4-23-2018

Emotion Classification based on Expressions and Body Language using Convolutional Neural Networks

Aasimah S. Tanveer
Lake Forest College, tanveeras@lakeforest.edu

Follow this and additional works at: https://publications.lakeforest.edu/seniortheses

Part of the Neuroscience and Neurobiology Commons

Recommended Citation
Emotion Classification based on Expressions and Body Language using Convolutional Neural Networks

Abstract
Recognizing human emotions based on a person's body language is a complex neurological process that humans often take for granted. The visual pathway that propagates from the eye to the occipital lobe, breaking down images from basic features and building upon the previous layer, can be replicated in a computer using artificially intelligent algorithms in the field of computer vision using convolutional neural networks (CNN).

The neural network, just like in the brain, models the interaction between neurons, with each neuron being represented by a mathematical function called a perceptron. Real-world images of humans in varying environments displaying emotion through variations in expression, body language, colorings, and features are manually labeled and used to train the CNN. This allows the algorithm to pick out features that occur in pictures of each emotion that will intelligently aid in relating body language, expressions and poses to an emotion, even if this relation is not defined previously. The accuracy of the classification is then analyzed by reviewing what features were extracted from the images and the network is retrained accordingly to output even more accurate results.

Document Type
Thesis

Degree Name
Bachelor of Arts (BA)

Department or Program
Mathematics and Computer Science

Second Department or Program
Neuroscience

First Advisor
Sugata Banerji

Second Advisor
Matthew R. Kelley

Third Advisor
Craig D. Knuckles

Subject Categories
Neuroscience and Neurobiology

This thesis is available at Lake Forest College Publications: https://publications.lakeforest.edu/seniortheses/144
Lake Forest College Archives

Your thesis will be deposited in the Lake Forest College Archives and the College's online digital repository, Lake Forest College Publications. This agreement grants Lake Forest College the non-exclusive right to distribute your thesis to researchers and over the Internet and make it part of the Lake Forest College Publications site. You warrant:

• that you have the full power and authority to make this agreement;
• that you retain literary property rights (the copyright) to your work. Current U.S. law stipulates that you will retain these rights for your lifetime plus 70 years, at which point your thesis will enter common domain;
• that for as long you as you retain literary property rights, no one may sell your thesis without your permission;
• that the College will catalog, preserve, and provide access to your thesis;
• that the thesis does not infringe any copyright, nor violate any proprietary rights, nor contain any libelous matter, nor invade the privacy of any person or third party;
• If you request that your thesis be placed under embargo, approval from your thesis chairperson is required.

By signing below, you indicate that you have read, understand, and agree to the statements above.

Printed Name: Aasimah S. Tanveer

Thesis Title: Emotion Classification based on Expressions and Body Language using Convolutional Neural Networks

This thesis is available at Lake Forest College Publications: https://publications.lakeforest.edu/seniortheses/144
LAKE FOREST COLLEGE

Senior Thesis

Emotion Classification based on Expressions and Body Language using Convolutional Neural Networks

by

Aasimah S. Tanveer

April 23, 2018

The report of the investigation undertaken as a Senior Thesis, to carry two courses of credit in the Department of Mathematics and Computer Science and the Neuroscience Program

_______________________
Michael T. Orr
Krebs Provost and Dean of the Faculty

_______________________
Sugata Banerji, Chairperson

_______________________
Matthew R. Kelley

_______________________
Craig D. Knuckles
ABSTRACT

Recognizing human emotions based on a person's body language is a complex neurological process that humans often take for granted. The visual pathway that propagates from the eye to the occipital lobe, breaking down images from basic features and building upon the previous layer, can be replicated in a computer using artificially intelligent algorithms in the field of computer vision using convolutional neural networks (CNN). The neural network, just like in the brain, models the interaction between neurons, with each neuron being represented by a mathematical function called a perceptron. Real-world images of humans in varying environments displaying emotion through variations in expression, body language, colorings, and features are manually labeled and used to train the CNN. This allows the algorithm to pick out features that occur in pictures of each emotion that will intelligently aid in relating body language, expressions and poses to an emotion, even if this relation is not defined previously. The accuracy of the classification is then analyzed by reviewing what features were extracted from the images and the network is retrained accordingly to output even more accurate results.
DEDICATION

For my family
ACKNOWLEDGMENTS

I would like to thank my advisor Dr. Sugata Banerji for allowing me to be a part of such an interesting project and teaching me everything I know about artificial intelligence. His constant insight, flexibility, and instruction was integral for the completion of this project. This project would not have come to fruition without him.

I would also like to thank Dr. Matt Kelley and Dr. Craig Knuckles for sitting on the defense committee for this thesis. Their guidance and revisions were great feedback in this endeavor. A special thank you to Dr. Naomi Wentworth for providing me with in-depth knowledge about the human visual system and providing resources to have the correct background information.

Thank you to my family for supporting me throughout my four years that have culminated in this work. They encouraged me to pursue such an interdisciplinary and custom field by providing both financial and emotional support. The invigorating discussions at home, passion for information technology, and push to strive for the best was the greatest motivator in pursuing this complex topic.

I would also like to thank my wonderful support system of friends: Logan Graham, Rosemary Thomas, Mohini Verma, Catie Brown, Rebecca Lynch, Krista Meuli, and Hamza Mahmood. These individuals, among so many others, provided the laughs, fun, and help that carried me through this thesis.

