Federal State Autonomous Educational Institution for Higher Education
National Research University Higher School of Economics

Faculty of Computer Science
Educational Program
Applied Mathematics and Information Science

TERM PAPER
(Research project)
Variational autoencoders for synthetic brain MRI images

Prepared by the student of group БПМИ192, 3rd year of study,
Voronin Andrei Andreevich

Supervisor:

4th year Ph.D. candidate in Skoltech, Moscow,
Kondrateva Ekaterina Andreevna

Abstract

Deep learning shows high potential for many medical image analysis tasks. Neural networks can work with full-size data without extensive preprocessing and feature generation and, thus, information loss. Recent work has shown that the morphological difference in specific brain regions can be found on MRI with the means of Convolutional Neural Networks (CNN). However, interpretation of the existing models is based on a region of interest and can not be extended to voxel-wise image interpretation on a whole image.

In the current work, we consider different methods of feature extraction and image compression applied to brain MRI images without any kind of preprocessed data. The main models that are studied in this work are convolutional autoencoders (AE) and variational convolutional autoencoders (VAE). These models are used to obtain low dimensional representations of original data (latent spaces) which are analyzed and used in problems of demographic signs of gender classification.

Key words

Magnetic resonance imaging • Deep Learning • 3D CNN • Convolutional Autoencoders • Variational autoencoders • Demographic signs of gender classification

Introduction

Deep learning has significantly influenced progress in image recognition, text recognition, and many other areas. Recently, the use of deep learning and neural networks in particular has contributed to medical diagnostics and analysis of medical images. Analysis of medical images allows you to make diagnoses, make predictions, monitor the effectiveness of treatment and stratify patients. The use of neural networks in this area simplifies the work of experts and sometimes even allows us to notice some things that cannot be noticed.

This paper discusses the application of convolutional neural networks to magnetic resonance imaging (MRI) images of the human brain. MRI images as an example of medical images are important for medicine. However, often the sizes of MRI scans are too large for straightforward computational analysis and require large computing power. In order to reduce the dimensions, various methods are used to extract important features from images. In the past, not only the MRI images but also regions of interest and other metadata were used for analysis, however modern technologies in deep learning and image recognition allow neural networks to work with only 3D images and extract features without extra work required. One of the well-known models of neural networks that allows us to reduce the dimensions of data with their subsequent reconstruction are autoencoders.

An autoencoder (AE) is a neural network architecture used to learn efficient representations of data in encoded form (embeddings). The autoencoder consists of two parts: an encoder and a decoder. The first one is responsible for compressing the data and extracting features, the second is responsible for correctly reconstructing data from the compressed format. Encoded data comes as the output of the encoder and forms a latent space, which is further used instead of the original data. There are various architectures and modifications for autoencoders that are considered in this work.

In recent years, a large variety of solutions have been presented that present outstanding results, however the main problem in neuromedical image analysis is the heterogeneity of data and thus the weak interpretability of the results. The problem is that the distribution of data differs from one medical scanner to another. Hence, it turns out that in practice some solutions give a low correlation with actual diseases or, on the contrary, give the same predictions for different diseases. In addition, various medical devices scan not only images but also other data that is used for predictions in lots of existing solutions. Also, it is worth mentioning that medical images are taken from real patients hence it turns out that often there is not enough data to train a generalized model, which is usually not the case in other areas of machine learning applications. Therefore, researchers often have to create synthetic data which adds to the problem of heterogeneity.

The current study aims to learn how to generate MRI data and study their low-dimensional representations in order to be able to bypass domain differences in prediction, or interpret pathologies using MRI. However, the global goal can be splitted into following subgoals:

- obtain low-dimensional representations of 3D medical images using autoencoder models
- study low-dimensional representations
- optimize the architecture and research

As an example of the practical use of autoencoders in this study, the classification of demographic signs of gender is solved and the results are compared between different architectures.

Literature review

There is a plenty of related works on topic of deep learning in neuroimaging and these are the most relevant:

1. DIVA: Domain Invariant Variational Autoencoders

This article tries to solve the problem of domain generalization. In medical imaging each person's data is presented as a separate domain and authors' goal is to generalize individual representation and to transform it to domain, unseen in the learning stage. Given not only brain images but also domain labels such as age, sex, etc. the solution in this article extractes features for a set of similar domains and disentangles these sources of variaton consequantly it is tranfered to a specific generalized domain which is why model achieves to work with examples that even were not presented in training data.

