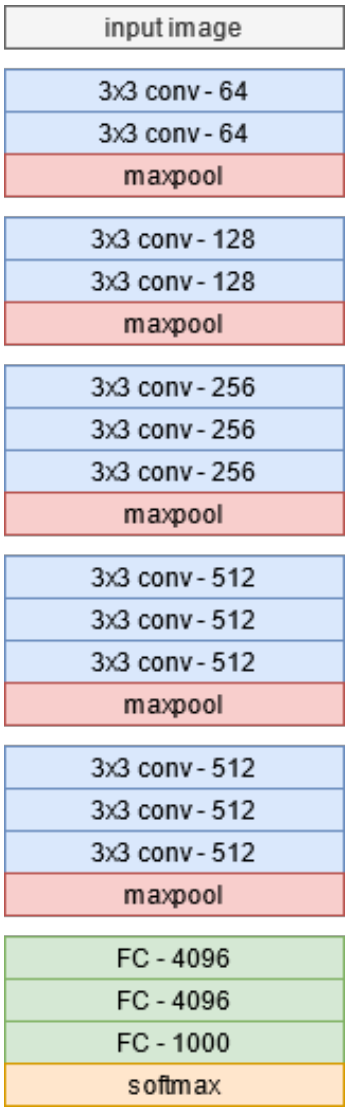

Figure 3.1. *16-layer VGG network architecture. [1]*

3.2 Residual neural network (ResNet)

Residual network (ResNet) was introduced a year after VGG, in 2015, by He et al. from Microsoft Research team in the paper *Deep Residual Learning for Image Recognition* [2]. The model shows clear improvement to VGG and it has won several competitions being one of the leading CNNs at the moment. ResNets are much deeper networks than VGGNets but still have lower complexity, which makes them both accurate and fast.

ResNet addresses the degradation problem in deep networks, which means that when more layers are added to the networks, at some point the accuracy stops increasing and rather decreases. The intuition would be that with more layers the layers learn progressively higher-level features and at some point the network is able to learn the data so well that the accuracies saturate, but a deeper network would always perform at least as well as a shallower one. However, experiments [2, 19] show that this is not the case. The

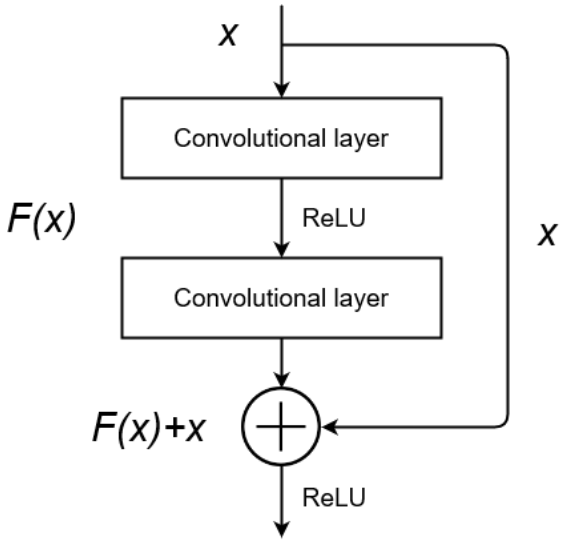

Figure 3.2. *A residual block. [2]*

degradation is not caused by overfitting, because that would mean that the network learns the training data so well that the testing error increases, but He et al. [2] showed that both the training error and testing error were higher with a 56-layer plain network compared to a 20-layer network.

ResNet alleviates the degradation problem by using residual learning, which can be implemented in neural networks by adding shortcut connections that skip one or more layers. A residual block is presented in Figure 3.2. The shortcut connection adds the input x to the original output mapping $F(x)$ by performing an identity mapping. ResNet has shown that deep networks that use this residual learning are easier to optimize and have better accuracy than their plain counterparts, and yet the shortcut connections do not add computational complexity. The errors of ResNets and corresponding plain networks were compared in [2], and with only 18 layers, there was not much difference between the plain network and ResNet, but when there were 34 layers, ResNet performed clearly better. The performance of ResNet increases when more layers are added, while it was decreasing with the plain networks.

