Data Augmentation for Deep Learning-based Gaze Estimation

Jorn Dijk[1]

Supervisor(s): Dr. Guohao Lan[1]**, Lingyu Du**[1]

[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 25, 2023

Name of the student: Jorn W. Dijk
Final project course: CSE3000 Research Project
Thesis committee: Dr. Guohao Lan, Dr. Xucong Zhang, Lingyu Du

Abstract

This study aims to provide insights in applying different data augmentation techniques to the input data of a convolutional neural network that estimates gaze. Gaze is used in numerous research domains for understanding and predicting emotions and actions from humans. Data augmentations consists of techniques to increase the size, variance and quality of training data to create better deep-learning models. Data augmentation is a widely used technique to reduce overfitting and increase accuracy of deep learning models. This research combines those two fields by first applying different individual data augmentations on the task of gaze estimation and after that combining the most useful methods to decrease the mean angular error even further. The results show that small geometric transformations, such as translating the image a portion of 15% or flipping the image horizontally 50% of the time give the most significant reductions in mean angular error. For individually applied data augmentation methods flipping got the best improvement, with 33% and 35% for both models in comparison to the baseline model. The best result is obtained by combining flipping with translation which got a mean angular error of 1.396 and 1.389 for both models. For obtaining the results a lot of training is necessary, which was the main limitation to conduct the experiments.

1 Introduction

This section will first go into the background of the subjects related to this paper and then elaborate on the gap between the existing research and the research question.

1.1 Background

The current progression of machine learning gives many advancements in various scientific fields. Gaze estimation is one of these fields. The gaze of someone indicates the direction they are looking at, but it also has numerous other applications. A gaze can help detect the emotions of a person [1] and expresses features of someone, such as desires and needs. These qualities make gaze estimation useful for various applications. Gaze estimation can improve applications that use eye-tracking by helping to develop better ways to measure where people are looking. Eye-tracking technology can aid in identifying the importance of different objects in websites [2] and help improve websites by providing information about what aspects hold the attention of a user [3]. Human-robot interaction has also gained significant importance due to the quick developments in robotic technologies. In this context, gaze estimation can assist by teaching the robots how and when to react [4] and identifying roles of people [5] through introducing implicit communication [6]. Furthermore, gaze estimation can help in processing social signals from humans [7], conducting psychological research [8; 9], predicting actions of humans [10], and various other tasks. Therefore it could be very beneficial to explore this field and seek improvements.

The research question of this paper combines gaze estimation with data augmentation. Data augmentation is already a widely used technique to improve the efficiency, accuracy or computational cost in various applications of convolutional neural networks (CNNs). Data augmentation is also often used to solve the problem of having limited data because augmentation can increase the size and quality of the data and use that to improve the model [11; 12]. In the same context, it can also help decrease the computational cost by relying on less training data [13]. Data augmentation in CNNs is often used on images but also proves useful in various other media, such as audio [14].

1.2 Related work

Much research has been done on data augmentation and deep-learning-based gaze estimation. This subsection will elaborate on three papers that proved important in these research fields.

Paper [15] studies the influence of using full-face images on the accuracy of gaze estimation. They describe a method using spatial weights CNN to estimate the gaze, which proved more robust when there was much variation in head pose and illumination. The paper highlights the importance of using full-face images for training the convolutional neural network. Full-face images will also be used in this research.

Gaze estimation is becoming more applicable in the world by enabling gaze estimation from from captured facial images from general-purpose cameras. This enables people with standard desktop webcams without additional camera equipment to utilize gaze estimation as described in [16]. This paper studies if people without expensive software or hardware could reliably track gaze from webcam images of one person. They concluded that when using only images from one person, the gaze could be estimated for full-face images with a minimum angular error of 1.14 degrees. This shows that nowadays gaze can be estimated from pictures taken by people without expensive equipment and highlights the importance of methods to increase the size and variation of input data.

Data augmentations can be beneficial to increase the variation and reduce the problem of overfitting as described in paper [17]. This paper studies the effects of different data augmentations on image classification, semantic segmentation and object detection tasks. They test the augmentations on many tasks with different models and datasets and see that using augmentations almost always increases the accuracy of the models. This paper shows that using data augmentations can increase the accuracy of many tasks, and our paper will investigate if they can also enhance the task of gaze estimation by decreasing the mean angular error.

1.3 Research questions and main contributions

In section 1.1, many useful applications of gaze estimation are shown. The importance of gaze makes it useful if we could find ways to improve gaze estimation. Convolutional neural networks are often used on

tasks involving images, such as gaze estimation [15; 18]. The previous subsections showed the importance of gaze estimation, how to use CNNs to estimate gaze and the benefits of data augmentation on CNNs, but not all three subjects combined. This paper combines those three subjects by using data augmentation to alter the mean angular error of CNNs used for gaze estimation. Data augmentation can do this by increasing the variation of the training set and using this to reduce overfitting. The paper will use data augmentation methods proven to work or not work on other problems and compare them to the results of using them on gaze estimation.
The problem tackled in this paper can be described in a single research question: *What effect do different data augmentations on images have on the mean angular error of gaze estimation using convolutional neural networks?*

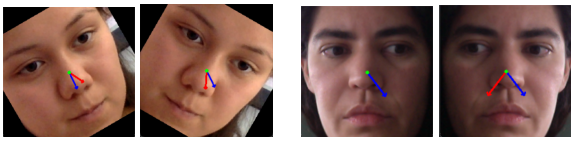
The following provides an overview of the structure of the rest of this research paper. Section 2 provides an overview of the methodology followed by this research. It will explain the preliminary knowledge to understand the research and provide an overview of the used data augmentations. Section 3 will then go more in-depth on the specific setup to conduct the methodology, so the parameters used and the division of the dataset. Section 3 also evaluates the results obtained by following the methodology. After section 3, the responsible research is elaborated on in section 4. This section will contain, among others, how to reproduce the experiments and possible ethical issues. Based on the obtained results and the methodology, the discussion (section 5) explains the main obstacles, discusses the obtained results and gives ideas for future work. Lastly, section 6 contains a summary of the completed research.