Lastly, I would like to thank Lake Forest College as well as Cory Stevens for advising my thesis.
Table of Contents

List of Figures ................................................. v

List of Tables .................................................. vii

1 Introduction .................................................... 1
2 Related Work ................................................. 1
3 Human Vision and Perception ................................. 3
  3.1 Eyeing the Visual System ................................. 3
  3.2 The Human Neuron ......................................... 11
4 Artificial Neural Networks .................................. 12
  4.1 The Artificial Neuron .................................... 12
  4.2 Artificial Neural Networks .............................. 14
  4.3 Convolutional Neural Networks ......................... 16
5 Proposed Method ............................................. 19
  5.1 Network Structure ....................................... 21
6 Experiments ................................................... 21
  6.1 Dataset ................................................... 21
7 Results ......................................................... 24
  7.1 Classification ............................................. 24
  7.2 Overfitting ............................................... 25
  7.3 Discussion ............................................... 26
8 Conclusions .................................................. 32

References ...................................................... 33
## List of Figures


2. Field of view and visual acuity.

3. A section of the human brain showing the visual cortices. Image source: McGill University.


5. A perceptron with a step activation function.

6. A fully-connected neural network with 3 input features, two hidden layers and four outputs.

7. A convolution filter of size $5 \times 5 \times 3$ being used on an image of size $32 \times 32 \times 3$.


9. A $2 \times 2$ maxpooling layer.

10. The dropout layer randomly disables about half of the neurons to improve generalization performance. Image source: https://leonardoaraujosantos.gitbooks.io/artificial-inteligence/content/dropout_layer.html.

11. A schematic representation of the CNN used in this work.

12. Sample images from the six emotion categories of our dataset.

13. The average of automatically detected faces from images of the six emotion classes.

14. The confusion matrix for emotion classification. The rows show the real emotions and the columns show the assigned emotions. The numbers indicate the classification accuracy for the six classes.
An example demonstrating the effect of overfitting where different crops of the same image were used as positive training samples for two different classes and later used for testing.

An input image and sample responses from some neurons of stages 1 to 5. The yellow areas show high activation while blue areas show little or no activation. The size of the activation area increases as we go deeper into the network.

The most responsive neurons in stage 5 for the six classes. The bar graphs plot the number of images where each particular neuron in stage 5 has had a high response for that class.

Patches extracted from each of the six emotion classes by the neurons most active for that class.

Some sample images from the dataset which the proposed system has difficulty classifying properly.
List of Tables

1. The structure of the convolutional neural network used for this work. 20
2. Class names and number of images in each class 23
1 Introduction

When my mom walks into my room with her hands on her hips and her eyebrows furrowed, I know that I’m in big trouble. Can a computer view this image and draw the same conclusion? Recognizing emotions from body language and poses is a remarkable feat of the human brain that we take for granted. Until a few years ago, it would have been impossible to create a computer capable of such a complex task. Today, artificial neural networks are being employed to enable computers solve many learning and recognition problems. Convolutional neural networks (CNNs) are a special class of deep artificial neural networks that have been found to be particularly effective at solving image classification problems. CNNs are can be designed to be similar to the human vision system in some respects, but are they able to solve this very human problem of emotion recognition from images? In this thesis, I have attempted to answer this question and I have detailed my successes and failures.

The remainder of this thesis is organized as follows; The work of other researchers on emotion recognition and CNNs is outlined in Section 2. The human vision system is briefly discussed in Section 3. Artificial neural networks and CNNs are described in Section 4. The experimental method proposed is described in Section 5. The experiments are described in Section 6. The results are discussed in detail in Section 7, along with a discussion of the findings. Finally, the Conclusions section lists all the work done in this thesis, and the future directions this research could move in.

2 Related Work

The problem of human emotion recognition has been generating interest for almost two decades [10]. The complexity of the tasks addressed by researchers has increased in complexity over the years. In the early years, the datasets used
were created by photographing a limited number of subjects in a controlled environment. One of the most significant datasets used for this purpose was the JAFFE Dataset [4]. More recently, uncontrolled images have been considered for emotion recognition. However, these datasets have been mostly limited to facial images for emotion recognition. Very recently, some work has been done on body language and pose estimation, but that uses input from 500 video cameras and not single still images [3].

In the past few years, the Computer Vision community has seen increased activity in the use of convolutional neural networks (CNNs) for many different classical problems. The initial breakthroughs were made possible by the availability of large labeled datasets (ImageNet [6], Places [17]) yielding great improvements on the object and scene classification tasks [11]. Since this initial success several strategies have been explored to adapt the network parameters or architecture to other tasks [7]. Typical convolutional neural networks used for categorization tasks are often concatenations of multiple convolution and pooling layers followed by two or three fully connected layers and a soft-max
classifier. Deep convolutional neural networks have also been shown to be good for feature extraction which can be used with different classifiers [1].

This thesis attempts to use a convolutional neural network for emotion recognition. However, it uses images downloaded from the Internet for this purpose. The goal of this work is twofold: first, to attempt to solve the emotion classification problem by training a CNN on images taken in the wild, and second, to try and gain insight on the internal representations and the emotion cues detected by the CNN during this process.

3 Human Vision and Perception

Before delving into how we represent human vision in a computer, it is important to understand the system we are attempting to model. The path to vision begins with the eye, a powerful and beautifully evolved organ that inputs, filters, and transfers light to the brain for processing. The eye must sense the presence and amount of light in its environment and focus the light in an optical array. The eye measures the brightness and color in each image and convert it into neural signals that can be sent to the brain for processing. In the brain, the signals are transferred from major visual structures to compose a visual scene allowing for interpretation of complex ideas, such as emotion to occur.