The architecture of ResNets is simple and the baseline was inspired by VGG. First there are convolutional layers of different sizes stacked and in the end there is a global average pooling layer and one fully connected layer with softmax. The convolutional layers mostly use 3x3 filters as in the VGG. This is the plain network that was compared with the ResNet, and the residual version is obtained by adding the shortcuts. ResNet has shortcut connections every other layer and they always skip two layers. The most common ResNet versions have 18, 34, 50, 101 or 152 layers in total. In this study, a 50-layer ResNet is used.

3.3 Proposed model and classification

The VGGNet and ResNet models from Torchvision, that were pretrained on ImageNet data [20], are used largely as such, since the study focuses on comparing the pretrained implementations instead of developing the models. Only hyperparameter optimization, including tuning of the fully connected layers, is applied, and that is discussed in Chapter 4. The classification is implemented using PyTorch framework.

The models are trained for 50 epochs and for each epoch, the training is performed so that normalized video frames are shuffled and fed to the model in batches of 32 or 64. The model output is an array that contains a probability for each of the classes, for each of the images. Using the outputs and the ground truth class labels, cross-entropy loss is calculated. Categorical cross-entropy loss contains softmax activation and can be defined as [21]

$$L_{CE} = -\log(\frac{\exp(x[class])}{\sum_j^C \exp(x[j])}), \tag{3.1}$$

where x is the probability vector and C is the number of classes. The weights of the model are then updated using stochastic gradient descent (SGD) optimizer with decaying learning rate, momentum, and Nesterov's accelerated gradient. It can be written as [22]

$$v_{t+1} = \mu v_t - \varepsilon \nabla f(\theta_t + \mu v_t) \tag{3.2}$$

$$\theta_{t+1} = \theta_t + v_{t+1}, \tag{3.3}$$

where $f(\theta)$ is the function to be minimized, ε is the learning rate, μ is the momentum, and v is velocity. Only the fully connected layers are trained while the weights of all the other layers are freezed. This way, the CNN works as fixed feature extractor. Accuracy of each epoch is calculated using the class with maximum probability as the predicted class. After the training is completed, the weights that gave the best validation accuracy are saved.

The evaluation of the trained model is done so that the model is used for 10 frames of each testing video, and the probabilities of the output array are averaged in order to obtain a vector that contains one probability for each class. The maximum value is then selected to be the prediction of the video clip. In the end, the classification accuracy and loss is calculated for each scene class.

4. EXPERIMENTS

VGG and ResNet were trained and evaluated using TAU Urban Audio-Visual Scenes 2021 dataset to see if they can classify urban city scenes effectively. The dataset is introduced in the next section. After that, optimization of the models is discussed. Finally, the results of the experiments are presented.

4.1 Dataset

The TAU Urban Audio-Visual Scenes 2021 is a newly collected dataset that is publicly available [3]. It contains more than 12 thousand video and audio files from 12 large European cities: Amsterdam, Barcelona, Helsinki, Lisbon, London, Lyon, Madrid, Milan, Prague, Paris, Stockholm, and Vienna. This study uses a development dataset [23] which has only 10 of the cities, Amsterdam and Madrid excluded. Also the audio files were not used in this project. There are 10 scene classes: airport, indoor shopping mall, underground metro station, pedestrian street, public square, street with traffic, traveling by tram, bus and metro (underground), and urban park. For each scene class, there are videos from multiple different locations in all or most of the 10 cities. The videos are 10 seconds long, and in total there are 34 hours of material, about 3 hours per scene class. The number of different locations for each class are presented in Table 4.1.

The purpose of the dataset is to provide consistent-quality video and audio data that is carefully planned and recorded under controlled conditions, unlike most other datasets [3]. The data was collected by Tampere University of Technology between May and November 2018 with a specific setup. The videos were recorded using a GoPro Hero5 Session action camera and the corresponding audio with a Zoom F8 audio recorder. The camera was mounted to the strap of a backpack and the recording person was instructed not to move. This results in video containing objects moving across a static background, with no interference from body or hand movements. In each location, a couple of few minutes long sessions were captured in slightly different positions and then split into segments of 10 seconds length. Each video file is characterized by the scene class, city and location ID. Post-processing of the videos includes removing personal information and recording errors, and automatic blurring of faces and car licence plates.

