WROCŁAW UNIVERSITY OF SCIENCE AND TECHNOLOGY
FACULTY OF ELECTRONICS

FIELD: Computer Science
SPECIALIZATION: Internet Engineering (INE)

MASTER OF SCIENCE THESIS

Research on methods of changing objects in images using Deepfake technology

Badania metod zmiany obiektów na obrazach z wykorzystaniem technologii Deepfake

AUTHOR:
Michał Zendran

SUPERVISOR:
Dr inż. Andrzej Rusiecki

GRADE:

WROCŁAW 2020

Contents

1 Abstract **3**

2 Introduction **4**
2.1 Motivation 4
2.2 Objective and assumptions 5
2.3 State of the art 5
2.4 Naming conventions and terminology 5

3 Theoretical background **6**
3.1 Artificial neural networks 6
3.2 Convolutional neural networks 7
3.3 Supervised, unsupervised and semi-supervised training 7
3.4 Haar feature-based cascade classifiers 7

4 Deepfake methods **8**
4.1 Autoencoder 8
4.2 Variational autoencoder 10
4.3 VAE-GAN 11
4.4 CycleGAN 11

5 Datasets **12**
5.1 Dataset description 12
5.2 Data pre-processing 13
5.3 Data augmentation 14

6 Technologies **16**
6.1 Software and Libraries 16
6.2 Hardware 16

7 Networks implementation **17**
7.1 Autoencoder 17
7.2 Variational autoencoder 17
7.3 VAE-GAN 17
7.4 CycleGAN 17

8 Results **18**
8.1 Autoencoder 18
8.2 Variational autoencoder 18
8.3 VAE-GAN 18
8.4 CycleGAN 18

9 Conclusions **19**

Bibliography **20**

List of Figures **22**

List of Tables **23**

Chapter 1

Abstract

To be written at the end of the work.

Chapter 2

Introduction

2.1 Motivation

Machine learning has found many, different applications in the field of image data processing and computer vision. From picture classification to image denoiseing and resolution enhancement, artificial neural networks has gained the opinion of exceptionally useful tool. But for some time, a new, controversial use-case has been getting more and more attention in both media and research circles. So-called "deepfake" technology has opened doors to many new possibilities of picture generation, but also raised many issues of moral and legal matters.

Deepfake is a technology from the field of machine learning designed to combine and overlay objects in images or videos creating deceptively realistic counterfeits. The name comes from combination of two terms: “deep learning” and “fake”, and has its origins in a Reddit user named “deepfakes”. Initially the term was associated only with face-swapping technology, but with time it was extended to all deep learning implementations of changing objects in images.

Deepfake technology has already found multiple applications such as changing seasons in the landscapes images, transforming horses into zebras or "repainting" images in styles of different artists. But the most controversial and impactful use-case so far is already mentioned face-swapping. In times of overwhelming amount of news it’s getting harder and harder to filter out fake ones from valuable peaces of information. People generally tend not to check sources of information but rather blindly follow hot stories in social medias and television. Such environment combined with capabilities of deepfake gives possibilities of influencing elections by misrepresenting politicians in forged videos to defame or blackmail theme. Another popular use-case of described technology is creating erotic videos by replacing faces of porn actress with faces of well-known celebrities. This application might have less dangerous consequences than influencing world politics but may be hurtful to people that became objects of such act.

Although there are many malicious ways of using deepfake technology it might also be used for good reasons such as helping people to cope with the loss of the loved once or in entertainment filed by de-ageing actors to play younger-selves. Besides, to be able to detect and fight harmful applications of deepfake it might be vital to deeply understand algorithms and techniques behind it. Therefore, conducting research on that part of machine learning field seems to have great meaning in incoming times.

2.2 Objective and assumptions

This project aims to implement and compare four different methods of changing objects in images with application of artificial neural networks. For sake of this research, human faces were chosen as an object of replacement, as it rises a complex issue of simultaneous color, texture and shape modification.

As there are no numerical methods of measuring the quality of images obtained from deepfake algorithms, the only way of appraising results of methods discussed in this research is visual evaluation. To be able to fairly rate each implemented technique, the same set of images will be used as a learning dataset for all cases. Therefor, effects of all approaches will be visually evaluated and compared with each other, which will result in the final assessment. This rating of methods is the expected outcome of the research.