2 Methodology

Researching the effects of data augmentation on gaze estimation requires multiple steps, outlined in this section. Subsection 2.1 explains the needed preliminary knowledge to understand the research and subsection 2.2 explain all applied data augmentations.

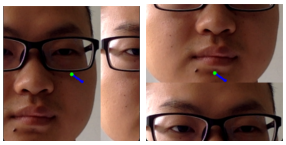
2.1 Preliminaries

The two main definitions used in this research are convolutional neural networks and data augmentation. This subsection explains those definitions.
Convolutional Neural Network (CNN): Many machine learning models exist to solve all kinds of problems. CNNs are a subset of these models commonly used to solve image-related tasks. They have multiple features which are useful for problems regarding images. Images are taken as input data and then passed through multiple convolutional layers, which extract specific features of the images important for the task that needs to be solved and reduce the dimensions without losing important data. This reduction of the dimensions of images makes a CNN efficient for images which consist of many pixels and thus much input data. Because of their practicality on images, CNNs are used to conduct the experiments for this research.



(a) Rotation with 30 (left) and -30 (right) degrees

(b) Normal image (left) and flipped image (right)



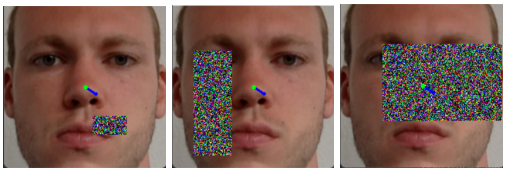
(c) Circular shift with horizontal (left) and vertical (right) shift

Figure 1: Images with applied geometric transformations with original labels (blue) and changed labels (red)

Data Augmentation: Data augmentation is a technique that consists of applying different transformations and augmentations on a dataset to create new samples from the existing dataset. It creates new samples useful for increasing the diversity or size of the dataset. This research uses data augmentation to increase the diversity of the dataset and see if it can alter the mean angular error of the used CNN.

2.2 Data augmentation

To study the effects of data augmentation, different data augmentation methods are applied to the images of the dataset. Data augmentation in CNNs can reduce overfitting, make the network more robust and improve the performance [17]. Several data augmentation methods are chosen, outlined below, to study their effect on the mean angular error of gaze estimation. Most methods have some randomization to generalize the augmentation, such as randomizing the degree of rotation in the case of rotation augmentation. Each randomization is applied to one image and passed to the neural network for training. This randomization ensures that the network is trained for the general augmentation method and not just a specific instance of the augmentation. For data augmentations that could potentially distort the labels, the model is trained on the instance of the images with corrected labels for the augmentation and trained with labels that are not changed. Training on both instances shows if the effect of the data augmentation changes when correcting the labels. This is particularly interesting for the task of gaze estimation because it is a regression problem where the labels depend on many aspects of the input image. If the images change even the slightest, they could suddenly represent a different gaze, harming the training. The augmentation is generally not applied during testing because we want to research if data augmentation helps the gaze estimation task on the original images. It is also essential to keep the other values constant between the baseline training and training on the different augmentations to get a truthful result not caused by coincidence.



(a) Erasure with three different possible randomizations

(b) Cropping with three different possible randomizations

Figure 2: Images with applied random erasure and random cropping with their labels in blue

2.2.1 Geometric transformations

Geometric transformations are transformations that change the geometry of the image without changing the pixel values. The pixel positions, however, can change in these transformations.

Flipping & rotation: The first augmentations tested are rotating the image a random number of degrees and flipping the image randomly. These two methods increase the performance of a model by reducing the impact of overfitting of image classification tasks [19] by adding variation to the dataset. Both augmentation methods use some randomization. Rotation is applied with a random number of degrees within -30 and 30, as seen in figure 1a, because too high rotations will not represent the images that could happen in the test set. Flipping will also be only done horizontally for the same reason as the low range of degrees for the rotation augmentation. Randomization on flipping decides if there will be a flip in the images, as seen in figure 1b. The random factor is 0.5, so 50% of all images will be flipped. Training is done two times for these two methods, once with correcting the labels and once without.

Circular shift: The next geometric transformation applied is a circular shift proposed by the paper [20] and shown in figure 1c. The method divides an image into seven parts and shifts the image horizontally or vertically a random number of parts and puts the removed part on the other side of the image. As seen in [20], this method increases the accuracy of all the used datasets and CNNs on the image classification task. For this research, we will test if this method also works for the regression problem of gaze estimation.