3.1 Eyeing the Visual System

The human eye is a spherical organ with a diameter of about 24 millimeters. Light first passes through a transparent membrane at the front of the eye known as the cornea which refracts the light. The iris, a muscle with an opening known as the pupil in the center, adjusts in size to control for the amount of light that passes into the eye in response to the brightness of the scene. In darker lit scenes, the pupil is dilated and contracted in bright scenes automatically
through its pupillary reflex. The pupil can also be manipulated voluntarily to focus attention on certain elements of a scene. This manipulation will influence how an image is perceived and which parts are blurred, less discernable which can be filtered out. Light then passes through the lens which is a transparent structure that further refracts light so that it can be focused on the innermost membrane of the eye. The power of the lens to refract light determines its focal length, which is the distance from the lens at which the image of an object is in focus when an object is far away. The angle of the lens can be manipulated with fibers, such as ciliary and zonule fibers, to better focus light, known as accommodation.

The eye can be divided into three chambers: the anterior chamber that is between the cornea and iris, the posterior chamber which is between the iris and the lens, and the vitreous chamber in the main interior portion of the eye. The anterior and posterior chambers contain aqueous humor and the vitreous chamber contains vitreous humor, humor being a clear, thin fluid. These humors also have a slight role in refracting light passing through the eye. Additionally, the eye is encased in three membranes from outer to inner, the sclera, choroid, and retina. The sclera is a tough, protective membrane that is the white portion of the eye. The choroid lines the interior of the sclera and contains the highest amount of the blood vessels that supply oxygen and nutrients to the inside of the eye. The retina is the inner most layer which is made of neurons and neuronal receptors. This is where light passing into the eye is converted to neuronal signals and passed onto the brain. Light processed from the cornea and focused on the retina for interpretation.

Because the purpose of the visual system is to eventually form a clear image of a scene and properly perceive it in the brain, the retina is an integral part of converting light into a signal that the brain can interpret, neural signals. Light
Fig. 2. Field of view and visual acuity

passing through the structures in the eye eventually hits the retina to form a retinal image which is inverted. The eye can be divided further into three main layers of the retina: the outer, inner, and ganglion cell nuclear layer. Each of the nuclear layers are separated by two synaptic layers, inner and outer, where the neurons in the retina synapse each other. Towards the back of the eye is a layer of inner and outer segments known as photoreceptors which are retinal neurons that transduce light into neural signals and have two classes: rods, cones, and photosensitive retinal ganglion cells. The outer layer contains a portion of the rods and cones while the inner layer contains various neuronal cells such as bipolar cells, which receive signals from photoreceptors and pass the signal onto amacrine cells, which pass signals back to bipolar cells and other amacrine cells, and horizontal cells that receive and send signals to rods and cones as well between themselves. The ganglion layer is the interface between the eye and the brain consisting of retinal ganglion cells (RGCs) that synapse among receptors and neurons then exit their axons, or projections from the cell body, at the optic disk which is also known as the "blind spot". The axons bundled together form
the optic nerve leading into the brain. This organization was first discovered by Masland [13].

The input of the visual system is pulled from the field of vision which allows for high-resolution depth-perception at a 110 deg angle in directly in front aspect of the visual field. Focus is a very important property of vision that manipulates how the scene is perceived. The place at which the light from objects at the center of vision and is focuses is called the fovea. There are no rods present here but a high density of cones. The eyes see objects in a scene at various level of fine detail, or acuity. Visual acuity is depended on certain optical and neural factors such sharpness of retinal focus, ability to which the retina functions properly, and sensitivity of interpretation in the brain [5]. It is most focused directly in front of the eyes and declines towards the periphery of the vision in an inverse-linear, hyperbolic fashion. This is shown in Figure 2. Depending on how light is refracted on the retina, there is a variation on how the image is seen and filtered in the brain. There are three main forms of acuity: spatial, temporal, and spectral. Spatial acuity is a function of location and brightness as the ability to resolve two points in space. Acuity is constant the periphery but much lower at the fovea while as brightness increases, it is harder to distinguish the edges of a gap. Temporal acuity is the ability to discriminate visual events in time. However, as later mentioned, our dataset used only static images, thus, will not account for this property. Lastly, spectral acuity distinguishes differences in wavelength stimuli, in other words, colors. Scenes with red, longer wavelength light require higher visual acuity than the shorter, blue wavelengths. Errors like ametropia, myopia, and hyperopia can interfere with the image on the eye side while disorders like retinal detachment and macular degeneration can change how the image is perceived on a neural level [16].
The primary visual cortex, often called the V1, is found in the occipital lobe in both cerebral hemispheres (Figure 3). It surrounds and extends into a deep sulcus called the calcarine sulcus. The primary visual cortex makes up a small portion of the visible surface of the cortex in the occipital lobe, but because it stretches into the calcarine sulcus, it makes up a significant portion of cortical surface overall. Visual information leaves the retina via the optic nerve and sent to a nucleus of the thalamus, known as the lateral geniculate nucleus (LGN) which specialized layers particularly the parvocellular cell layer, which is important for spatial resolution, visual acuity, and the detailed analysis of shape, size, and color, as well as the koniocellular cellular layer who are thought to be involved with some aspects of color vision. The LGN is a relay center for the visual pathway as
the main central connection between the optic nerve and occipital lobe. Neurons of the LGN send their axons through the optic radiation, a direct pathway to the primary visual cortex. The LGN allows for both temporal correlations as well as spatial correlations and these signals are important to achieve a three-dimensional representation of an object in space. The output from the LGN is a signal that provides a way to control where objects in space converge. Another signal is provided to focus the eyes per the distance of the object and more importantly anatomical computations are achieved to determine the position of every major element in object space relative to the principle plane. Through subsequent motion of the eyes, a larger stereoscopic mapping of the visual field is achieved. The LGN also has a very important role in focusing attention on the most important portion of a scene via its peri-reticular nucleus.