The development dataset is split so that approximately 70% of the data are included in

Table 4.1. *Number of filming locations of each scene class in the development dataset. [3]*

Scene	Locations
Airport	29
Bus	53
Metro	53
Metro station	46
Park	40
Public square	41
Shopping mall	31
Street pedestrian	45
Street traffic	42
Tram	51
Total	431

a training set and 30% in a test set [3]. Files from same locations are included into the same subset. A validation set is also needed, and in this study, the training set is further split so that all files from Helsinki and Lisbon constitute the validation set and the others are left for training. This generates about 80/20 ratio between the training and validation sets and makes them unrelated to each other, while containing videos from all the scenes evenly.

In order to use the videos in the CNN models, they need to be extracted to image frames. Here only 4 frames per video are used, from timestamps 0s, 3s, 6s, and 9s. This makes the training faster and does not affect the accuracy much, since the videos are quite static and the content does not change a lot during the clip. For testing, 10 frames per video are used. For both training and testing, the frames are also resized from 720x1280 to 256x512 and normalized with a specific transform needed for pretrained ResNet and VGG.

4.2 Model optimization

In order to achieve the best possible performance of pretrained neural networks, hyperparameter optimization is needed. Hyperparameters are the training variables for which a value is set manually before the training, for instance learning rate and batch size. Model design components such as the number of layers and the optimizer are often considered hyperparameters as well. For simplicity, here they are referred to as hyperparameters. In this project, the optimization is done manually, but there are different methods to do it automatically, such as Bayesian optimization [24]. Manual optimization is done so that one parameter value is changed at a time, and if the validation accuracy is better than

Table 4.2. *Hyperparameter optimization. V=VGG, R=ResNet*

Parameter	Tested values	Selected value
FC layers	Different combinations of 1-3 layers with output sizes 4096, 2048, 1024, 512, and last layer output always 10	R: 3 layers with outputs 512 - 1024 - 10
Dropout	0.0, 0.1, 0.2, 0.3, 0.4, 0.5	R: 0.4
Optimizer	SGD+momentum, SGD+momentum+Nesterov, Adam, AdamW, Adam+AMSGrad	SGD+momentum+Nesterov
Batch size	32, 64	V: 32, R: 64
Learning rate	0.01, 0.001, 0.0005, 0.0001, 0.00001	V: 0.0001, R: 0.001
Step size	5, 7, 10	V: 10, R: 5

with the previous value, the new value is selected. Experimenting with different hyperparameters showed that changing a hyperparameter value can change the accuracy even several percentage points.

The optimization results are presented in Table 4.2. It shows all the tested hyperparameter values and the value that gave the best result and was selected. Different architectures of the fully connected layers were only tested for ResNet, since it originally has only one FC layer with input size 2048 and output 1000. That is replaced by 3 layers with output sizes 512, 1024, and 10, with dropout of 0.4 in between. The VGG model already has 3 layers, with input size 4096 and outputs of 4096, 4096, and 1000, and dropout of 0.5 in between. That is used as is and only the final layer output is changed to 10 as there are 10 classes.

Optimizer is the algorithm that changes the weights of the neural network in order to reduce the loss. Here, different variants of stochastic gradient descent (SGD) and Adam (adaptive moment estimation) optimizers were tested. SGD is simpler and more general, while Adam is faster and often performs the best, which is why it has become very popular. However, [25] suggests that because of poor generalization ability, the use of adaptive optimization methods, such as Adam, should be reconsidered. In their experiments, SGD performed better than adaptive methods, and the experiments of this project gave similar results. Even though Adam had better training accuracies and was faster, SGD had better validation performance. SGD with momentum of 0.9 and Nesterov's accelerated gradient method were chosen for both models, although some of the optimizers were tested only on ResNet.

Batch size of 32 was selected for VGG and 64 for ResNet. The initial learning rate for VGG was 0.0001 and it was reduced by 0.1 each 10 epochs. ResNet started from 0.001