There are two main factors that will be taken into a consideration during a results evaluation process. First of them is a resemblance of the faked image, to the appearance of the imitated person. The more striking similarity, the better. The other crucial aspect is preservation of original facial expression and pose, as the believable deepfake must capture the source material movements. Resultant of those two factors will be the main feature to be rated. It is assumed that neither of mentioned characteristics should outweigh the other one, but rather, the final effect should be well-balanced composition of both aspects.

2.3 State of the art

While there are many, great articles and papers that elaborately explain all mentioned approaches of generating deepfakes, no comparison of those methods were found. To be written ...

2.4 Naming conventions and terminology

Below, all abbreviations and naming conventions used in this research are listed and explained:

- Deepfake – name of the deep learning technology of swapping objects in images or an end result generated witch such technology
- ANN – artificial neural network
- CNN – convolutional neural network
- AE – autoencoder
- VAE – variational autoencoder
- GAN – generative adversarial network
- VAE-GAN – variational autoencoder-generative adversarial network
- Activation function – transfer function
- GPU – graphics processing unit

Chapter 3

Theoretical background

3.1 Artificial neural networks

An artificial neural network is a computing system inspired by the neural structure of the biological brain and its way of processing information. Such structure is able to "learn" how to solve certain problems or recognize patterns without being pre-programmed with rules how to do it. Artificial neural networks are main tool used in deep learning, which is part of bigger family of machine learning [1].

Figure 3.1: Deep learning belongingness

Fundamental unit of artificial neural network is so-called "artificial neuron" which is a mathematical function modeled after the structure of biological neuron [2]. Such artificial neuron has a number of inputs $\{x_0, x_1, x_2, ..., x_n\}$ with assigned, separated weights $\{w_0, w_1, w_2, ..., w_n\}$ for each of them to be multiplied by. In the simplest case, all product are summed and passed to a transfer function φ. Additionally a bias value b is added to the sum of products which allows to shift the activation function up or down. Then the output y of the neuron is calculated as follows:

$$y = \varphi\left(\sum_{j=0}^{n} x_j w_j\right) \tag{3.1}$$

General idea of artificial neuron model is illustrated in figure 3.2. Typically, neurons are organized into layers which may perform different operations. The output of neurons from one layer is passed to the input of the next layer or might be a part of output vector of whole network. Modern models of artificial neural networks may consist of tens or even hundreds of different layers, on which depends how well given network will cope with certain tasks. Those layers, during process of network training, can adjust theirs weights to produce output values that correctly solve given problem [2] [3].

Figure 3.2: Artificial neuron model

Artificial neural networks have found many, different applications in various fields, in both, academic and business environments. For some time, they are successfully used to solve problems of system identification and control (processes in industrial sector), image classification (medical diagnosis), financial forecasting, face recognition (security), sequence recognition (speech recognition), pattern recognition (data generation) and many more [4].

3.2 Convolutional neural networks

Explain how it works, what are main use-cases and so on.

3.3 Supervised, unsupervised and semi-supervised training

Description of both and what are main differences and when to use which.

3.4 Haar feature-based cascade classifiers

As in section name

Chapter 4

Deepfake methods

4.1 Autoencoder

Autoencoders are the most basic approach to the problem of deepfake generation. In fact, all mentioned methods, except for CycleGAN, are just different variations of this idea. Autoencoder is a type of artificial neural network that learns to reproduce given input in an unsupervised manner. The goal is to train functions $A : \mathbb{R}^n \rightarrow \mathbb{R}^p$ (encoder) and $B : \mathbb{R}^p \rightarrow \mathbb{R}^n$ (decoder) to satisfy condition given in equation 4.1 as described in [5],

$$arg\,min_{A,B} E[\Delta(x, B \circ A(x))] \tag{4.1}$$

where E – expectation over the distribution of x and Δ – reconstruction loss function, which measures the distance between given input and the output of the decoder. General idea of autoencoder model is illustrated in figure 4.1. Typically, architecture of autoencoder consists not only of input and output layers, as this would result in simple coping pixels from the input to the output of the network, but also contains single or multiple hidden layers in between, with the number of neurons lesser than the number of pixels in the input image. Such structure causes bottleneck effect and creates so-called compressed representation at the output of the encoder part, known also as "feature map" or in case of deepfake "latent face". Such compression causes feature map to preserve only information most relevant for later reconstruction and gets rid of unnecessary data.