A parallel circuit stop on the visual pathway is the superior colliculi which is located in the midbrain as part of the tectum. Neurons here receive direct input from the retina as well and respond only to visual stimuli. The superior colliculi directs behavioral responses towards specific points in space and each layer contains a topographic map of the surrounding world in retinotopic coordinates, and activation of neurons at a particular point in the map evokes a response directed toward the corresponding point in space. This is done by directing eye movements to the visual stimulus as well as particular points of attention in a scene. While the superior colliculi is not directly implicated in object recognition, it aids the human brain to call attention to specific objects and modulate a behavior response based on these stimuli.

Neurons from the LGN carry a signal in a tract often called the optic radiation, which curves around the wall of the lateral ventricle in each cerebral hemisphere and reaches back to the occipital lobe. The axons included in the optic radiation terminate in the primary visual cortex in what is called a retino-
topic manner, meaning that axons carrying information from a specific part of the visual field terminate in a location in V1 that corresponds to that location in the visual field. Neurons travel on three distinct pathways from the LGN, magnocellular and parvocellular were already mentioned above, and the third being the koniocellular. These different types of neurons preferentially respond to different types of visual stimuli, thus it seems these pathways are each somewhat specialized for specific categories of stimuli.

Areas surrounding the primary visual cortex of the occipital lobe, called the extra striate cortex functions to pass information from using the dorsal and ventral pathways to the primary visual cortex per Milner and Goodale. The ventral stream is particularly of interest due to it being colloquially known as the “what” pathway which is involved in object and visual identification and recognition. The pathway is sensitive to high spatial frequencies such as details, stores long term visual representations and is a very allocentric frame of reference. This is an intense form of visual processing that requires a high level of consciousness and is very foveal in that focus is paid attention to. It has strong connections to the medial temporal lobe which stores long-term memories, the limbic system which controls emotions, and the dorsal stream which deals with object locations and motion. The ventral stream gets its main input from the parvocellular layer of the LGN of the thalamus. These neurons project to V1 sublayers and then from there, the ventral pathway goes through V2 and V4 to areas of the inferior temporal lobe which include the PIT (posterior inferotemporal), CIT (central inferotemporal), and AIT (anterior inferotemporal). Each visual area contains a full representation of visual space. That is, it contains neurons whose receptive fields together represent the entire visual field. Visual information enters the ventral stream through the primary visual cortex and travels through the rest of the areas in sequence. Moving along the stream from V1 to AIT, receptive
fields increase their size, latency, and the complexity of their tuning. All the areas in the ventral stream are influenced by extraretinal factors in addition to the nature of the stimulus in their receptive field. These factors include attention, working memory, and stimulus salience. Thus, the ventral stream does not merely provide a description of the elements in the visual world. It also plays a crucial role in judging the significance of these elements. These elements are further illustrated by the observation that damage to the ventral stream can cause inability to recognize faces or interpret facial expression.

The primary visual cortex or the V1 is highly specialized for processing information about static and moving objects and is excellent in pattern recognition. The visual cortex is organized in distinct layers discovered by Hubel and Wiesel [9] and each layer functions to form a complete picture of the scene. It organizes spatial information in vision and transforms a retinotopic map into a full visual image. The visual information relayed to V1 is not coded in terms of spatial imagery but rather described as edge detection. As an example, for an image comprising half black on one side and half white on the other, the dividing line between the black and the white has the strongest local contrast and is encoded, while few neurons code the brightness information. As information is further relayed to subsequent visual areas, it is coded as increasingly non-local frequency/phase signals. Note that, at these early stages of cortical visual processing, spatial location of visual information is well preserved amid the local contrast encoding. Edge detection aims to identify edges by looking at image brightness that contrasts sharply. It is a simple step into comprising a picture or scene along with contrasting lines, known as ridge detection. Edge detection is affected by factors such as focal blur, penumbral blur which is caused by shadows, and shading of smooth and rough objects.
3.2 The Human Neuron

A neuron is a cell that can be excited that receives, processes, and transmits information via electrical and chemical signals. The human neuron works as an input-output system. A neuron receives inputs which either signal excitation or inhibition. If an excitatory response is received, the signal will continue to the cell body then to the axon hillock where an action potential will begin. An action potential is a change in membrane potential that allows for an electrical signal to propagate along a neuron’s axon. Interplay between ions allow for this change in potential to occur. The signal reaches the axon terminals where neurotransmitters, chemical signals, can be released into the synapse and go onto carry the signal to the next neuron. A hallmark characteristic of a neuron is that it carries an all-or-nothing response. A set threshold must be overcome, or the signal will not go through at all. Neurons interacting with one another is what constitutes a neural network. A human neuron is shown in Figure 4.