Random erasure & random cropping: Occlusion on images means that some parts of the image, which can be useful for the task, are hidden. Strong CNN models should be able to do the task even when some image features are occluded. When occluding facial images in the training set, the model might learn the importance of all features in the image and not learn on specific parts of the image, such as only the eyes or nose. Occluding parts of the images can generalize the model and reduce overfitting. This paper researches three methods are researched random erasure, shown in figure 2a, random cropping, shown in figure 2b, and combining these as described in [21]. Random erasure is actually not a geometric transformation but still included in this section due to the effectiveness together with random cropping as seen in paper [21]. Random erasure will use the hyperparameters and method proven to work best for the image classification task in [21]. Cropping is performed randomly between a scale of 0.4 and 1 because cropping too much will remove certain features, such as whole eyes, making it harder to predict the gaze, resulting in worse results. The aspect ratio is kept at one because if the image gets a different aspect ratio, the label will probably also change, resulting in higher mean angular error and less meaningful results.

Figure 3: Images with three possible translations according to the translation augmentation and their labels in blue

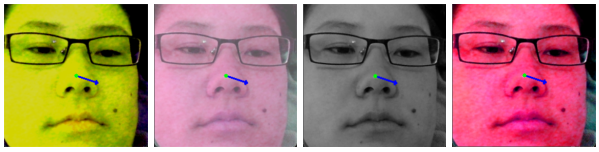
Translation: Images of full faces can contain different face positions. People can have their heads slightly to the left, right, down or up while still having most of their face in the view of the camera. This research applies a translation operation to the images to account for these changes. 75% of the images will have a random translation of -15% - 15% both horizontally and vertically. The experiments use 75% to ensure a greater number of complete images than translated images in the training data because complete images represent the test set better. The translated images represent small changes in the test set for people who shift slightly to a specific direction, so the model can still learn from people even when specific features of one side of the images are not in the picture. 15% keeps the visible portion in the images large enough so that important features like the nose and eyes are always in the picture, even with the greatest possible translation. The part that is shifted away is filled with pixel value 0, which gives a black colour as seen in figure 3.

2.2.2 Appearance transformations

Appearance transformations affect the appearance of the input images. Images often consist of 3 channels for RGB values. With these channels, we can do several augmentations, like removing channels, converting to grayscale or changing the brightness and contrast.

Colour jitter & gaussian blur: For this paper, we tested the methods described in [22]. This involved colour jitter with a strength of 1 using the algorithm from appendix A of [22] and from the same appendix Gaussian blur. In the paper,

(a) Blurred images with three possible strengths



(b) Images with four different possible color jitters

Figure 4: Images with different possible appearance transformations and their labels in blue

Figure 5: Images with applied noise injection with variance 0.005 (left), 0.01 (middle) and 0.1 (right) and their labels in blue

they defaulted to combining those two augmentations, so this is tested to see if that has more effect on the mean angular error even when applying them individually does not give a positive result. These data augmentations are shown in figure 4.

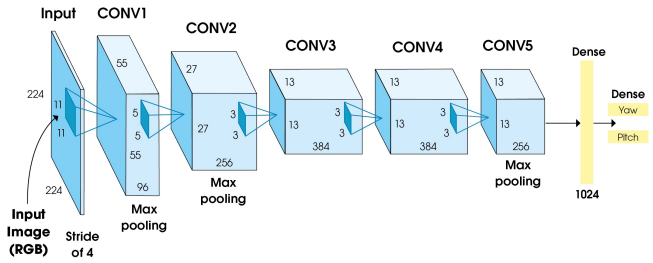
Noise injection: Images often contain imperfections, like black spots, poor camera quality, etc. Noise injection can simulate these imperfections [23] to prevent overfitting. When adding noise, the images become more general, making the network less prone to overfitting. There are many noise injection methods with all different performances on tasks for different CNNs [24]. However, the one that will be applied and tested here is Gaussian noise because it is the most common method for noise injection. The Gaussian distribution has two parameters, the mean and the variance. The mean will be 0 because otherwise, there will be a bias towards increasing or decreasing the pixel values. The variance will be tested with three different values, 0.1, 0.01 and 0.001, shown in figure 5, to see if adding more or less will result in a different mean angular error.

3 Experiments

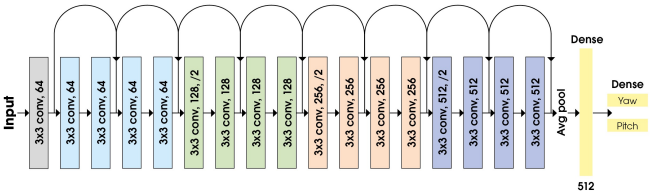
This section explains the experimental setup, which others can use to replicate the experiments, the results obtained by running the models and the evaluation of those results.

3.1 Experimental setup

This subsection describes the experimental setup. The first part elaborates on the baseline implementation of the models,



(a) Version of AlexNet used to train the models



(b) Version of ResNet18 used to train the models

Figure 6: Models used to train and test the network

such as parameters used, training time and used models. Subsection 3.1.2 explains the dataset used and the division of the data. Together they give a thorough guide for reproducing the experimental setup.