Figure 4.1: General idea of autoencoder

Generating deepfakes using autoencoders approach consists of three major steps. Described process is illustrated in figure 4.2. Let us assume that X is a set of face images of person x and Y is a set of face images of person y. Firstly, as shown in figure 4.2a, an encoder is trained to produce feature maps for images of faces from both classes. Afterwards, two decoders are trained separately to reproduce original images from latent

faces generated by pre-trained encoder, as presented in figure 4.2b. Finally, decoders are switched to produce images from one class based on feature maps from the other class, which was illustrated in figure 4.2c.

(a) Training autoencoder

(b) Training decoders X and Y

(c) Generating X from Y and Y from X

Figure 4.2: Three steps of generating deepfakes

4.2 Variational autoencoder

Unlike basic autoencoders, which purpose is to reproduce given input by "memorizing" it, variational autoencoders aim to produce latent vectors (feature maps) that follow a unit Gaussian distribution [6]. Such modification allows to generate new images by sampling a latent vector from the unit Gaussian distribution and pass it as an input of the decoder network. The goal in VAE model training is to optimize the loss function given in equation 4.2 [7]. First component of this function is a reconstruction loss $\mathcal{L}_{rec}$, which, by comparing pixels from image x, encoded to a latent vector $z = Encoder(x) \sim q(z|x)$, and reconstructed image $\bar{x} = Decoder(z) \sim p(x|z)$, measures expected marginal log-likelihood of the observations in x. Second component, so called Kullback–Leibler divergence [8], measures the extent to which one probability distribution differs from another. Applying KL divergence to VAE loss function allows to control distribution of the latent vector z.

$$\mathcal{L}_{vae} = \mathcal{L}_{rec} + \mathcal{L}_{kl} \tag{4.2}$$

$$\mathcal{L}_{rec} = -E_{q(z|x)}[\log p(x|z)] \tag{4.3}$$

$$\mathcal{L}_{kl} = D_{kl}(q(z|x)||p(z|x)) \tag{4.4}$$

Presence of KL divergence in VAE loss function generates problem with applying backpropagation to the encoder network, as decoder randomly samples from latent vector z and backpropagation cannot flow through random node as in figure 4.3a. As a solution, a "reparameterization trick" is introduced [9]. This method maps the output of an encoder to separate vectors, mean μ and standard deviation σ, instead of a z-dimensional latent vector. Such representation allows to calculate z in a following way:

$$z = \mu + \sigma\epsilon \tag{4.5}$$

where ϵ is sampled from a multivariate standard normal distribution ($Normal(0, 1)$) and therefore acts as a stochastic component. Such reparameterization allows to run backpropagation process as now latent vector is not purely random, but consists of deterministic and stochastic nodes as illustrated in figure 4.3b.

(a) Nodes before reparameterization (b) Nodes after reparameterization

Figure 4.3: Reparameterization trick

By applying reparameterization trick to equation 4.4, Kullback–Leibler divergence takes following form [10]:

$$\mathcal{L}_{kl} = -\frac{1}{2}\sum_{i=0}^{z}(1 + \log(\sigma_i^2) - \mu_i^2 - \sigma_i^2) \tag{4.6}$$

Generating deepfakes using variational autoencoders approach is based on the same algorithm as autoencoders approach illustrated in figure 4.2. The only difference between those two methods is the way the latent face (latent vector) is obtained.

4.3 VAE-GAN

Idea behind deepfake generated by GAN actually "VAE-GAN".

4.4 CycleGAN

Describe what is it, what it consists of, what are its applications, why I thought it should work for deepfake. Explain how it works exactly. Show learning process and results (good ones: horses to zebras and bad ones: face to face). Idea behind deepfake generated by CycleGAN. Explain why I'm assuming it should it work?

Chapter 5

Datasets

5.1 Dataset description

Creating a high quality, well-balanced dataset is an essential step in the process of deepfake generation. Poorly prepared training set might cause that even great network architectures and state of the art algorithms will produce disappointing results. The most efficient approach to this problem ,in case of face-swapping technology, is obtaining images of targeted people from video recordings, which allows to produce great number of images that cover different facial expressions and head positions. There are several factors that should be taken into a consideration in order to construct proper dataset. First of all, videos of people with at least slight resemblance should be chosen. The most important feature in this matter is skin tone but the more similar appearances, the more deceiving deepfake might be achieved. Another important factor is overall quality of source material. Videos with high resolution and large number of frames per second are best suited for this purpose, but also require more computational power and time to properly train necessary networks. Finally, after a deconstruction of videos into single images, all set should be revised to remove pictures that depict faces which are blurry, deformed or in some ways covered, for example by hand.