The key feature here is splitting resources into three subjects: there are three separated autoencoders for domain label, class label and any residual variations in the input x. Each autoencoder is trained for its specific purpose separately to predict their own domains. At the end all three models are combined to show impressive results in generalizing the original domain.

However, even though the model was constructed to train in both supervised and semi-supervised way, it still works with specific data which is presented in not only images but also specific labels. In current work our goal was to work with only 3D MRI images without meta-data but we could still use these techniques in our research.

2. 3D StyleGAN for Three Dimensional Medical Images

This article presents another popular architecture in CNN and Image recognition which is Generative Adversarial Net (GAN). GAN is a network that learns representations of data in a training set and creates new data considering these representations. This is achieved by creating two separate models: generator and discriminator. The first model learns to generate data and the second learns to classify generated and original data. The combination of cooperative training of these two models leads to the result that the eventual network learns representation of original data which is used to synthesize new data from the same latent space.

This study in particular considers StyleGAN2 by NVIDIA and adapts it to work with 3D brain images. All 2D convolutions, upsampling and downsampling are replaced with 3D versions. Finite model consists of the original StyleGAN generator with (De)Mod operators and discriminator which is a 3D residual network. To unite training of these two models was used the logistic loss function without regularization for the discriminator component.

The presented approach achieves image projection that minimizes distance between an input image I and a generated image. As the result model is capable of both generating new adequate brain images and projectising unseen images so that reconstructed images are almost identical to the input. Even though it is hard to distinguish the generated data from original data, this model is highly sensitive to hyperparameters of both generator and discriminator and training data itself, which should be fairly detailed to assure the high quality of generated images.

3. Three-dimensional convolutional autoencoder extracts features of structural brain images with a "diagnostic labelfree" approach: Application to schizophrenia datasets

This study is the most relevant in the scope of current research. It presents convolutional autoencoders for extracting features from MRI images with the following classification comparison. The presented models managed to both extract features and successfully reconstruct images in high resolution. Authors of this article use block type of architecture and experiment with parameters such as number of blocks in the whole autoencoder or number of channels in convolution functions. Apart from that the article presents practical usage of autoencoders applied to schizophrenia dataset. Latent space is used to classify brain images of schizophrenic patients and healthy people. Quality evaluation is measured by success of regression on different targets. With this final goal study compares different parameters for their architecture of autoencoders. Despite the fact that their best autoencoders achieve to successfully extract features and then reconstruct images, this research followed one practical purpose – to achieve good quality in classification. With this information, we could assume that this study is narrowly purposed consequently this requires further research in scope of global imaging generalization.

4. MRI Denoising using Progressively Distribution-based Neural Network, Magnetic Resonance Imaging

This study presents an autoencoder for denoising in brain imaging. Denoising is another widespread usage of autoencoders. In particular in brain MRI recognition denoising is fairly useful in spite of common inaccuracy of medical equipment. Architecture of the presented autoencoder allows it to extract features from original images however the final purpose of this extraction is to detect noise in these images, not to compress with saving main statistical parameters and key features that could be used later. Moreover, this study used the help of medical experts who evaluated anatomic accuracy and signal quality of achieved results. Apart from that, further research in practical usage of the denoising autoencoder showed that the combination of GAN with this autoencoder provided high correlation and high resolution of generated pictures. To conclude, this work is not clearly related to current study, however it demonstrates another practical usage of autoencoders in neuroimaging.

5. Deep convolutional autoencoders for reconstructing magnetic resonance images of the healthy brain

This work compares different state of art technologies in image recognition and their usage in brain MRI recognition. The main point of interest in the scope of our study are 5 types of autoencoders which are: original shallow residual, full pre-activation shallow residual, skip connection, Myronenko AE, residual U-NET. Authors compare and describe their architecture and evaluate visual similarity of reconstructed and original images. WHat is interesting is that none of these models use max pooling and upscaling methods in the decoder part. The last autoencoder which is residual U-NET – is a U-NET type architecture with full pre-activation blocks. (U-NET – is a skip connection network with concatenation of layers' results). Residual U-NET shows best results by chosen metrics. Despite the fact that all reconstructed images are visually similar to the original, there are no examples of generative or practical regression usage of these autoencoders in this paper.