3.1.1 Implementation details

A baseline model for the convolutional neural network is necessary to research the effects of data augmentation on the result of gaze estimation. Altered versions of AlexNet [25] and ResNet18 [26] are used to train and test the data as seen in figure 6. AlexNet is used due to the low computational cost and proven effectiveness on image datasets [25]. AlexNet works well for problems with large image datasets as input, and this research extends this model to make it useful for a regression task, by removing the softmax for the output. The task of gaze estimation is a regression task that takes an image as input and returns a 2-dimensional output. This network takes in images of people with three colour channels (RGB) and a width and height of 224 pixels and returns a 2-dimensional vector of the pitch and yaw of the gaze. The output has only two channels instead of the 1000 in the original AlexNet model. So, the number of hidden layers is lowered to one with 1024 neurons compared to the two hidden layers with 4096 neurons each in the AlexNet model [25]. Lowering the number of neurons and hidden layers decreases the computational costs, which is necessary to train all the networks with different data augmentations within the allocated time of the research. Lowering the neurons also decreases overfitting and reduces the mean angular error of the model.

The second model used to study the effect of the data augmentations is an altered version of ResNet18 [26]. The model used in this research has a lower the number of neurons in the dense layer compared to the original ResNet18 reducing the mean angular error and the training time. The number of neurons in the dense layer is reduced to 512 with two output channels for pitch and yaw.

The selected models have several hyperparameters which influence their outcomes. These hyperparameters and other

choices for the models are described below. In all the experiments, the Adam optimizer is used [27] with an initial learning rate of 0.0001 and a batch size of 64. Training takes ten epochs, after which calibration goes on for 200 epochs. These values are chosen through trial and error. A higher learning rate prevented the ALexNet model from learning for some data augmentations, and a lower calibration would give suboptimal results for both models. The training uses ten epochs because after 2-3 epochs, the training without calibration does not decrease the mean angular error anymore, and ten epochs gives a great margin to consider the slower convergence of some data augmentation methods. The loss function used is the L1 loss function, which is the sum of the absolute differences between predicted and actual labels. Seeds keep the randomized starting weights and the shuffled training data constant between the baseline model and the models with augmented images as input. Three iterations with three different seeds are done to see if the augmentation has an effect and that it is not a coincidence based on randomized weights or other random factors, which ensures that the results are valid and robust.
The task of the research question is not to optimize the model for gaze estimation [15], so a perfect model with the lowest test error is unnecessary. The models used are thus not optimized for the task, which saves time. That ensures that there is enough time to train all the models while still demonstrating the effects of the data augmentation on the mean angular error of the network.

3.1.2 Dataset

The dataset used in this research is the normalized version of the MPIIFaceGaze dataset [15], which is based on the MPIIGaze dataset [18]. MPIIFaceGaze is a dataset used specifically to conduct research for gaze estimation. It consists of 15 persons, with each 3000 images resulting in a total of 45000 images and a label describing the gaze for each person. The images consist of 3 channels, Red, Green and Blue (RGB), with 448 $\times$ 448 pixels each. The labels consist of 2 values, yaw and pitch, which describe the rotation in radians around the y and x-axis of the image.
This data is not used in the original form. All the images are reduced to 224 $\times$ 224 pixels using bicubic interpolation, so training and testing require less computational time. After reducing the size of the images, the data needs to be divided into training and testing data. Training is done on the first 14 persons and then calibrated on 10% of the 15th person to improve the model [28]. We considered two methods for calibration. The first method first trains on the first 14 persons for ten epochs and then calibrates the trained model for 200 epochs on the 300 images of the 15th person. 200 epochs are chosen due to the slower convergence to the optimal results for some data augmentation methods. The second method described in [28] does the calibration during training and adds the 10% of the 15th person to the training data and trains on the whole training data for 20 epochs instead of 10 to make sure the models converge to their optimal results. The first method is chosen for this research as it has a lower mean angular error for the test set of the baseline model than the first method shown in table 1.

Calibration	Model	Angular error
True	AlexNet	2.146±0.015
	ResNet18	2.516±0.038
False	AlexNet	3.148±0.132
	ResNet18	3.101±0.042

Table 1: Angular error with and without calibration on Alexnet and ResNet18

3.2 Results

Many results are achieved while doing the research. This subsection explains how the results compare to the baseline model (3.2.1), how to decide that the results have a significant difference to the results of the baseline model (3.2.2) and shows the results of applying different data augmentation methods and their evaluation (3.2.3).

3.2.1 Comparison baseline

A baseline model is necessary to compare the effects of different data augmentations on the performance of a neural network. When comparing the results of both models from section 3.1.1 the main difference is that ResNet18 starts with a higher error but converges a lot quicker. The sample variance of AlexNet is 0.0004, with a mean angular error of 2.146 for the test set and 0.116 for the training set. Moreover, the sample variance of ResNet18 is 0.0027, with a mean angular error of 2.516 for the test set and 0.093 for the training set. Despite the low mean angular error, the models are still overfitting quite a bit with a difference of more than 2 degrees for both models. We want to solve this using data augmentations on the input data. The following sections show the results of applying the different data augmentation techniques from the methodology to see if they reduce the overfitting problem and reduce the mean angular error. The comparison uses the angular error between the actual and predicted labels. The angular error is calculated from the yaw and pitch values by first converting them to x, y and z values using equation 1, 2 and 3.

$$x = -cos(pitch) + sin(yaw) \tag{1}$$

$$y = sin(pitch) \tag{2}$$

$$z = cos(pitch) + cos(yaw) \tag{3}$$

The angular error uses these coordinates in 4 with v1 as the predicted coordinates and v2 as the actual coordinates.

$$Angular\,error = \arccos \frac{\vec{v1} \cdot \vec{v2}}{|\vec{v1}| \times |\vec{v2}|} \times \frac{180}{\pi} \tag{4}$$

Each model trains for three iterations, and the final angular error used to compare the results is the mean of the angular errors of the three iterations. In each table with results the ± symbol means the mean absolute difference between the mean of the angular error of the three iterations and the angular error of the three iterations individually.