![Fig. 4. A human neuron. Image source: Wikipedia.](image-url)
4 Artificial Neural Networks

4.1 The Artificial Neuron

A computer counterpart of the neuron is the artificial neuron, also known as the perceptron [14]. Just as a biological neuron takes an electrical input at its dendrites and creates an action potential which yields a response that propagates to another neuron, the perceptron takes numeric input and propagates a response to other perceptrons. The simplest perceptron has $m$ input values (which correspond with the $m$ features of the examples in the training set) and one output value. This is shown in Figure 5. If the sum of the products between the feature value $x_i$ and weight-factor $w_i$ is larger than zero, the perceptron is activated and 'fires' a signal (+1). Otherwise it is not activated. Mathematically, the weighted sum between the input-values and the weight-values is the scalar-product $<w, x>$. This scalar-product is passed through the signum function $sgn()$; it maps the output to +1 if the scalar-product is positive, and it maps the output to −1 if the scalar-product is negative. Thus, this perceptron can mathematically be modeled by the function

$$y = sgn(b + <w, x>)$$  \hspace{1cm} (1)

Here $b$ is the bias term, i.e. the default value when all feature values are zero.

To train a perceptron, we need a set of training data points labeled with the information about whether they belong to a particular class (positive) or not (negative). Initially, the bias-value and all the elements in the weight-vector are set to small random values or zero. Then we pass each training data point through the perceptron. The actual output value obtained from each training example should be either +1 (if it belongs to the positive class) or −1 (if it does not belong to the positive class). The activation function value from Equation 1
is the predicted output value. If for a training sample $j$, the predicted output $y_j$ matches the actual (desired) output $d_j$, the program moves on to the next sample. If the predicted output $y_j$ does not match the desired output $d_j$, the weight-vector $w$ and bias $b$ is updated according to the equations given below.

$$w_{\text{new}} = w_{\text{old}} + (d_j - y_j) \times x_j \quad (2)$$

$$b_{\text{new}} = b_{\text{old}} + (d_j - y_j) \quad (3)$$

This weight update is performed for each training data point until it gives the correct output, or a predetermined number of iterations have passed.

The $\text{sgn}()$ function in Equation 1 could be replaced by other functions depending on the use of the perceptron. Some other popular activation functions are given below.
The binary step function

\[ f(x) = \begin{cases} 
0, & \text{for } x < 0 \\
1, & \text{for } x \geq 0
\end{cases} \]  \hspace{1cm} (4)

The Sigmoid function

\[ f(x) = \frac{1}{1 + e^{-x}} \]  \hspace{1cm} (5)

Rectified Linear Unit (ReLU)

\[ f(x) = \begin{cases} 
0, & \text{for } x < 0 \\
x, & \text{for } x \geq 0
\end{cases} \]  \hspace{1cm} (6)

Deep convolutional neural networks, like the one used in this work, usually use the ReLU activation function. After training, a perceptron learns to activate for certain input patterns and not activate for other patterns. In other words, each perceptron is a detector for certain input patterns.

In the rest of this thesis, the terms “neuron” and “perceptron” have been used interchangeably.

4.2 Artificial Neural Networks

An artificial neural network (ANN) is an interconnected group of perceptrons arranged in the form of layers. In Figure 6, each circle represents a perceptron as described in Section 4.1. The first, or the input layer, gets the input features. The last layer, or the output layer has a perceptron for each class label. Ideally, when an input belongs to class \( k \), only the \( k \)th perceptron of the output layer will have a high activation output, and all others will have low activation outputs. In cases of ambiguity, more than one of these neurons may have high activation values. An example of such a network is shown in Figure 6. The ANN also usually has
one or more internal or "hidden" layers between the input and the output layers. These layers are also made of perceptrons whose inputs are the combinations of the outputs of the perceptrons in their previous layers. Thus each neuron in an ANN is a detector for patterns that are a combination of patterns detected by the neurons of the previous layer. If a neural network has more than one hidden layer, it is known as a deep neural network.

A neural network can be a feedforward neural network which is trained by applying the method described in Section 4.1 on each neuron after each training input, or it can be trained by the back-propagation algorithm. The backpropagation algorithm introduced by Werbos [15] in 1974 calculates a gradient that is needed in the calculation of the weights to be used in the network. It is commonly used for training deep neural networks. The backpropagation algorithm works by moving the training data forward through the network first, then calculating the error and, and finally updating the weights by moving backwards through the network. It is much faster than the forward-propagation algorithm which must pass the same training samples again and again through the network for training.
4.3 Convolutional Neural Networks

Convolutional neural networks (CNNs) are deep artificial neural networks inspired from the mammalian visual system that are specially designed for image classification. Hubel and Wiesel [8] showed in the 1950s and 1960s that mammalian visual cortices contain neurons that each respond to small areas of the visual field known as its receptive field. Neighboring cells have similar and overlapping receptive fields. Similarly, a CNN is also made up of artificial neurons with overlapping receptive fields. Yann LeCun used a CNN to recognize handwritten digits in 1989 [12] but CNNs did not become popular for another two decades since they required very large labeled datasets for training and such datasets were not available at the time. With the creation of large labeled digital image datasets like Imagenet [6] in 2009, it became possible to apply CNNs to the task of classifying complex images into thousands of categories. This was first done successfully in 2012 by the AlexNet [11], a CNN designed by Alex Krizhevsky et al.

There are three main types of layers in a typical CNN: convolution, max-pooling, and fully connected. Other than these, there may be normalization, ReLU and dropout layers added in between. The final layer before the output layer is usually a softmax classifier. Some of these layers are described below.