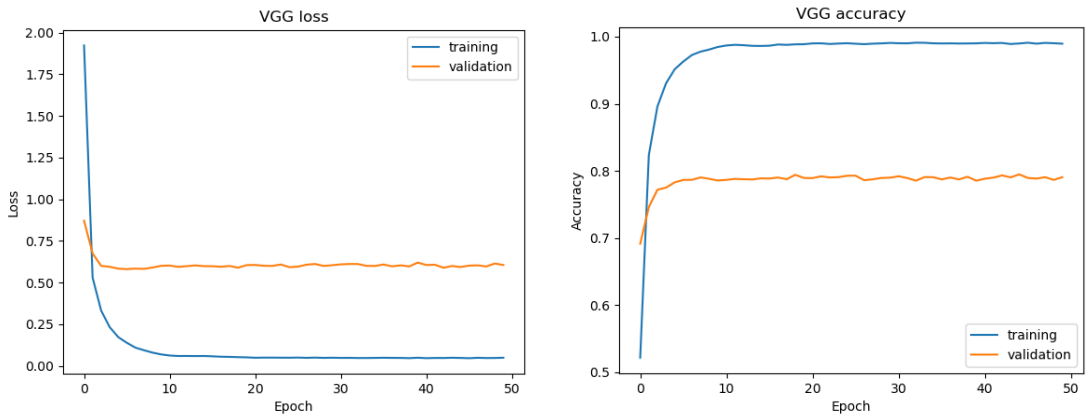

Figure 4.1. *Learning curves of the VGGNet model*

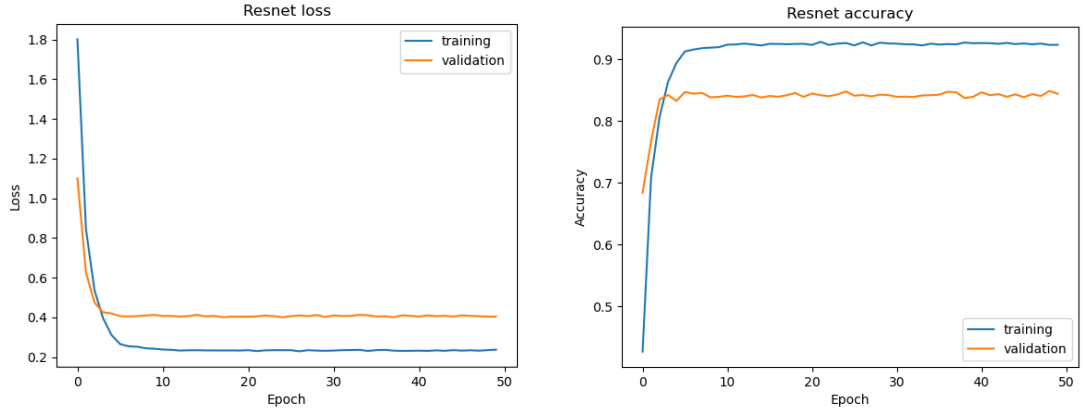

Figure 4.2. *Learning curves of the ResNet model*

and was decayed each 5 epochs. Learning rate should be reduced when the loss stops decreasing. Too high learning rate causes fluctuating as it jumps over minima and too low rate can take long to converge or get stuck in a local minimum.

4.3 Results

The learning curves of the optimized VGG and ResNet models are shown in Figures 4.1 and 4.2 respectively. The shapes of loss and accuracy are similar for both models, and they converge very fast. However, for VGG the difference between training and validation is bigger and the validation accuracy and loss do not reach as good values as with ResNet. This means that VGG learns the training data well but does not manage to generalize the model for new data that well. That is called overfitting, but it is not too severe here, because the validation loss does not start increasing. After the training, the weights that gave the best validation accuracy were saved for the model.

The evaluation of the trained models was performed as described earlier and the obtained results for each scene class are presented in Table 4.3. As the learning curves and previ-

Table 4.3. *Test accuracies per scene class*

Scene class	VGG accuracy (%)	ResNet accuracy (%)
Airport	64.1	68.7
Bus	93.6	90.2
Metro	79.7	90.0
Metro station	**95.3**	**96.6**
Park	87.3	86.5
Public square	78.3	67.4
Shopping mall	85.6	79.8
Street pedestrian	83.6	94.5
Street traffic	77.1	85.6
Tram	**51.6**	**63.0**
Total	79.6	82.2

ous study suggested, also the test accuracy was in average a little better for ResNet than VGG. For some scenes the VGG performed as well as Resnet or even better, but for some it was clearly worse. The best and worst classified scenes were still the same for both models, metro station being almost always correctly labeled, while tram is misclassified in almost half of the cases with VGG. The total test accuracy was close to the validation accuracy with VGG while ResNet performed slightly poorer in testing.

