FACIAL EMOTION RECOGNITION USING TRANSFER LEARNING

Abstract

Vision is one of the most important senses and we rely on vision quite a lot- from navigating in the physical world to recognising object to interpreting emotions and thus vision is huge part of our lives. Thanks to evolutionary changes we have trained ourselves with 540 million years worth of data and this same neuroscience could be applied to achieve artificial computer vision. Facial emotion recognition is an important topic in the fields of artificial intelligence and computer vision due to its significant academic and commercial potential. Using the same very concept we intend to study the "Facial Emotion Recognition System" as an application Convolution Neural Network. However, these networks may be difficult to train where the available amount of data is limited. The purpose is to classify images of peoples facial expressions into certain distinct emotion classes using transfer learning on a relatively small dataset. In transfer learning, we take the pre-trained weights of an already trained model(one that has been trained on millions of images belonging to 1000's of classes, on several high power GPU's for several days) and use these already learned features to predict new classes. We also intend to refer and study the past models in the same field starting from LeNet 5 - a 7 layered convolution network, AlexNet, VGG and ResNet-a 152 layered convolution networks architectures were pre-trained on ImageNet, a relatively large dataset containing 1.2 million images with 1000 categories.

Keywords: Computer vision, Convolution Neural Network, Transfer learning, LeNet 5, AlextNet, VGG and ResNet

1. Introduction

Facial Emotions are important factors when it comes to human communication and human-computer interaction. Computer vision is a way in which Recognition of facial emotions could be or rather have been automated (These facial emotions may many a times be difficult to perceive) . We humans interact among ourselves with the help of a universal language- Emotions. Emotions are beyond races, gender, and cultural diversity.It gives the mental state of a person that directly relates to his intentions or the physical efforts that he must be applying for performing tasks. FER is thus an important topic in the fields of artificial intelligence and computer vision due to its significant academic and commercial potential. Therefore it has high importance in fields such psychological studies, virtual reality, robotics and many more. Interest in this particular field has also been increased since the past decade due to the introduction and development of various artificial intelligence techniques such as Virtual Reality, Augmented Reality and Human- computer Interaction.
The purpose is to classify images of peoples facial expressions into certain distinct emotion classes using transfer learning on a relatively small dataset. In transfer learning, we take the pre-trained weights of an already trained model(one that has been trained on millions of images belonging to 1000's of classes, on several high power GPU's for several days) and use these already learned features to predict new classes.
Starting from the 1990s up until today, many models were proposed in the same field. In 1990s Yann LeCun, Leon Bottou, Yosuha Bengio and Patrick Haffner proposed a neural network architecture for handwritten and machine-printed character recognition which was called LeNet-5.
This was comparatively simple and straightforward and was a 5 layered Convolution Neural Network.
The introduction of CNN and RNN spawned an array of variations and improvements to these designs, such as AlexNet, VGG, and ResNet. These new models have been increasing in complexity and performance. Using CNN has now been proven to be a much more complexed task. Traditional machine learning methods are good for specific task with limited amount of data but when there is a whole lot of data, building models from scratch becomes a tedious task. And thus we use transfer learning that is utilising knowledge acquired for one task to solve related ones. AlexNet provides a good start

for Transfer Learning. Since many classes the net was trained on, there is an expectation that universal features had been learned by the network.

2.Related Work

A brief review on the related work done using new and recent methodologies to minimise losses and increase the accuracy of Facial Emotion recognition.
The first approach that we discuss used the local binary pattern algorithm for the purpose of classification of the grayscale images into six said classes (namely happiness, sadness, anger, fear, disgust and surprise). Happy, George and Routray before using this LBP for minor changes in facial expression, preprocessed the initial data using the Haar Cascade face detection algorithm. They used LBP and principal component analysis algorithm and training result was stored to classify features.
The second approach is the Boosted LBP approach which was used to extract the most discriminant LBP features and this particular approach was proposed by Shan et al. Databases like JAFFE, MMI and Con-Kanade databases were used for experimentation purposes. The Principal Component Analysis algorithm was also used for classification purposes. This approach also classifies Grayscale images like the first one and classifies the images into six different classes. The whole process is finally tested and classified using multi class SVM- classifiers.
The third approach, which was proposed by Arunugam who used Radial Basis Function Network to detect faces from images instead of using Hair Cascade Algorithm. Also, this approach classifies images into Three classes- happy, anger and Disgust.
All the above approaches used real time images clicked using camera instead of using pre clicked images.