For sake of this research the "VoxCeleb2" dataset was used. As described in [11], VoxCeleb2 consists of over 1 million utterances of over 6000 celebrities derived from videos from YouTube platform. Source data is diversified in terms of recorded people's genders, ethnicities, ages and accents, but also in terms of videos quality, lighting conditions, stability and lengths. Each recording has a resolution of 224 by 224 pixels and depicts closeup shot of a character's face and shoulders, which is close enough to clearly capture subtle facial expressions.

Figure 5.1: Sample images from VoxCeleb2 dataset

5.2 Data pre-processing

To prepare dataset best suited for purpose of this research, following preparations were made, in accordance with guidelines described in section 5.1. Sample images of pre-processed data are presented in figure 5.2.

1. Two actors (Leonardo DiCaprio and Robert Downey Jr.), further called subject A and subject B, were chosen as targets of face replacement. This choice was dictated by how well their faces are known and recognizable, which facilitates the final assessment, and by factors mentioned in section 5.1.

2. From the set of all videos of chosen subjects available in "VoxCeleb2" dataset, those which presented subjects in similar age, and had good recording quality, were selected.

3. From each video, every tenth frame was extracted to limit the amount of nearly identical images. Additionally, through Haar feature-based cascade classifiers described in section 3.4, the face itself was cut out from each extracted frame to discard unnecessary parts of pictures.

4. Resolutions of all obtained face images varied, therefore data had to be rescaled to a common resolution of 160 by 160 pixels.

5. Final step, was to save obtained images as arrays in a form of a single file, in uncompressed ".npz" format. This allows easy data transferring and simplifies the process of loading data during the training of artificial neural networks.

(a) Sample images of subject A (b) Sample images of subject B

Figure 5.2: Sample images of pre-processed data

5.3 Data augmentation

While generating results using autoencoders method, a specific relationship was observed. It was noticed that generated deepfakes look like composition of two faces, one directed to the left and the other to the right. This phenomenon might be seen in figure 5.3a, where first column from the left presents original images, second contains images reconstructed from latent spaces by decoder XY and third presents actual deepfakes obtained by decoding latent faces encoded from subject X, with decoder Y, as explained in section 4.1.

(a) Sight directed to the left

(b) Sight directed to the right

Figure 5.3: Models trained without data augmentation

Undoubtedly, deepfakes presented in figure 5.3a are of much worse quality then those, generated from the same, but diverted images, presented in figure 5.3b. Such phenomenon occurred, as significant majority of faces extracted from subject Y in created database was directed to the right. That kind of imbalance resulted in models learning, that "proper" reconstruction of subject Y should be directed to the right, regardless of position of original face.

To solve the issue of unbalanced data, so called data augmentation technique was used. This approach assumes that more information might be extracted from collected data, through different augmentation methods [12]. When processing images, following methods might be applied: image inversion, rotation, scaling, zooming, translations or adding Gaussian noise. In case of deepfake generation, significant transformations of original picture would cause, that the generated images would not resemble the originals. Therefore, only one augmentation technique, reflection, was applied. As presented in figure 5.4, networks trained with original data extended by augmented one perform much better when processing images depicting different head positions.

(a) Sight directed to the left

(b) Sight directed to the right

Figure 5.4: Models trained with data augmentation

Chapter 6

Technologies

6.1 Software and Libraries

All scripts and programs necessary for data pre-processing described in section 5.2 were implemented in Python 3.7.6 programming language [13] from Anaconda distribution [14]. This platform provides over 7,500 data science and machine learning packages such as OpenCV (Open Source Computer Vision Library), an open source computer vision and machine learning software library built to provide a common infrastructure for computer vision applications [15], or NumPy, a library that provides a multidimensional arrays and matrices objects, and a large set of high-level mathematical functions to operate on these arrays [16].

Models of artificial neural networks, its training process and validation were implemented on a Google Colaboratory platform that allows to write and execute arbitrary python code through the browser, and is especially well suited to machine learning and data analysis [17]. The most important library used in development process was Tensorflow 2.0, an end-to-end, open source, python-friendly platform designed for complex computations, machine learning and deep learning [18]. Additionally, to simplify networks implementation process, Keras library was used. Keras is an open-source, high-level neural networks API, written in Python and supported in Tensorflow's core library. Created as a higher-level, more user friendly interface to significantly simplify development of deep learning models [19].