The reason for poor accuracy of tram is that it is often classified as a bus and sometimes as a metro, as they of course have quite similar features. Metro is also sometimes classified as a bus, more often with VGG, while ResNet tells them apart quite well. The metro station is quite different from other classes, as it is underground and often includes the metro, and that is why it has the best accuracy. Airport is sometimes misclassified as a metro station, though, or as a shopping mall, as they are all indoors and contain many people. Park is also quite unique as it usually has much greenery, but it is sometimes confused to a street with traffic, which can also have trees. Public square is often misclassified as a pedestrian street, especially with ResNet, as they are quite similar places and both have people. Pedestrian streets, streets with traffic and public squares are also sometimes confused with each other, as they are urban outdoor areas.

While experimenting with the hyperparameters, the testing gave corresponding accuracies for different scenes as with the final model, which means that the models learn the different classes similarly regardless of the parameters. This is expected, as only the final layers are trained, and the convolutional layers search for similar features every time. However, the very first test with ResNet with no parameter optimization gave total accuracy of 77 %, which compared to the 82 % accuracy of the optimized model, shows that the optimization does help the model to learn the features better, improving the per-

formance significantly. On the other hand, even without any optimization the accuracy is very high considering that the data is different than what the models were originally trained with, which further proves how useful it is to use pretrained models. They can be utilized easily without needing to spend a lot of time designing and optimizing the model.

5. CONCLUSION

Classification is a challenging but important task, and there are many different ways to do it with machine learning. This thesis studied classification with pretrained convolutional neural networks, using a video dataset of urban city scenes. The selected neural networks, VGGNet and ResNet, are popular and successful models, and on the basis of previous related work they were expected to be suitable for the scene classification task as well.

The experiments showed that both of the proposed models are a reasonable choice for the task, each reaching about 80 % classification accuracy. ResNet was slightly better, but there was little difference. However, there were big differences between the 10 categories, more distinct scenes being more accurately classified. This means that the features of more similar scenes were not learned well enough in order to tell them apart. Still none of the class accuracies were under 50 %, which proves that the CNN models are indeed useful for scene classification.

The experiments made could be extended to other pretrained models as well, to see if they give similar results and to better determine which models are the best. The parameter optimization could also be done more thoroughly and systematically with automated methods, which would likely increase the performance yet a little more, in addition to saving time. Different divisions of training and validation data should also be experimented with, and maybe use cross-validation. Furthermore, the motion information of the videos could be utilized, as here it was completely ignored and the video frames were processed individually as images. Probably even more useful would be to use the audio, which was also ignored. The audio contains additional information about the scene, and that could be utilized in the classification.

REFERENCES

[1] Simonyan, K. and Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *3rd International Conference on Learning Representations, ICLR 2015* (2015).

[2] He, K., Zhang, X., Ren, S. and Sun, J. Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

[3] Wang, S., Mesaros, A., Heittola, T. and Virtanen, T. A Curated Dataset of Urban Scenes for Audio-Visual Scene Analysis. *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2021. URL: https://arxiv.org/pdf/2011.00030.pdf.

[4] *Machine Learning*. IBM Cloud Learn Hub. July 15, 2020. URL: https://www.ibm.com/cloud/learn/machine-learning (visited on 03/17/2021).

[5] Kotsiantis, S. B., Zaharakis, I. D. and Pintelas, P. E. Machine Learning: a Review of Classification and Combining Techniques. *The Artificial intelligence review* 26.3 (2006), pp. 159–190. DOI: 10.1007/s10462-007-9052-3.

[6] Patel, H. and Mewada, H. Analysis of Machine Learning Based Scene Classification Algorithms and Quantitative Evaluation. *International Journal of Applied Engineering Research* 13.10 (2018), pp. 7811–7819. ISSN: 0973-4562. URL: https://www.ripublication.com/ijaer18/ijaerv13n10_74.pdf.

[7] Goodfellow, I., Bengio, Y. and Courville, A. *Deep Learning*. MIT Press, 2016. URL: http://www.deeplearningbook.org.

[8] *Convolutional Neural Networks*. IBM Cloud Learn Hub. Oct. 20, 2020. URL: https://www.ibm.com/cloud/learn/convolutional-neural-networks (visited on 03/24/2021).