3. Computer Vision

Vision is the most important senses and vision seems easy to us since we had 540 million years worth of data to train on and then including the bipedal movement and human language(100 thousand years of data). Since Vision being so important, by the 1960s, this led to the scientists wondering how the neuroscience could be applied to artificial intelligence. One of the most popular breakthrough in this particular field was by the two Harvard scientists- Hubel & Wiesel.
The experiment was pretty simple and it used the visual cortex of cats. They displayed the stimulus on the screen and identified particular patterns.
The answer to the question that what do computers actually “see" is simple. The computer sees images as nothing but numbers. Images are made up of pixels and these pixels could be represented as numbers (in case of grayscale images) or as 3D matrices (in case of RGB images).
There are basically two types of computer vision tasks- regression and classification. Regression is when output takes a continuous values where as classification is when the output takes a certain class label.

4. Deep Learning

Deep learning is a subset of machine learning that empowers a machine to imitate the workings of a human brain. It uses multi-layered artificial neural networks to create patterns by processing massive amounts of data. Deep learning is a data-oriented domain where the accuracy of the result is directly proportional to the size of dataset being used to train the model. The highly flexible architectures of the deep learning technique can learn directly from raw data and can increase their predictive accuracy when provided with more data.
Deep learning uses its multi-layered architecture to extract higher level features from the raw input provided to the model and reach the output. In deep learning every layer transforms its input into a more condensed and composite form. In deep learning the architectures can be built by using a greedy layer by layer mechanism. Deep learning approaches can be mainly classified into supervised, semi-supervised or partially supervised, unsupervised, and Reinforcement Learning (RL) or Deep RL (DRL).

Deep learning technique is mainly employed in the areas where the presence of a human expert is not feasible (navigation on Mars),problems with many possible outcomes (weather prediction),the problems where the solution needs to be adaptive (biometrics, personalization) or the scenarios where the problem size is too huge for our limited reasoning abilities (sentiment analysis, featuring personalized ads on Facebook). Due to its application in almost all the areas deep learning is often referred as a universal learning approach.

5. Feature Extraction:

Feature extraction is a dimensionality reduction technique in which the useful and interesting parts of an image are extracted and stored in a more condensed form of a feature map or a feature vector. One of the main reasons that provides deep learning an edge over its traditional machine learning counterparts is its more complex feature extraction or feature learning mechanism.

In a rule-based approach features are hand designed to produce an output , in traditional machine learning methods hand designed features are simply mapped to obtain output whereas in deep learning approach first simple features are extracted from the input image and condensed to form complex features which are then mapped to obtain the output. In deep learning feature extraction is an automatic process which is represented hierarchically in multiple levels.

6. Convolution Neural Network

6.1 Overview

In case of a fully connected neural network, for example, lets take a 2D image as an input. Here in, each neuron in the hidden layer of the fully connected neural network would be connected to tall neutron in the input layer. This way all the spatial arrangement of the input would be totally lost and thus its not very feasible to use this method. The solution here could be connecting the patches of the input represented as a 2D array to neutron in hidden layers. This way the neutron in the hidden layer only sees particular region.
This patchy operation of feature extraction is known as convolution. Thus it is safe to say that the Convolution Neural Network (CNN) has fewer parameters when compared to fully connected Neural networks.
His type of network structure was first proposed in 1988 and then again in 1990s when gradient based algorithms (which minimised an error criterion) were applied to CNNs which led to successful results for handwritten digit classification. CNNs are highly optimised in structure for processing 2D and 3D images.

6.2 CNN Architecture:

The CNN architecture consists of two main parts- the feature extractor and the classifier part. There are three main operations in CNN namely- Convolution, Non-Linearity and pooling.