3.2.2 Significance of results

As seen in section 3.2.1, the final angular error is obtained by taking the angular error from each of the three iterations and averaging those results. The significance is determined using

Augmentation	Model	Variable	Angular error	Imprvovement (%)	p-value
Baseline	AlexNet	-	2.146±0.015	0	1
	ResNet18	-	2.516±0.038	0	1
Flipping	AlexNet	gaze unflipped	11.585±0.053	-439.862	<**0.0001**
		gaze flipped	1.445±0.016	32.650	<**0.0001**
	ResNet18	gaze unflipped	13.194±0.136	-424.385	<**0.0001**
		gaze flipped	1.623±0.06	35.494	<**0.0001**
Rotation	AlexNet	gaze unrotated	1.925±0.039	10.319	**0.0023**
		gaze rotated	1.868±0.046	12.945	**0.0017**
		test rotated	4.674±0.032	-117.801	<**0.0001**
	ResNet18	gaze unrotated	2.076±0.070	17.503	**0.0019**
		gaze rotated	2.065±0.058	17.931	**0.0015**
		test rotated	4.568±0.039	-81.545	<**0.0001**
Noise Injection	AlexNet	Var(0.1)	4.958±0.283	-131.023	**0.0003**
		Var(0.01)	2.551±0.020	-18.887	<**0.0001**
		Var(0.001)	2.278±0.167	-6.174	0.3579
	ResNet18	Var(0.1)	5.901±0.192	-134.535	<**0.0001**
		Var(0.01)	2.671±0.034	-6.165	**0.0180**
		Var(0.001)	2.587±0.015	-2.810	0.0914
Circular Shift	AlexNet	-	2.110±0.019	1.663	0.1183
	ResNet18	-	2.222±0.037	11.697	**0.0052**
Erasing	AlexNet	-	2.205±0.068	-2.738	0.3531
	ResNet18	-	2.464±0.080	2.066	0.5054
Cropping	AlexNet	-	1.843±0.061	14.139	**0.0031**
	ResNet18	-	1.868±0.053	25.735	**0.0002**
Blur	AlexNet	-	1.980±0.038	7.742	**0.0080**
	ResNet18	-	2.404±0.046	4.451	0.0815
Color Jitter	AlexNet	-	2.131±0.023	0.709	0.5225
	ResNet18	-	2.536±0.055	0.0794	0.7324
Translation	AlexNet	-	1.809±0.065	15.702	**0.0014**
	ResNet18	-	1.868±0.041	25.751	**0.0001**

Table 2: Angular error of the applied data augmentation methods and improvement in comparison to the baseline model

the p-value under the null hypothesis that the data augmentation does not affect the mean angular error of a model. For the null hypothesis, the mean of the baseline models is compared to the means of the different data augmentation methods to see if the changes are significant enough to discard the null hypothesis. A p-value lower than 0.05 is used to consider whether a result is significant because this value is generally used in scientific research to determine significance [29]. An unpaired two-tailed t-test computes the p-value with the three iterations as the independent samples.

3.2.3 Influence of data augmentations

This subsection will outline and evaluate the results of the applied data augmentations.

Flipping (F): The results for flipping are interesting because all p-values are lower than 0.05 and have significance. Flipping the input data but leaving the gaze unchanged gives a significantly worse mean angular error. This is likely since flipping some images will result in similar images with completely different gazes, which confuses the model. Flipping the labels accordingly gives a significantly better result. Images of faces of people are not symmetrical due to lighting, setting and facial features, meaning that flipping increases the variation of the input data and the generality of the models. This reduces overfitting and reduces the mean angular error, as seen in table 2. With 28 and 34 percent improvements, we can conclude that flipping the input data while flipping the labels can help predict the gaze.

Rotation (R): Rotation gives interesting results. Both rotating with and without the labels gives a significant reduction in mean angular error. Applying rotation together with changing the labels increases the variation of the dataset and reduces overfitting. People who typically look slightly down will now also have images that make them look in a different direction, increasing their ability to learn from these images. We would expect that rotating without rotating the labels accordingly would worsen the results because the gaze direction would not correspond with the rotated image anymore. This is not the case, as in row four of table 2. The models can most likely learn from the amount of black space inserted when rotating images because test images have no black space, and images in the training set without black space have correct labelling. To test this, the test images are also rotated, with their labels accordingly, while the training labels are unrotated. This increases the mean angular error for both models by more than 80% because now images