6. Generation of 3D Brain MRI Using Auto-Encoding Generative Adversarial Networks

This article presents an autoencoder-based GAN which is meant to train on a small amount of data and still achieve impressive results in image generation. In particular, authors trying to solve the problem of scarcity of data present a new version of GAN by exploiting benefits of both Variational autoencoder (VAE) and GAN. In fact, they use labeled data to train their model which is not exactly what it is meant to achieve in current study. However, they show practical usage of their model by applying it to brain tumor classification dataset. According to the authors, it is difficult for their model to generate images for specific diseases such as Alzheimer.

7. Learn to Ignore: Domain Adaptation for Multi-Site MRI Analysis

This study presents another approach in usage of autoencoders for feature extraction with further usage in classification problems. The problem the authors tried to solve was scarcity of data in a real world scenario. They tried to design a model to use in classification purposes where the training data is combined of different datasets where class ratio is imbalanced from dataset to dataset making only small amounts of data consist of equal amounts of each classes' examples. This is why they decided to implicate a classificator inside of an autoencoder and thus making constraints on latent space to achieve their goal. Basically the presented model learns to ignore differences of medical scanners features and focus more on disease features. Despite the fact that this model outperformed others in multiple sclerosis classification, the presented solution is fairly narrow-purposed and is not tested in other applications of neuroimaging.

As it is shown, there are plenty of good solutions in brain MRI analysis. However, most practices in neuroimaging depend on the data they are using which usually differs from one medical scanner device to another. Which is why an in depth research is required every time someone tries to solve a problem in this sphere.

Methods

1. Training and estimation data analysis

The original MRI images were taken from the database Human Connectome Project (HCP) which contains M111 data from 1113 subjects: 507 men and 606 women of ages 22-36. T1 images were explored and preprocessed with HCP-pipelines2. For the morphometry data analysis we used Freesufrer 3 preprocessed features from section Expanded FreeSurfer Data for the same 1113 subjects. The morphometry characteristics as number of vertices, volumes, surface areas, and others were computed for 34 cortical regions according to Desikan-Killiany Atlas and for 45 subcortical areas according to Automated Subcortical Segmentation Atlas summing up in 935 vectorized features.

In the following experiments all models use only brain MRI images of size [58, 70, 58] for training. For equality of experiments for each model the same stop condition was used. The stop condition was to check validation loss for the last 10 epochs of training and compare if the mean of it was not greater than loss in current training epoch. This stop condition allowed control overfitting and at the same time the size of the window was enough to let loss function be optimized.

For testing classification accuracy labels of sex of each patient were used. As the baseline for classification the support vector machine classifier with rbf kernel (SVM) classifier from sklearn was trained on the full-size data reshaped to the 1-dimensional vector. To estimate every model linear regression classifier from sklearn was applied to latent space. Each classification estimation was performed with 5-folds cross validation to ensure stability.