1. Convolution: for every neutron in the hidden layer, its inputs are those neurons in the patch of the input layer it is connected to. We apply the matrix which goes through linear and non- linear activations to form output feature map, and then these output feature maps are combined accordingly and then an additive bias is given.
2. Non- Linearity: The next step is to add a non-linearity operation to the output since the data is highly non linear. This operation to be applied after every convolution. The most common operation used is the ReLU (Rectified Linear Unit) operation that performs a pixel by pixel operation and replaced all the negative values in the output by zero.
3. Pooling: The last key operation is Pooling. This operation is used to reduce dimensionality and preserve spatial invariance. The most popular technique is max pooling where we take the maximum value in the patch.

The above three operations mainly constitute as a part of the feature extraction part of the CNN architecture. The second part is that, that how are we actually performing the classification. The previous layers output high level features of the input data which could be fed into fully connected layers to perform the classification task.

The output of the fully connected layer in practice is a probability distribution function for an input image's membership over possible classes. A common way to do this is to use a function called softmax function where the output represents a categorical probability distribution.

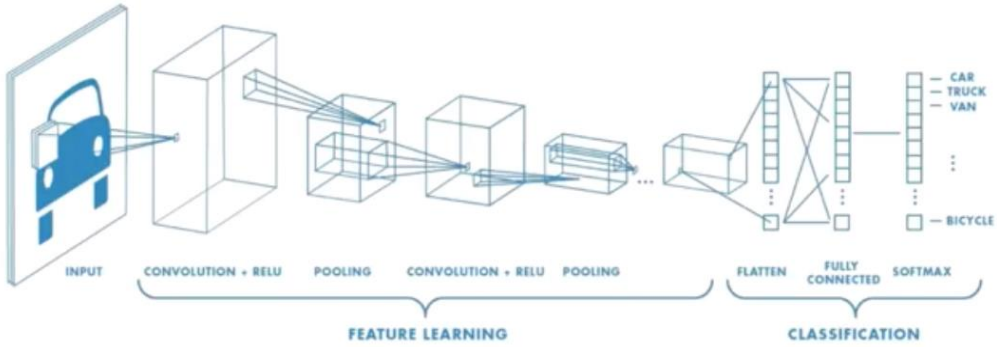

Fig:CNN architecture

7. POPULAR CNN ARCHITECTURES

We will now go through some of the popular CNN in order to get a better understanding of the Facial Emotion Recognition system and it's evolution through the past many years.
Some of the popular CNN architectures we'll study and compare are- AlexNet, VGG, Xception, ResNet and Inception.
These architecture mainly consists of stacks of several convolution layers, max pooling layers, sub sampling layers and soft max layers. The basic components of all these architectures is almost the same.

7.1 AlexNet (2012) :-

AlexNet was first proposed by Alex Krizhevesky, Ilya Sutskever, and Geoffrey E. Hinton with the publication of the paper ÏmageNet Classification and Deep Convolution Neural Networks"and this model was the winning entry in ILSVRC 2012.
Input: the input to the ALexNet is an RBG image of size 256x256. That means that all the images in the training as well as the target data set should be of size 256x256.
Architecture:
AlexNet consists of in total 8 layers- 5 convolution layers and 3 fully connected layers. The First convolution layer performs convolution and max pooling where 96 different filters of size 11x11 are used. The max pooling operation is performed using filters of size 5x5.
The second convolution layer performs the same operations of convolution and max pooling with 5x5 filters where as the third, fourth and the fifth convolution layers perform the same operations using 3x3 filters.
Then we have two fully connected layers followed by a Softmax layer at the end. Two new concepts used in this network are Local Response Normalisation and dropout.

7.2 VGGNET (2014) :-

The Visual Geometry Group (VGG), is a deep convolutional network for object recognition. It was developed and trained by Oxford's renowned Visual Geometry Group and was the runner up of the 2014 ILSVRC. It was specifically trained for a more than a million type of images with respect to "ImageNet" database. One of the significant additions of this model is-it demonstrates that depth of a network is a critical component to achieve better recognition and classification accuracy in CNNs.