Combination	Model	Angular error	Improvement (%)	p-value
F + C	AlexNet	1.657±0.037	22.785	<**0.0001**
	ResNet18	1.830±0.100	27.257	**0.0013**
F + R	AlexNet	1.519±0.018	29.220	<**0.0001**
	ResNet18	1.695±0.078	32.626	**0.0002**
F + T	AlexNet	1.396±0.039	34.927	<**0.0001**
	ResNet18	1.389±0.070	44.794	<**0.0001**
T + R	AlexNet	1.987±0.046	7.395	**0.0183**
	ResNet18	1.994±0.064	20.737	**0.0011**
J + B	AlexNet	2.218±0.032	-3.362	0.0662
	ResNet18	2.579±0.037	-2.499	0.2104
C + E	AlexNet	2.169±0.018	-1.088	0.2665
	ResNet18	2.260±0.195	10.160	0.1649
F + S + R	AlexNet	2.066±0.073	3.706	0.2298
	ResNet18	1.945±0.174	22.680	**0.0178**
F + T + R	AlexNet	1.749±0.079	18.497	**0.0030**
	ResNet18	1.887±0.161	24.999	**0.0085**

Table 3: Angular error of the applied combinations of data augmentation methods, with the codes taken from table 2

which look the same have different labels.

Noise injection: Injecting noise gives varying results depending on the variance used for injecting noise. More variance gives a higher difference between the original and augmented images. Too high variance, as seen in table 2, increases the mean angular error because the training set is not representative for the test set anymore. Lowering the variance gives results similar to the baseline model because the image will look more similar to the original image.

Circular shift (S): Circular shifting got interesting results because we can see from table 2 and 3 that only ResNet18 significantly improves from applying circular shift. This shows that different models react differently to the circular shift operation. Circular shifting is basically the same as translation, with a higher portion shifted away and a different fill value. This could explain why ResNet18 does give an decrease as translation reduces the mean angular error of both models. The main difference between AlexNet and ResNet18 is the number of convolutional layers, so it could be that more convolutional layers are needed to extract useful information from the circular shift operation.

Random erasing (E): Erasing does not give a significant increase or decrease in mean angular error. Erasing deletes random parts of the images and does reduce overfitting, but it also increases the mean angular error of the training set. The mean angular error on the training set is 1.357 for AlexNet and 1.108 for ResNet18, so while the overall angular error does not change, we do see less overfitting due to the smaller difference between training and testing error. This is likely since erasing can hide features important for learning but also keeps the images more general by hiding different parts for each image.

Translation (T): As seen in table 2, the translation method described in section 2.2 significantly decreases the mean angular error of both models. It increases the variation of the training set by 75% of the time removing small portions of all sides of the images. The model now learns on images without those sides and can better predict images that are shifted to a side due to head position.

Random cropping (C): As seen in table 2, cropping significantly reduces mean angular error for both trained models. With around 14% and 26% increase, this method is useful to reduce overfitting for the task of gaze estimation. Erasing deletes certain parts of the image, while cropping enhances certain parts by making them bigger. This adds variation to the training data, decreasing the mean angular error of the test data. People have different sizes for nose, eyes, head, chin, cheeks, etc., due to genetics, camera position or head pose, which can be more accurately represented in the training data by cropping. The model learns more variation in the size of certain features or whole faces, which helps the model generalize for the test data.

Blur (B): Blurring the images according to the methodology described in section 2.2 has a small but significant impact on the AleNet model. This could be because AlexNet is less complicated than ResNet18 due to the smaller number of convolutional layers.

Color Jitter (J): Applying colour jitter to the input images does not result in significant outcomes. This can be because the test set has similar colour features as the training set. The colour features that are important to identify the gaze are most likely already generalized in the training data, so adding more difference by increasing brightness, contrast, hue, saturation or converting to grayscale will not decrease the mean angular error of the test set.

Combinations: Table 3 shows the results of applying combinations of data augmentation methods. Some data augmentation methods are chosen to study because they significantly

increase accuracy for other tasks, such as classification, as described in section 2.2. Those papers also gave insights into combining specific methods, which are also tested here for gaze estimation. These include J + B, C + E and F + CS + R. We see from table 3 that methods that did not work individually, such as colour jitter, blurring or erasing, also do not work in combination. For other combinations, we tried combining methods that, when applied individually, have a high impact on the mean angular error. Flipping had the greatest decrease in mean angular error, so it would make sense that combining this method with another would result in the best result in mean angular error. Table 3 shows that combining methods with flipping reduces the mean angular error significantly, but flipping alone is still better, except for combining flipping with translation. Other combinations add too much variation and change in the input data, making it less representative for the test data. Flipping together with translation gives the lowest mean angular error, with 1.396 for AlexNet and 1.389 for ResNet18.

4 Responsible Research

This section gives insights in the scientific integrity (4.1), reproducibility (4.2) and the ethical aspects (4.3) of the research

4.1 Scientific Integrity

This research involved collaboration with students C. Feng, Y. Reda, T. Nguyen and T. Penning. Each person researched a subject in the field of deep-learning-based gaze estimation. Baseline model architectures and parameters were shared to make sure everybody got similar results for the baseline model. Furthermore, to avoid plagiarism, all used literature is cited.

4.2 Reproducibility

The reproducibility of the research is achieved by providing all the information necessary in the different sections and references in this paper. Section 2 explains the different data augmentation methods in detail. Section 3 can then be used to set the parameters and architecture for the models. For hardware requirements it would be recommended to use a GPU to avoid long training times.

4.3 Ethics

The dataset used contains images of whole faces from 15 different subjects. These images can be considered sensitive information, but the work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. The license allows the material to be freely shared, copied, redistributed, and built upon. Except for the dataset, no personal or privacy-sensitive information was used for this research.