**Convolution** Since images are not a single-dimensional string of numbers like the typical perceptron’s input but spread over 3 dimensions (length, width and color planes), the patterns that a CNN’s perceptrons detect must also be 3-dimensional. To achieve this, the weights of these neurons are arranged in the form of small 3-dimensional filters. The input image undergoes convolution with each of these filters, which is essentially the same operation as described before: a pixel-wise multiplication followed by summing of the products. As the filter is slid over the width and height of the input image, it produces a 2-dimensional
activation map that gives the responses of that filter at every spatial position. Intuitively, the network will learn filters that activate when they see some type of visual feature such as an edge of some orientation or a patch of some color. As the image progresses through the network, the filters of later layers activate on detecting combinations of these patterns that form complex shapes like eyes, hands and faces. The process of convolution is explained in Figure 7. Figure 8 shows the filters formed by the first convolution layer of the AlexNet [11] convolutional neural network trained on the ImageNet [6] image dataset.

**Fig. 7.** A convolution filter of size $5 \times 5 \times 3$ being used on an image of size $32 \times 32 \times 3$.

Max Pooling A max pooling layer takes the 2-dimensional output from its previous layer and scales it down by downsampling. To scale the output by a factor on \( n \), for instance, the layer replaces every \( n \times n \) group of values with the maximum of those values. This reduces the size of the data, as well as makes the features less sensitive to spatial displacement. This operation is shown in Figure 9.

Fully Connected Fully connected layers are simple hidden multi-layer perceptrons as described in Section 4.2. After several layers of convolution and pooling, the data is input into the fully connected layers. They are called fully connected because every neuron in each of these layers is connected to every neuron in the previous and next layers. Sometimes, the fully connected layers may contain dropout layers in between.

Dropout A dropout layer randomly deactivates the outputs of half the neurons from the previous layer. Dropout layers are added to reduce overfitting and to improve the generalization performance of a network. Basically disabling certain neurons (detectors for certain concepts) forces the network to learn to classify based on other concepts. Dropout is shown Figure 10.
Softmax  In mathematics, the softmax function, or normalized exponential function [2], is a function that transforms a K-dimensional vector of arbitrary real values to a K-dimensional vector of real values in the range (0, 1) that add up to 1.

AlexNet has five convolutional layers and three fully connected hidden layers. It also has eight normalization layers, eight ReLU layers, three max-pooling layers and a softmax layer. The network used in the current work is a very slightly modified version of AlexNet. This network is described in Section 5.1.

5 Proposed Method

Images of people displaying different emotions were downloaded from the Internet manually using a search engine and these were used both for training our network and testing it. The network was initialized with random values and trained by backpropagation. Different network designs were tried and the design that worked best with all classes was chosen. The network itself is described in Section 5.1. Apart from the classification output, the activation regions from

Fig. 10. The dropout layer randomly disables about half of the neurons to improve generalization performance. Image source: https://leonardoaraujosantos.gitbooks.io/artificial-intelligence/content/dropout_layer.html
<table>
<thead>
<tr>
<th>Layer</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Image Input</td>
<td>224x224x3 images with ‘zerocenter’ normalization</td>
</tr>
<tr>
<td>2</td>
<td>Convolution</td>
<td>96 11x11 convolutions, stride [4 4], padding [0 0 0 0]</td>
</tr>
<tr>
<td>3</td>
<td>Normalization</td>
<td>Batch normalization</td>
</tr>
<tr>
<td>4</td>
<td>ReLU</td>
<td>ReLU</td>
</tr>
<tr>
<td>5</td>
<td>Max Pooling</td>
<td>3x3 max pooling, stride [2 2], padding [0 0 0 0]</td>
</tr>
<tr>
<td>6</td>
<td>Convolution</td>
<td>256 5x5 convolutions, stride [1 1], padding [0 0 0 0]</td>
</tr>
<tr>
<td>7</td>
<td>Normalization</td>
<td>Batch normalization</td>
</tr>
<tr>
<td>8</td>
<td>ReLU</td>
<td>ReLU</td>
</tr>
<tr>
<td>9</td>
<td>Max Pooling</td>
<td>3x3 max pooling, stride [2 2], padding [0 0 0 0]</td>
</tr>
<tr>
<td>10</td>
<td>Convolution</td>
<td>384 3x3 convolutions, stride [1 1] and padding [0 0 0 0]</td>
</tr>
<tr>
<td>11</td>
<td>Normalization</td>
<td>Batch normalization</td>
</tr>
<tr>
<td>12</td>
<td>ReLU</td>
<td>ReLU</td>
</tr>
<tr>
<td>13</td>
<td>Convolution</td>
<td>384 3x3 convolutions, stride [1 1], padding [0 0 0 0]</td>
</tr>
<tr>
<td>14</td>
<td>Normalization</td>
<td>Batch normalization</td>
</tr>
<tr>
<td>15</td>
<td>ReLU</td>
<td>ReLU</td>
</tr>
<tr>
<td>16</td>
<td>Convolution</td>
<td>256 3x3 convolutions, stride [1 1], padding [0 0 0 0]</td>
</tr>
<tr>
<td>17</td>
<td>Normalization</td>
<td>Batch normalization</td>
</tr>
<tr>
<td>18</td>
<td>ReLU</td>
<td>ReLU</td>
</tr>
<tr>
<td>19</td>
<td>Max Pooling</td>
<td>3x3 max pooling, stride [2 2], padding [0 0 0 0]</td>
</tr>
<tr>
<td>20</td>
<td>Fully Connected</td>
<td>4096-sized fully connected layer</td>
</tr>
<tr>
<td>21</td>
<td>Normalization</td>
<td>Batch normalization</td>
</tr>
<tr>
<td>22</td>
<td>ReLU</td>
<td>ReLU</td>
</tr>
<tr>
<td>23</td>
<td>Dropout</td>
<td>50% dropout</td>
</tr>
<tr>
<td>24</td>
<td>Fully Connected</td>
<td>1024-sized fully connected layer</td>
</tr>
<tr>
<td>25</td>
<td>Normalization</td>
<td>Batch normalization</td>
</tr>
<tr>
<td>26</td>
<td>ReLU</td>
<td>ReLU</td>
</tr>
<tr>
<td>27</td>
<td>Dropout</td>
<td>50% dropout</td>
</tr>
<tr>
<td>28</td>
<td>Fully Connected</td>
<td>6-sized fully connected layer</td>
</tr>
<tr>
<td>29</td>
<td>Normalization</td>
<td>Batch normalization</td>
</tr>
<tr>
<td>30</td>
<td>ReLU</td>
<td>ReLU</td>
</tr>
<tr>
<td>31</td>
<td>Softmax</td>
<td>Softmax</td>
</tr>
<tr>
<td>32</td>
<td>Output</td>
<td>6 numbers representing the probabilities of the image being in the 6 emotion classes</td>
</tr>
</tbody>
</table>