2. First experiments

To begin experiments and research on latent space, we had to obtain an autoencoder to create latent space. The first version of autoencoder for brain MRI analysis was kindly provided by Skoltech students which was their past project in generating MRI scans. This version of autoencoder was taken as introductory material and was supposed to be modified in future. The architecture was common for autoencoders: encoder and decoder were symmetrical, consisted of 5 standard convolutional blocks, each block consisted of Convolutional layer with ReLU activation function following by BatchNorm layer in encoder and Upsample layer instead of Convolutional layer in decoder. The amount of channels in convolutional layers was constant on the whole structure of both encoder and decoder and equaled 64 channels. With the amount of channels and convolutions the dimension of latent space was (1, 256). According to knowledge that our shapes of original images were (58, 70, 58) this resulted in reduction of original images by 919 times.

```
self.encoder = nn.Sequential(
    nn.Conv3d(1, 64, kernel_size=3, padding=1),
    nn.ReLU(),
    nn.BatchNorm3d(64),
    nn.MaxPool3d(2),
    nn.Conv3d(64, 64, kernel_size=3, padding=1),
    nn.ReLU(),
    nn.BatchNorm3d(64),
    nn.MaxPool3d(2),
    nn.Conv3d(64, 64, kernel_size=3, padding=1),
    nn.ReLU(),
    nn.BatchNorm3d(64),
    nn.MaxPool3d(2),
    nn.Conv3d(64, 64, kernel_size=3, padding=1),
    nn.ReLU(),
    nn.BatchNorm3d(64),
    nn.MaxPool3d(2),
    nn.Conv3d(64, 64, kernel_size=3, padding=1),
    nn.ReLU(),
    nn.BatchNorm3d(64),
    nn.Conv3d(64, hidden_dim, kernel_size=3, padding=1),
    nn.AdaptiveAvgPool3d((1, 1, 1)),
    nn.Flatten()
)
```

```
self.decoder = nn.Sequential(
    nn.Unflatten(1, (hidden_dim, 1, 1, 1)),
    nn.Upsample(size=(3, 4, 3), mode='trilinear',),
    nn.Conv3d(hidden_dim, 64, kernel_size=3, padding=1),
    nn.ReLU(),
    nn.BatchNorm3d(64),
    nn.Upsample(size=(7, 8, 7), mode='trilinear'),
    nn.Conv3d(64, 64, kernel_size=3, padding=1),
    nn.ReLU(),
    nn.BatchNorm3d(64),
    nn.Upsample(size=(14, 17, 14), mode='trilinear'),
    nn.Conv3d(64, 64, kernel_size=3, padding=1),
    nn.ReLU(),
    nn.BatchNorm3d(64),
    nn.Upsample(size=(29, 35, 29), mode='trilinear'),
    nn.Conv3d(64, 64, kernel_size=3, padding=1),
    nn.ReLU(),
    nn.BatchNorm3d(64),
    nn.Upsample(size=(58, 70, 58), mode='trilinear'),
    nn.Conv3d(64, 64, kernel_size=3, padding=1),
    nn.ReLU(),
    nn.BatchNorm3d(64),
    nn.Conv3d(64, 1, kernel_size=3, padding=1)
)
```

3. Block architecture experiments with constant channels

The purpose of this part of experiments was to achieve visually acceptable result for reconstructed images to begin research of latent space. Moving from previous model, the architecture of blocks was considered to be investigated. However another type of base block was chosen. New block for encoder consisted of 2 Convolutional layers with ReLU activation functions and followed with BatchNormaliztion and a MaxPooling layer at the end of block. For decoder part, the block was fully symmetrical with reversed layers for Convolutional and MaxPooling layers which are Transposed Convolution and UnPooling. The number of blocks and channels in autoencoder was compared using these metrics: MSE loss for reconstruction quality and classification accuracy applying the classifier to each models' latent space. At this point of time was chosen to crop the background of images and upscale them to save dimensions in order to use all resources of model only on brain features.

layer	ch.	kernel	stride	pad.	act.
conv_1	n	3,3,3	1	1	Relu
conv_2	n	3,3,3	1	1	Relu
pool_1	n	2,2,2	2	0	-
dcnv_1	n	3,3,3	1	1	Relu
dcnv_2	n	3,3,3	1	1	Relu
unpool_1	n	2,2,2	2	0	-

4. Constant embedding size with different number of channels

To experiment with different channel sizes was chosen to fix embedding size. Using the same block structure from the previous part the number of blocks was limited to 3 in both encoder and decoder. The purpose of this constraint was to fix the receptive field of the encoder. Despite the fact that embedding size decreased with the number of convolutions, the large amount of blocks lead the receptive field to be more than 80% of the brain which resulted in the model not learning small details on the brain image. On the opposite side, small number of blocks and consequently small number of convolutions resulted in large sizes of embeddings and models tend to not fully compress and generalize the original image. As the main purpose of autoencoder usage is to extract features and thus decrease the amount of data required to store original images, the small amount of blocks did not suit our research objective.

With a fixed amount of blocks, the numbers of channels were chosen to get embeddings with shape 35 times lower than original images. Different approaches were tested, such as decreasing or increasing the amount of channels with each convolution.