5 Discussion

This section will discuss the research by explaining the limitations and obstacles, giving an explanation of the result and providing possibilities for future work.

5.1 Limitations

The main limitation and bottleneck of the experiments was the training time. A network trains for all augmentation techniques and some combinations, which could not be done on the laptop available. DelftBlue is chosen to the experiments due to the high processing power and access to GPUs. However, these GPUs need to be allocated to run experiments on them, which can take much time when many other people also want to access the GPUs. Running the experiments and waiting for the results took much time, which resulted in suboptimal use of available time. This is not ideal for conducting research in the 8-9 weeks allocated. This also prevented running more iterations of the models. It would be ideal to have at least 5 or more iterations to base the conclusions on, but this was not possible due to the long training times.

5.2 Results

If we look at the results and their evaluation from section 3.2.3, we see that geometric transformations have the highest impact on the mean angular error. The reason for this could be, that most geometric transformation represent changes in the images that could also happen coincidentally when taking pictures. Cropping represents head being closer to the camera or bigger facial features, rotation represents tilting the head, flipping gives variation in facial features happening on both sides for a person and translation and circular shifting represents people that are shifted a bit to a side in an image.
Appearance transformations do add variation but do not have a high impact on gaze estimation. This could be because they do not add variation that didn't already happen in the training set. Colour jitter for example adds many colour features but the colour features useful for generalizing the model, such as changing the colours to other skin colours, already happens by using 15 different persons. We can conclude that small changes using geometric augmentations give the best results for decreasing the mean angular error for gaze estimation.
Another noticeable aspect of the results is the better decrease in mean angular accuracy of the ResNet18 model compared to the AlexNet model. For almost all significant reductions we see that ResNet18 has much more improvement than AlexNet. AlexNet only seems to have a better performance for the blurring augmentation method.

5.3 Future work

For future work, one thing that can be considered is the effect of other data augmentation techniques. Not all techniques and combinations are researched in this paper, so it would be interesting to see combinations with more complicated techniques, such as combining images.
For the data augmentations that were applied, many parameters could be changed. The parameters used for this research were mostly taken from other related research or were found by trial and error. However, they are not optimized to get the best parameter values for the mean angular error of gaze estimation. Future work could delve deeper into the effect of changing the parameters of data augmentation methods on the task of gaze estimation, such as changing the rotation degree or the level of blurring.

A question that arises from this work is the difference in effect on different convolutional neural networks. This research studies two different CNNs, and they have different results as discussed in the previous subsection. The reason could be the number of convolutional neural layers or some other underlying feature of the CNN. Future work could study the effect of convolutional neural network features on the results obtained by this research.

6 Conclusion

The main goal of this research is to study the effect of different data augmentation methods on the mean angular error of gaze estimation using a convolutional neural network. Data augmentation methods alter input data to increase the size and variation. This research uses data augmentation to study its influence on the mean angular error of gaze estimation to find some insights that could be useful for this task. There are many data augmentation methods, but based on past research, the augmentation methods of flipping, cropping, rotation, translation, colour jitter, blurring, noise injection and circular shifting are chosen to study their effects. The mean angular error over three training iterations is used to study the results. They showed that small changes using geometric transformations give the best results for decreasing gaze estimation mean angular error. With flipping over 30% increase, rotation over 10%, cropping between 10% and 30% depending on the model and translation around 20% compared to the results of the trained model without augmentations.

After analyzing the results, combinations were chosen to improve the mean angular error of the models further. Combinations of methods that individually decrease the mean angular error, such as flipping and cropping or translation and rotation do also decrease the model when combined, but not by more than when applied individually. Only flipping in combination with translation gives a better result than applying flipping or translation alone. This combination gives the best result with a improvement of 35% for AlexNet and 45% for ResNet18 resulting in mean angular errors of 1.396 and 1.389.

From the results we see that applying data augmentation on gaze estimation using convolutional neural networks can benefit the performance when correct augmentation methods are used. This could be useful for many applications using gaze estimation, such as human robot interaction or eye-tracking technologies.

References

[1] Y. Zhao, X. Wang, and E. Petriu, "Facial expression analysis using eye gaze information," Univ Ottawa, Ottawa, USA, 2011, pp. 7–10.

[2] J. P. Velasquez, "Combining eye-tracking technologies with web usage mining for identifying website key-objects," *ENGINEERING APPLICATIONS OF ARTIFICIAL INTELLIGENCE*, vol. 26, no. 5-6, pp. 1469–1478, MAY-JUN 2013.

[3] Q. Wang, S. Yang, M. Liu, Z. Cao, and Q. Ma, "An eye-tracking study of website complexity from cognitive load perspective," *DECISION SUPPORT SYSTEMS*, vol. 62, pp. 1–10, JUN 2014.

[4] D. Das, M. G. Rashed, Y. Kobayashi, and Y. Kuno, "Recognizing gaze pattern for human robot interaction," in *HRI'14: PROCEEDINGS OF THE 2014 ACM/IEEE INTERNATIONAL CONFERENCE ON HUMAN-ROBOT INTERACTION*, ser. ACM IEEE International Conference on Human-Robot Interaction. ACM SIGCHI; ACM SIGAI; IEEE Robotics & Automation; HFES; AAAI; ACM; IEEE, 2014, pp. 142–143, 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Bielefeld, GERMANY, MAR 03-06, 2014.