[9] LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. and Jackel, L. D. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation* 1.4 (1989), pp. 541–551. ISSN: 0899-7667. DOI: 10.1162/neco.1989.1.4.541.

[10] LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P. Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324. ISSN: 1558-2256. DOI: 10.1109/5.726791.

[11] Liu, S., Tian, G. and Xu, Y. A Novel Scene Classification Model Combining ResNet Based Transfer Learning and Data Augmentation with a Filter. *Neurocomputing*

338 (2019), pp. 191–206. ISSN: 0925-2312. URL: https://www.sciencedirect.com/science/article/pii/S0925231219301833.

[12] Fu, Y., Xin, R. and Ye, O. A Novel Attention-Based Neural Network for Video Scene Classification in Complex Background. *Proceedings of the 32nd International Conference on Computer Animation and Social Agents*. CASA '19. 2019, pp. 85–88. ISBN: 9781450371599. DOI: 10.1145/3328756.3328768.

[13] Zhang, F., Du, B. and Zhang, L. Scene Classification via a Gradient Boosting Random Convolutional Network Framework. *IEEE Transactions on Geoscience and Remote Sensing* 54.3 (2016), pp. 1793–1802. DOI: 10.1109/TGRS.2015.2488681.

[14] Li, L.-J., Su, H., Lim, Y. and Fei-Fei, L. Objects as Attributes for Scene Classification. *Proceedings of the 11th European conference on Trends and Topics in Computer Vision*. Vol. 1. 2010, pp. 57–69. ISBN: 978-3-642-35748-0. URL: http://vision.stanford.edu/documents/LiSuLimFeiFei_ECCV2010.pdf.

[15] Ahmad, I., Hamid, M., Yousaf, S., Shah, S. T. and Ahmad, M. O. Optimizing Pretrained Convolutional Neural Networks for Tomato Leaf Disease Detection. *Complexity* 2020 (2020). ISSN: 1076-2787. URL: https://downloads.hindawi.com/journals/complexity/2020/8812019.pdf.

[16] Szymak, P., Piskur, P. and Naus, K. The Effectiveness of Using a Pretrained Deep Learning Neural Networks for Object Classification in Underwater Video. *Remote Sensing* 12 (2020). DOI: 10.3390/rs12183020. URL: https://res.mdpi.com/remotesensing/remotesensing-12-03020/article_deploy/remotesensing-12-03020.pdf.

[17] Paszke, A., Gross, S., Chintala, S. and Chanan, G. *PyTorch*. 2016. URL: https://pytorch.org/ (visited on 04/07/2021).

[18] Krizhevsky, A., Sutskever, I. and Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. *Communications of the ACM* 60 (2012), pp. 84–90. ISSN: 0001-0782. DOI: 10.1145/3065386.

[19] Srivastava, R. K., Greff, K. and Schmidhuber, J. Training Very Deep Networks. *Advances in Neural Information Processing Systems*. Vol. 28. 2015.

[20] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. and Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.

[21] *Cross-Entropy Loss*. 2019. URL: https://pytorch.org/docs/stable/generated/torch.nn.CrossEntropyLoss (visited on 06/12/2021).

[22] Sutskever, I., Martens, J., Dahl, G. and Hinton, G. On the Importance of Initialization and Momentum in Deep Learning. *Proceedings of the 30th International Conference on Machine Learning*. Vol. 28. 3. PMLR, 2013, pp. 1139–1147. URL: http://proceedings.mlr.press/v28/sutskever13.pdf.

[23] Mesaros, A., Heittola, T. and Virtanen, T. *TAU Urban Audio-Visual Scenes 2021, Development Dataset*. Zenodo, 2021. DOI: 10.5281/zenodo.4477542.

[24] Snoek, J., Larochelle, H. and Adams, R. P. Practical Bayesian Optimization of Machine Learning Algorithms. *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'12. 2012, pp. 2951–2959.

[25] Wilson, A. C., Roelofs, R., Stern, M., Srebro, N. and Recht, B. The Marginal Value of Adaptive Gradient Methods in Machine Learning. *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. 2017, pp. 4151–4161. ISBN: 9781510860964. URL: https://proceedings.neurips.cc/paper/2017/file/81b3833e2504647f9d794f7d7b9bf341-Paper.pdf.