6.2 Hardware

Data pre-processing described in section 5.2 were entirely developed on a local computer with Windows 10 operational system, NVIDIA GeForce GTX® 960M GPU, quad core processor Intel® Core™ i7-6700HQ with base frequency 2,60 GHz, 16 GB RAM memory SO-DIMM DDR4 Synchronous with frequency 2133 MHz and 256 GB SSD memory disc SK Hynix SC308.

Google Colaboratory platform described in section 6.1 allows a free access to powerful GPUs such as Nvidia Tesla K80 [20] which is compatible with CUDA technology – a parallel computing platform and programming model developed by NVIDIA for general computing on graphical processing units. With CUDA, it is possible to significantly speed up computing applications by using the power of GPUs [21].

Chapter 7

Networks implementation

Detailed description of implementation of each method. What are the topologies, what callbacks were used, why those parameter, why those batches itp

Describe data loading from npz file (normalization)

7.1 Autoencoder

7.2 Variational autoencoder

7.3 VAE-GAN

7.4 CycleGAN

Chapter 8

Results

To ensure comparability of results obtained from each implemented method, all details related to network topologies and training process, except for those that are specific for analyzed method, should be unified. It is especially important in case of methods which incorporate autoencoder idea, as dimensionality of latent space, may have great impact on final results, and therefore should be kept the same for each method. Such assumptions ensure that all differences in results arise from exceptional quality of particular method.

8.1 Autoencoder

8.2 Variational autoencoder

8.3 VAE-GAN

8.4 CycleGAN

Chapter 9

Conclusions

Bibliography

[1] V. Shah, "Difference between artificial intelligence,machine learning and deep learning," 2017. [Online]. Available: https://buzzrobot.com/difference-between-artificial-intelligence-machine-learning-and-deep-learning-ccfd779eca7b

[2] K. Suzuki, *Artificial Neural Networks - Methodological Advances and Biomedical Applications*, 04 2011.

[3] B. Mehlig, "Artificial neural networks," 2019.

[4] O. Abiodun, A. Jantan, O. Omolara, K. Dada, N. Mohamed, and H. Arshad, "State-of-the-art in artificial neural network applications: A survey," *Heliyon*, vol. 4, 11 2018.

[5] D. Bank, N. Koenigstein, and R. Giryes, "Autoencoders," 2020.

[6] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2013.

[7] X. Hou, L. Shen, K. Sun, and G. Qiu, "Deep feature consistent variational autoencoder," 2016.

[8] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, 1951.

[9] D. Kamath, "Generating new faces with variational autoencoders," 2020. [Online]. Available: https://towardsdatascience.com/generating-new-faces-with-variational-autoencoders-d13cfcb5f0a8

[10] S. Odaibo, "Tutorial: Deriving the standard variational autoencoder (vae) loss function," 2019.

[11] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *INTERSPEECH*, 2018.

[12] C. Shorten and T. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, 12 2019.

[13] D. Kuhlman, *A Python Book: Beginning Python, Advanced Python, and Python Exercises*, 12 2013.

[14] "Anaconda." [Online]. Available: https://www.anaconda.com/about-us

[15] "OpenCV." [Online]. Available: https://opencv.org/about/

[16] "NumPy." [Online]. Available: https://numpy.org/

[17] "Google Colaboratory." [Online]. Available: https://research.google.com/colaboratory/faq.html

[18] "TensorFlow." [Online]. Available: https://www.tensorflow.org/

[19] "Keras: The Python Deep Learning library." [Online]. Available: https://keras.io/

[20] "NVIDIA TESLA K80." [Online]. Available: https://www.nvidia.com/en-gb/data-center/tesla-k80/

[21] "CUDA." [Online]. Available: https://developer.nvidia.com/cuda-zone

List of Figures

3.1 Deep learning belongingness 6
3.2 Artificial neuron model 7

4.1 General idea of autoencoder 8
4.2 Three steps of generating deepfakes 9
4.3 Reparameterization trick 10

5.1 Sample images from VoxCeleb2 dataset 12
5.2 Sample images of pre-processed data 13
5.3 Models trained without data augmentation 14
5.4 Models trained with data augmentation 15

List of Tables