5. variational autoencoders

Variational autoencoder (VAE) is a modification of an autoencoder that provides a probabilistic way to represent the observations in the latent space. Unlike an autoencoder, the encoder of the VAE describes the probability distribution of each latent attribute. It is achieved by modifying loss function and output of encoder to consider distribution of embeddings while learning to compress and reconstruct original images. This makes latent space continuous which allows easy random sampling and interpolation. Output of the decoder of VAE is two vectors: a vector of means μ of data samples and a vector of standard deviations σ. The embedding is sampled using these two vectors. In fact, any distribution could be learned using VAE however most common is to constrain latent space to standard normal distribution.

To make the model approximate latent space to specific distribution the Kullback-Lebler divergence is added to the original MSE. Kullback-Lebler divergence allows to count “distance” between two distributions. By optimizing divergence, the model tends to learn embeddings with distribution close to normal.

$$\text{loss} = ||x - \hat{x}||^2 + \text{KL}[\,N(\mu_x, \sigma_x), N(0, I)\,] = ||x - d(z)||^2 + \text{KL}[\,N(\mu_x, \sigma_x), N(0, I)\,]$$

There are plenty of ways to build architecture for VAE however in this study we kept the block architecture from normal autoencoders as it has proven to be successful and also performed experiments on the number of blocks and channels.

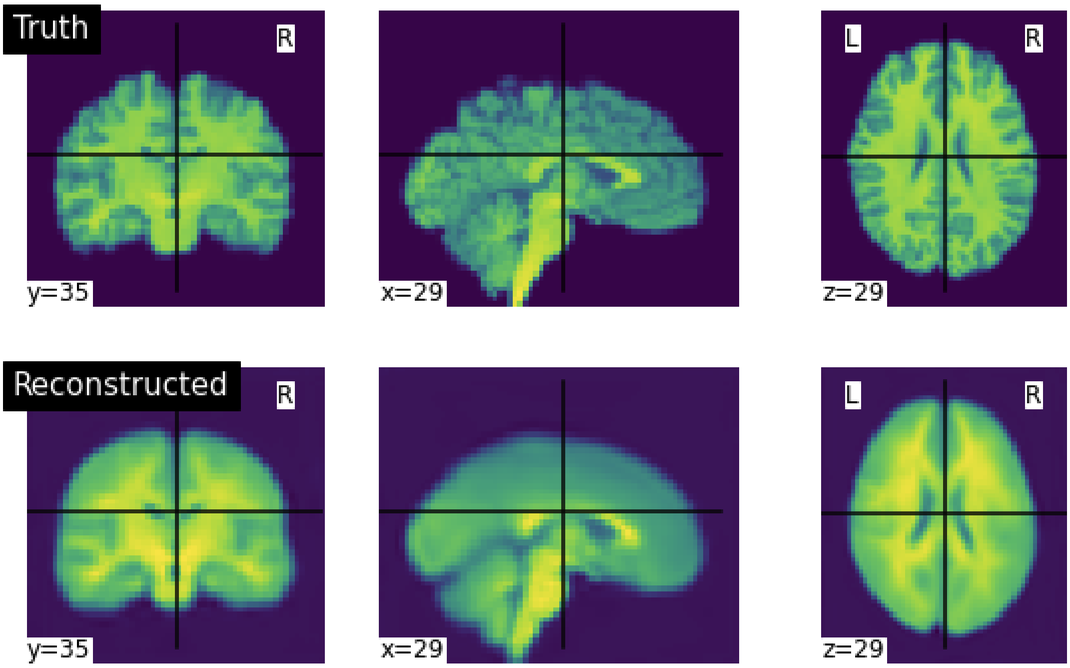
Results

1. Results of estimation data analysis

The baseline model for classification which was SVM applied to original brain images reshaped to vetor achieved the accuracy score of 0.902 (as the mean of 5 folds). This accuracy is considered as baseline for all autoencoders.

2. Results of first experiments

The first version of the autoencoder consisted of a large amount of convolutions and thus the receptive field was too large to learn details of the brain, however this model managed to accurately reconstruct the smooth shape of the original brain. Despite the fact that reconstructed images are visually far from perfect the obtained latent space allowed to achieve 0.863 accuracy in the classification. The comparison of original and reconstructed samples of images are shown on the figure.