[5] B. Mutlu, T. Shiwa, T. Kanda, H. Ishiguro, and N. Hagita, "Footing in human-robot conversations: How robots might shape participant roles using gaze cues," 03 2009, pp. 61–68.

[6] O. Palinko, F. Rea, G. Sandini, and A. Sciutti, "Eye tracking for human robot interaction," in *2016 ACM SYMPOSIUM ON EYE TRACKING RESEARCH & APPLICATIONS (ETRA 2016)*, S. Spencer, Ed. ACM; ACM SIGGRAPH; ACM SIGCHI, 2016, pp. 327–328, 9th Biennial ACM Symposium on Eye Tracking Research and Applications (ETRA), Charleston, SC, MAR 14-17, 2016.

[7] A. Vinciarelli, M. Pantic, H. Bourlard, and A. Pentland, "Social signal processing: State-of-the-art and future perspectives of an emerging domain," *MM'08 - Proceedings of the 2008 ACM International Conference on Multimedia, with co-located Symposium and Workshops*, 10 2008.

[8] M. L. Mele and S. Federici, "Gaze and eye-tracking solutions for psychological research," *COGNITIVE PROCESSING*, vol. 13, no. 1, SI, pp. S261–S265, AUG 2012.

[9] K. Kampe, C. Frith, R. Dolan, and U. Frith, "Psychology: Reward value of attractiveness and gaze," *Nature*, vol. 413, pp. 589–589, 10 2001.

[10] V. Novak and R. Riener, "Enhancing patient freedom in rehabilitation robotics using gaze-based intention detection," *IEEE ... International Conference on Rehabilitation Robotics : [proceedings]*, vol. 2013, pp. 1–6, 06 2013.

[11] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, 2019, cited By :3794. [Online]. Available: www.scopus.com

[12] A. Mikołajczyk and M. Grochowski, "Data augmentation for improving deep learning in image classification problem," 2018, p. 117 – 122. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85050025181&doi=10.1109

[13] S. Bargoti and J. Underwood, "Deep fruit detection in orchards," 2017, p. 3626 – 3633. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85028002275&doi=10.1109

[14] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, p. 279 – 283, 2017, cited by: 941; All Open Access, Bronze Open Access, Green Open Access. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85015238568&doi=10.1109%2fLSP.2017.2657381&partnerID=40&md5=32e51ebbf6466f513006b595c52de08c

[15] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "It's written all over your face: Full-face appearance-based gaze estimation," vol. 2017-July, 2017, Conference paper, p. 2299 – 2308, cited by: 205; All Open Access, Green Open Access. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85030228048&doi=10.1109%2fCVPRW.2017.284&partnerID=40&md5=f1e83f1920e0afa0235ad6d80256ae3a

[16] M. F. Ansari, P. Kasprowski, and P. Peer, "Person-specific gaze estimation from low-quality webcam images," *Sensors*, vol. 23, no. 8, 2023. [Online]. Available: https://www.mdpi.com/1424-8220/23/8/4138

[17] S. Yang, W. Xiao, M. Zhang, S. Guo, J. Zhao, and F. Shen, "Image data augmentation for deep learning: A survey," 2022.

[18] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Appearance-based gaze estimation in the wild," *CoRR*, vol. abs/1504.02863, 2015. [Online]. Available: http://arxiv.org/abs/1504.02863

[19] Y.-L. Chang, T.-H. Tan, W.-H. Lee, L. Chang, Y.-N. Chen, K.-C. Fan, and M. Alkhaleefah, "Consolidated convolutional neural network for hyperspectral image classification," *Remote Sensing*, vol. 14, no. 7, 2022. [Online]. Available: https://www.mdpi.com/2072-4292/14/7/1571

[20] K. Zhang, Z. Cao, and J. Wu, "Circular shift: An effective data augmentation method for convolutional neural network on image classification," in *2020 IEEE International Conference on Image Processing (ICIP)*, 2020, pp. 1676–1680.

[21] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," 2017.

[22] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., vol. 119. PMLR, 13–18 Jul 2020, pp. 1597–1607. [Online]. Available: https://proceedings.mlr.press/v119/chen20j.html

[23] Y. Jiang, R. Zur, L. Pesce, and K. Drukker, "A study of the effect of noise injection on the training of artificial neural networks," in *2009 International Joint Conference on Neural Networks, IJCNN 2009*, ser. Proceedings of the International Joint Conference on Neural Networks, 2009, pp. 1428–1432, 2009 International Joint Conference on Neural Networks, IJCNN 2009 ; Conference date: 14-06-2009 Through 19-06-2009.

[24] M. E. Akbiyik, "Data augmentation in training {cnn}s: Injecting noise to images," 2020. [Online]. Available: https://openreview.net/forum?id=SkeKtyHYPS

[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., vol. 25. Curran Associates, Inc., 2012. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[27] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 12 2014.

[28] N. Bandyopadhyay, S. Riou, and D. Schwab, "Effect of personalized calibration on gaze estimation using deep-learning," *CoRR*, vol. abs/2109.12801, 2021. [Online]. Available: https://arxiv.org/abs/2109.12801

[29] C. Andrade, "The p value and statistical significance: Misunderstandings, explanations, challenges, and alternatives," *Indian Journal of Psychological Medicine*, vol. 41, p. 210, 05 2019.