Different internal layers were also analyzed to see what body language and expression cues the network is using for classification.
5.1 Network Structure

The CNN used in this project has 32 layers in all. Table 1 shows the details of these layers. The network was implemented using MATLAB. The input images were resized to $224 \times 224$ pixels and the dataset was augmented by producing more images from the existing ones. For this purpose, new images were created by transforming the existing images by random translations in both horizontal and vertical directions, random scalings from 80% to 120% along both horizontal and vertical axes, up to 5 deg rotation (both clockwise and anti-clockwise) and reflecting about the vertical axis. Additional images were also created by transforming the color images to grayscale. Figure 11 shows a schematic representation of this network.

6 Experiments

6.1 Dataset

Six emotions were used for the purposes of this project: angry, disgusted, fearful, happy, sad, and surprised. Images for these six emotions were collected by downloading images from the internet using Google Image Search. Each emotion was
searched with search strings like "happy people", "happy men", "happy women", "happy children", "happy faces", etc. and images were downloaded manually from the results. An effort was made to make the dataset diverse in terms of gender, age, race and facial features. Most images have the subjects faces visible, but images where the subject’s face is not visible were selected too. Duplicate images were avoided as much as possible. Some sample images from this dataset are shown in Figure 12. The number of images in each class are given in Table 2. Just as to demonstrate how consistent or inconsistent the facial expressions for an emotion category were, we ran an automatic face detector on every image and averaged these face images from each class. The result of this operation can be seen in Figure 13. It can be seen here that the happy and surprised classes produce most uniform facial features across images.

**Training** The neural network was trained by backpropagation on a random sample of the downloaded images. Neural Networks typically require a high
volume of input data and so we used 150 images per class for training. The rest of the images were used for testing. The input images were resized to $224 \times 224$ pixels and the dataset was augmented by producing more images by transforming the existing images. The operations used for this were random translations in both horizontal and vertical directions, random scalings from 80% to 120% along both horizontal and vertical axes, up to 5 deg rotation (both clockwise and anti-clockwise), reflecting about the vertical axis, and transformation of the color images to grayscale. Training usually took about 30 hours.

**Table 2.** Class names and number of images in each class

<table>
<thead>
<tr>
<th>Class Name</th>
<th>Image Count</th>
<th>Training Set</th>
<th>Test Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angry</td>
<td>172</td>
<td>150</td>
<td>22</td>
</tr>
<tr>
<td>Disgusted</td>
<td>165</td>
<td>150</td>
<td>15</td>
</tr>
<tr>
<td>Fearful</td>
<td>205</td>
<td>150</td>
<td>55</td>
</tr>
<tr>
<td>Happy</td>
<td>275</td>
<td>150</td>
<td>125</td>
</tr>
<tr>
<td>Sad</td>
<td>294</td>
<td>150</td>
<td>144</td>
</tr>
<tr>
<td>Surprised</td>
<td>354</td>
<td>150</td>
<td>204</td>
</tr>
</tbody>
</table>
Testing For testing, images different from the training set were passed through the trained network and the predicted label was matched to the actual label of the test images to measure accuracy. The training and test samples were selected randomly. The test images were scaled to the same size as the training images.

7 Results

7.1 Classification

The image classification accuracy of the proposed system is shown in Figure 14 in the form of a confusion matrix. The matrix represents the most successful neural network that we tested with a mean classification accuracy of 48.7%. While this may not look like a very high rate, it is important to note that randomly assigning one of 6 class labels to images would have a success rate of approximately 16.7% and this system achieves about three times that rate. In the figure, the
rows represent the actual labels of the test images and the columns represent the assigned labels. So high values (warm colors) in the main diagonal indicate correctly classified images while high values in other cells indicate confusion. The figure also shows the actual fraction of all images for each class that were correctly classified. As can be seen, happiness is the most accurately detected emotion with a success rate of 67% followed by anger (59%), disgust (53%) and surprise (47%). Fear and sadness were the most difficult to detect correctly, each being detected in only 33% of test cases.

Figure 14 also shows bright squares at row 5 and column 2, and at row 3 and column 4. The first one indicates that a lot of images from the sad category were mislabeled as disgusted. The second one indicates that many images from the fearful category are mislabeled as happy. Some of these images are shown in Figure 19 and Section 7.3 attempts to look at the causes of such confusions.