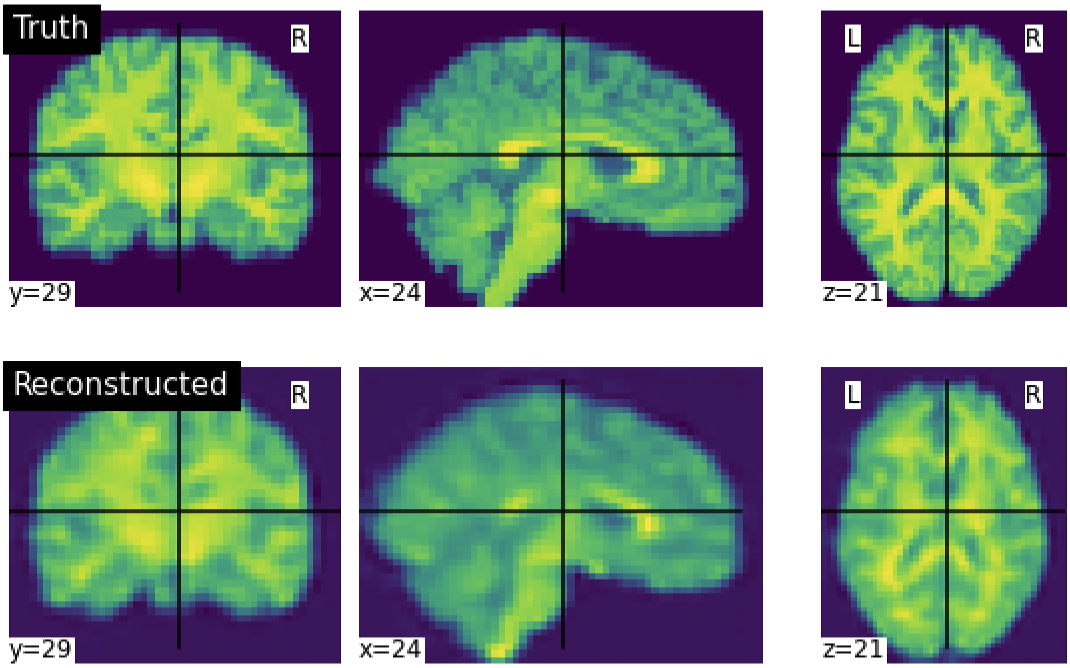
These observations led to experimenting with block types, number of channels and embedding shape. First version of autoencoder was considered as introductory material and required further enhancement.

3. Results of block architecture experiments with constant channels

Originally the chosen sizes of convolution layers allowed the autoencoder to have no more than 4 blocks. As it was explained and expected, the number of blocks highly affected not only the look of reconstructed images and the sizes of embeddings. The increase in the number of channels in each convolution resulted in an increase in the quality of the reconstructed image and embedding shape at the same time. Considering the fact that we wanted to meaningfully compress original images, we had to balance between the size of the embedding and the quality of the reconstructed image, which was the number of channels and blocks in terms of architecture. The results of experiments are presented in table 2.