VGG architecture comprises of two convolutional layers which uses ReLU activation function and are followed by a single max pooling layer and several fully connected layers which also comprises of ReLU activation function and the final layer of the model which is a Softmax layer is used for classification.

Three VGG models were proposed namely VGG-11, VGG-16, and VGG-19 which had 11,16, and 19 layers respectively. All the proposed models had three fully connected layers at the end with flexible convolution layers which varied according to the proposed model. VGG-11 consisted of 8 convolution layers, VGG-16 consisted of 13 convolution layers and VGG-19 which was computationally the most expensive model of all consisted of 16 convolution layers.

7.3 GoogLeNet (2014)

The model GoogLeNet was proposed by Christian Szegedy of Google in 2014 to reduce computation complexity when compared to other CNN models. This model was introduced to use Inception Layers that had variable receptive fields. These receptive fields created operations that captured sparse correlation patterns in the new feature map stack.
The concept of these inception layers in this model was to improve the recognition accuracy using a stack of such layers. These extra added 1x1 kernels allowed for dimensionality reduction. GoogLeNet consisted of 22 layers in total, which was far greater than any network before it. However, the number of network parameters GoogLeNet used was much lower than its predecessor AlexNet or VGG.

7.4 ResNet (2015) :-

Kaiming He, developed Residual Network (ResNet) with the aim of vanishing gradient problem that its predecessors had by employing ultra-deep networks in his model. ResNet was the winner of ILSVRC 2015. ResNet consists of a conventional feed forward network which comprises of a residual connection. A regular *ResNet* model has double- or triple- layer skips in its implementation that contain nonlinearities (ReLU) and batch normalization in between.

Unlike the conventional network architectures such as AlexNet and VGG, ResNet employs a complex and composite architecture that relies on micro module architecture which is also referred as network-in-network architectures (NiN). The micro module architecture attributes to the set of building blocks which are used to construct the network. Each residual or building block can have a different set of operations. Several evolved versions of ResNet have been proposed which constitutes of both Inception and Residual units. This improved version of the Inception-Residual network is known as PolyNet.

ResNet is a flexible model which was built to work on different number of layers ranging from 34, 50,101 to 152 and even 1202. One of the most popular model of ResNet is ResNet50 which comprises of 49 convolution layers and 1 fully connected layer at the network end. ResNet models consists of much more layers and are much deeper than VGG16 and VGG19 but, the model size is significantly smaller due to the usage of global average pooling instead of the fully-connected layers which helps in reducing the model size down to 102MB for ResNet50.

8. Comparison Study of the CNN models

Table 1: Factual comparison between few popular CNN Models

Models	**AlexNet**	**VGG**	**GoogLeNet**	**ResNet**
Proposed By	Alex Krizhevesky	Visual Geometry Group	Christian Szegedy	Kaiming He
Year	2012	2014	2015	2016
No. Of Conv layers	5	16	21	50
Input	227x227	224x224	224x224	224x224
No of Fully connected Layers	3	3	1	1
Total no. of layers	8	19	22	51
Size of Filters	3,5,11	3	1,3,5,7	1,3,7
Top 5 errors	15.3	7,3	6.67	5.3

9. Transfer Learning

9.1 Introduction: Transferring knowledge is a very common task among humans. Humans transfer knowledge in various ways to transfer knowledge between tasks. We humans recognise and then apply the relevant knowledge from the past learning experience and apply it to new tasks. Transfer learning here thus can be defined as a machine learning technique that focusses on storing knowledge that was gained by while solving a past problem and then applying this stored information or data to solve different relevant problems. Transfer learning basically involves freezing certain layers.

9.2 Why transfer learning?

There are many deep neural networks that are trained on natural images. In such cases, the first few layers have common tasks (like learning features similar to colour blobs or etc) irrespective of the dataset being used that is they perform general tasks. Thus these first larger features are said to be general.
In transfer learning we first train a base network on a base dataset and task, and then we "transfer" the learned features to a second target network to be trained on a target dataset and task. This process will tend to work if the features are general, that is, suitable to both base and target tasks, instead of being specific to the base task.