### 7.2 Overfitting

One problem with training a CNN with too few images is overfitting. Overfitting occurs when a network “memorizes” the samples belonging to the different classes and fails to generalize well on unseen images. Even though we augment our dataset heavily, the total number of training images is still very low for a task as complex as emotion recognition. So the network does display the effects of overfitting to some extent. One such example is shown in Figure 15. Different crops of the same image (one of very few instances where the same image is present in different classes, by mistake) were used as positive training samples for two different classes, angry and sad. Later, when presented with these same images for testing, the system classifies the angry crop into the angry class, and the sad crop into the sad class, showing that it had ”memorized” them. The only way to reduce overfitting would be to use more training images.
Fig. 15. An example demonstrating the effect of overfitting where different crops of the same image were used as positive training samples for two different classes and later used for testing.

7.3 Discussion

Apart from the classification task, the experiments conducted also generated a rich array of visualizations that help shed some light on the internal workings of the neural network and the intermediate representations learnt by it. This section discusses these results, and tries to understand the shortcomings of the system as well.

Figure 16 shows the responses of neurons at each stage of the network seen as heatmap visualizations. Neurons in each stage are detecting features going from the simplest features seen through filters like edge detection in stage one, to more complex regions of activity seen in stage five. The warmer regions show where the neuron is responding the most in an image and how the overall feature is extracted by the neuron. Please note that these are the responses of just five random neurons out of the millions in the entire network.

Figure 17 visualizes the number of images with high responses from each neuron in stage five. These important neurons indicate the visual patterns the
Fig. 16. An input image and sample responses from some neurons of stages 1 to 5. The yellow areas show high activation while blue areas show little or no activation. The size of the activation area increases as we go deeper into the network.

Fig. 17. The most responsive neurons in stage 5 for the six classes. The bar graphs plot the number of images where each particular neuron in stage 5 has had a high response for that class.
The system thinks are most important in classifying an image into an emotion class. The variation in responses across classes means that the neuron is looking for different features for different classes, which is in accordance with what is expected from the network and verifies the network is working correctly.

The feature patches visualizations shown in Figure 18 are the most significant features isolated in the last stage of processing in the neural network. The black portions have been blocked by the network as it has been deemed irrelevant to classification and the patches showing the image are the areas with the highest responses which are being used for classification. Not all features make outward sense to humans but some features are easier to interpret than others. In almost all of the emotions, facial features took prominence. In case of the most successfully identified emotion, happy, wide angles of the arms, shoulders thrown backward, cupping of the face, and smiling were isolated by the network. This is remarkable due to the fact that humans also use these indicators to evaluate emotion. A promising aspect of the network is that in group scenes, smiles and the same wide arm gestures were isolated in all subjects in the image.

Angry patches yielded pointing gestures, widening of the palms, furrowing of the eyebrows, and the angular setting of arms on hips were most commonly isolated. When looking at the disgusted patches, almost all major features isolated were facial. Furrowing of the brows, upward pinching of the lips, display of tongue, and crinkling of the nose are seen throughout. Some body language poses were displayed as well, such as pinching of the nose with a hand and putting hands in front with palms wide in a blocking stance. Like disgusted, surprised was also face heavy in features extracted. Almost all images had widening of the eyes and most displayed a circular, wide mouth shape. Hand touching the face was isolated, mainly covering the mouth or cupping the cheeks, which is a hallmark feature of surprise.
The most confused emotions of fearful and sad isolated an impressive amount of images with body language, however, the body language had less variety across images. Fearful images had the facial hallmark of the wide mouth that
may indicate screaming but also had extensive images with the covering of the face with both hands or extending hands in front, with one in front of the other. One feature isolated was even nail biting. Closing of the shoulders inward was also a promising feature isolated. Sad had very similar body language poses in that almost all had some form of covering the face with the hands. Downward curling of the lips was a facial feature that was extracted if the face was visible. A distinct body position of cradling was seen in the data which is very interesting and provided the most complete body language picture in that it also involved the lower body.

The final visualization, Figure 19, shows some images that the system misclassified. For many of these images shown here, the source of the misclassification is easy to interpret. For example, the confusion matrix in Figure 14 indicated that a large number of fearful images were interpreted as happy, which seems unusual. On looking at the images, however, we can see why this is so. These fearful images are full-body images of people with open gestures and they could easily be mistaken for happy images if the facial expression is disregarded. Similarly in some images that are labeled fearful, the emotion could easily be mistaken as surprise. In other images originally labeled surprise, the emotion could be seen as happy and this is what the system sees. Another issue that may be worth mentioning here is that most of these images show actors pretending to display emotions rather than real people displaying emotions. This may be affecting the performance of the network because it is difficult to say if all the body language cues in the images are consistent with real emotions.

Overall, the system seems to fixate more on facial expressions than body poses, but that is what we humans do as well. In future, an experiment could be designed where the faces of the models would be blurred and the system would
Fig. 19. Some sample images from the dataset which the proposed system has difficulty classifying properly.

be forced to rely on other cues. It would be interesting to see the features learned by such a neural network.
8 Conclusions

This thesis tries to solve the problem of classifying human emotions based on expressions, body language and poses from photographic images. For this purpose, we first download images from the Internet and build a dataset with 6 emotion categories. Next, we train a deep convolutional neural network with these images and test the classification performance of this network with close to 49% accuracy. Finally, we analyze these results and generate a rich set of visualizations to try and understand the intermediate representation and the visual cues learned by the system. In future, this work could be extended by expanding the dataset, and focusing more on the bodies and less on the faces of the subjects.
References