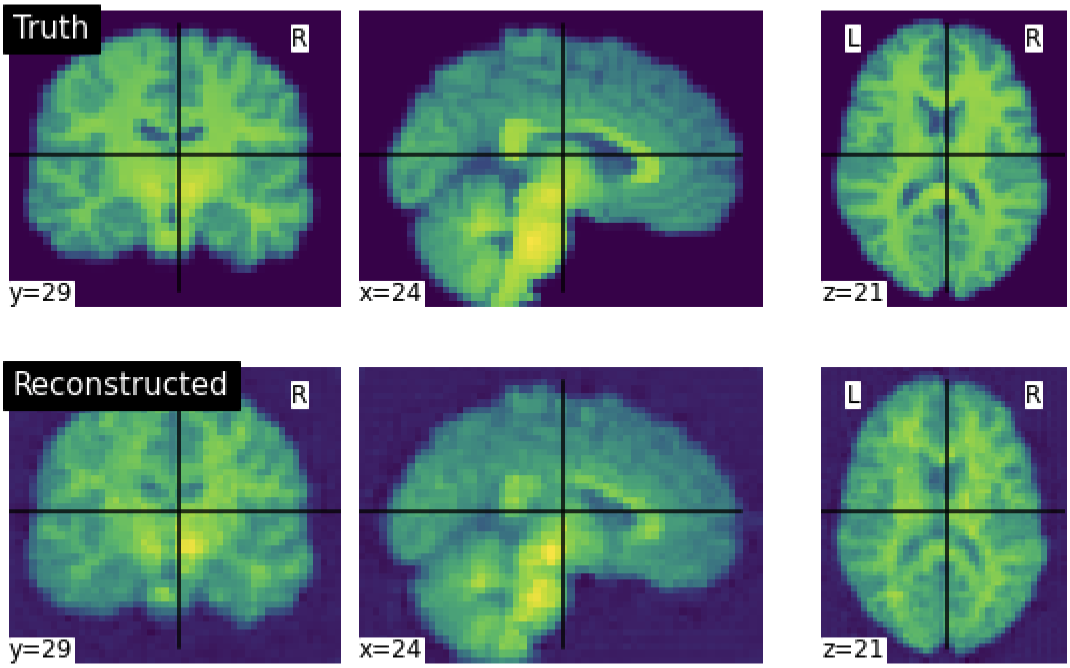
blocks	channels	MSE	accuracy
1	16	0.00245	0.922
1	32	0.00237	0.927
1	64	0.00231	0.931
2	16	0.00475	0.907
2	32	0.00451	0.910
2	64	0.00396	0.914
3	16	0.01706	0.902
3	32	0.01424	0.902
3	64	0.01398	0.913
4	16	0.05423	0.876
4	32	0.05376	0.883
4	64	0.05243	0.885

Considering the size of embedding, the best model was chosen to be 3 block 16 channels with embedding/original image ratio 35 : 1. Comparison of reconstructed and original images for this model is shown at picture.



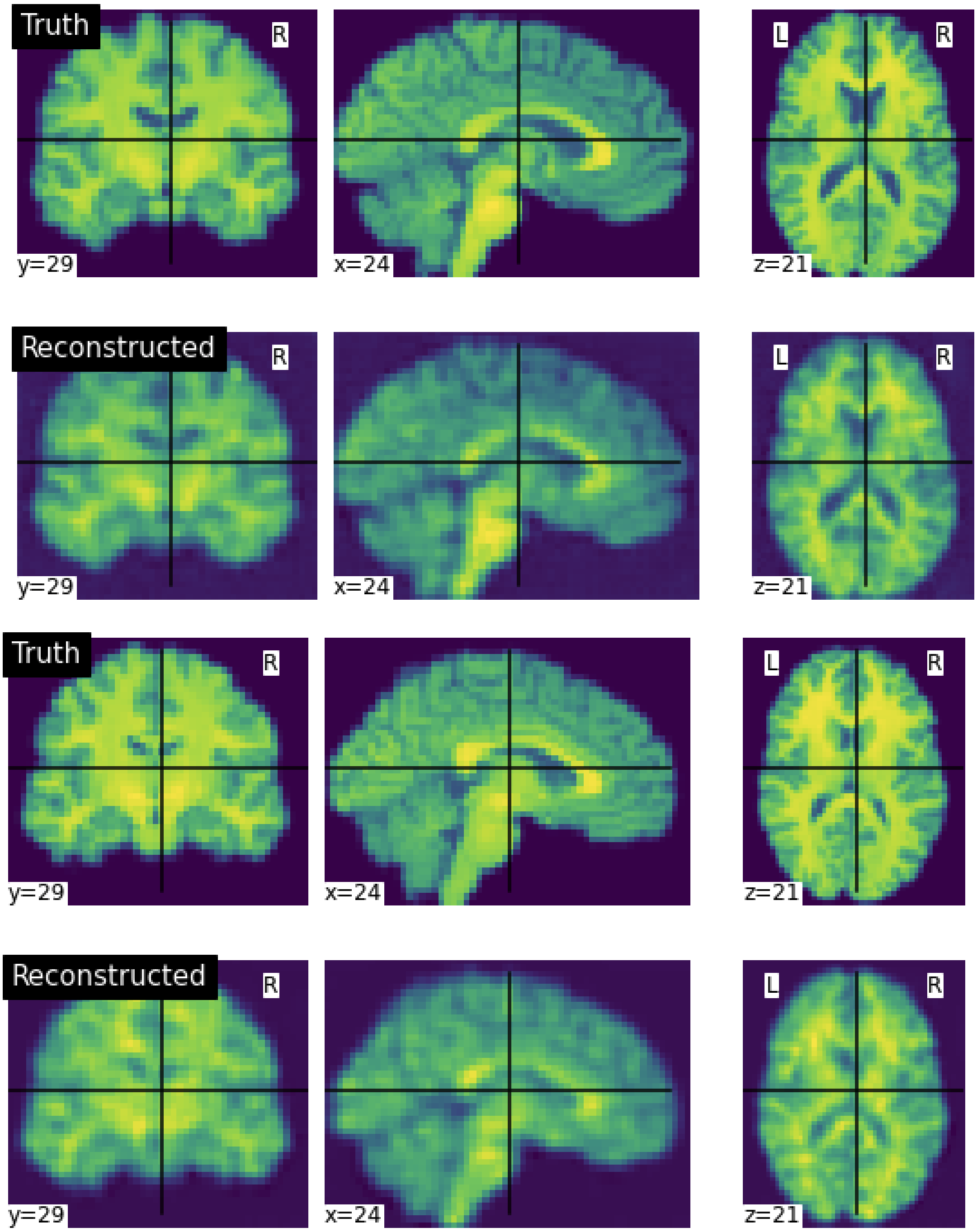
4. Results of constant embedding size with different number of channels