10. Comparison of some other CNN architectures through Experimental study

Table 2: Facts about some CNN architectures

	VGG16	**Inception v1**	**ResNet50**	**Xception**
Year	2014	2014	2015	2016
No. Of layers	13+3	22	50+1	71
Parameter	138M	5M	26M	23M
Default Size	224x224	299x299	224x224	229x229

Table 3: Experimental Results After Training the above models

Data Set	**Model**	**Loss Function**	**Training Accuracy**	**Validation/test**
FER2013	VGG16	Categorical cross Entropy	99.8	43.7
FER2013	Inception v1*	Categorical cross Entropy	98.9	50.3
FER2013	ResNet50*	Categorical cross Entropy	99.2	51.2
FER2013	Xception	Categorical cross Entropy	99.8	49.8

Note:
*While training these models there was some data losses.
The seed value was different while splitting of data

11. Output Justification:

VGG 16: The difference is accuracy is due to the problem of overfitting. Also, 43.7% accuracy is not to be expected when using real world entities unless the images are pre-processed.

12. Conclusion

In this survey, we have studied the facial emotion recognition system in details by understanding all the necessary basic concepts. We have also briefly explained Deep learning and its significant in Facial emotion recognition. We have studied the step by step procedure involved in the Convolution Neural Network and how different CNN architectures evolved and improved the Facial Emotion Recognition System in terms of minimising the losses and increase in accuracy. The problem of training all such models from scratch, which is a tedious and a complex task, can be solved using the concept of Transfer learning. Lastly, we have also done an experimental study by training the models and comparing the outcomes.

13. References

1] Soad Almabdy , Lamiaa Elrefaei : Deep Convolutional Neural Network-Based Approaches for Face Recognition

2] Deep Learning For Beginners Using Transfer Learning In Keras URL: https://towardsdatascience.com/keras-transfer-learning-for-beginners-6c9b8b7143e

3] Dive into. Deep Learning URL: https://d2l.ai/chapter_preface/index.html

4] Nithya Roopa. S : Emotion Recognition from Facial Expression using Deep Learning

5] OM Parkhi, A Vedaldi, A Zisserman : Deep face recognition

6] Vincent Fung. An Overview of ResNet and its Variants.
https://towardsdatascience.com/an-overview-of-resnet-and-its-variants-5281e2f56035

7] Bernard Marr.What Is Deep Learning AI? A Simple Guide With 8 Practical Examples.
https://www.forbes.com/sites/bernardmarr/2018/10/01/what-is-deep-learning-ai-a-simple-guide-with-8-practical-examples/#6d5e5acf8d4b

8] He Jun ; Li Shuai ; Shen Jinming ; Liu Yue ; Wang Jingwei ; Jin Pen. Facial Expression Recognition Based on VGGNet Convolutional Neural Network Published in: 2018 Chinese Automation Congress (CAC)

9] Suresh Dara ; Priyanka Tumma. Feature Extraction By Using Deep Learning: A Survey
Published in: 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)

10] A Comprehensive Guide to Convolutional Neural Networks — the ELI5 way URL: https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

11] Understanding AlexNet URL: https://www.learnopencv.com/understanding-alexnet/

12] David Orozco1, Christopher Lee2, Yevgeny Arabadzhi3, and Deval Gupta : Transfer learning for Facial Expression Recognition

13] Dhwani Mehta, Mohammad Faridul Haque Siddiqui, and Ahmad Y. Javaid : Facial Emotion Recognition: A Survey and Real-World User Experiences in Mixed Reality

14] Transfer Learning Introduction URL: https://www.hackerearth.com/practice/machine-learning/transfer-learning/transfer-learning-intro/tutorial/

15] Byoung Chul Ko : A Brief Review of Facial Emotion Recognition Based on Visual Information

16] Yingying Wang [1] , Yibin Li [1,*], Yong Song [2] and Xuewen Rong : Facial Expression Recognition Based on Auxiliary Models