In the series of previous experiments it was decided that we should keep the embedding/original ratio of 1:35. The best model in this experiment was found to be a 3 block autoencoder with transition from 16 to 32 channels in convolutions. Comparison of reconstructed and original images for this model is shown in the picture. Classification accuracy score of this model is 0.909 and MSE loss is 0.017.

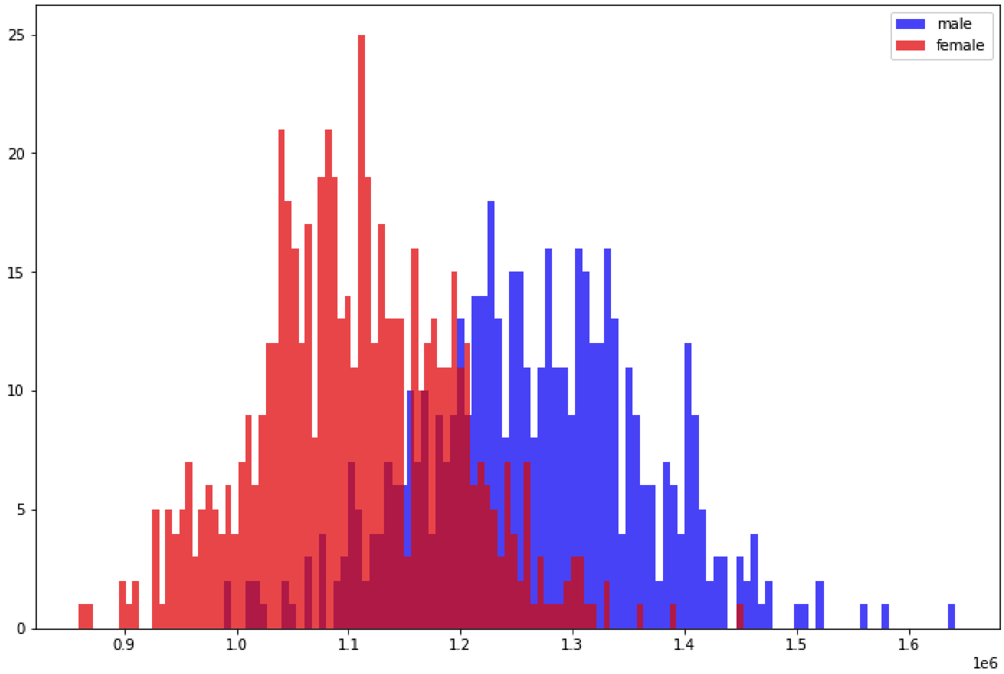


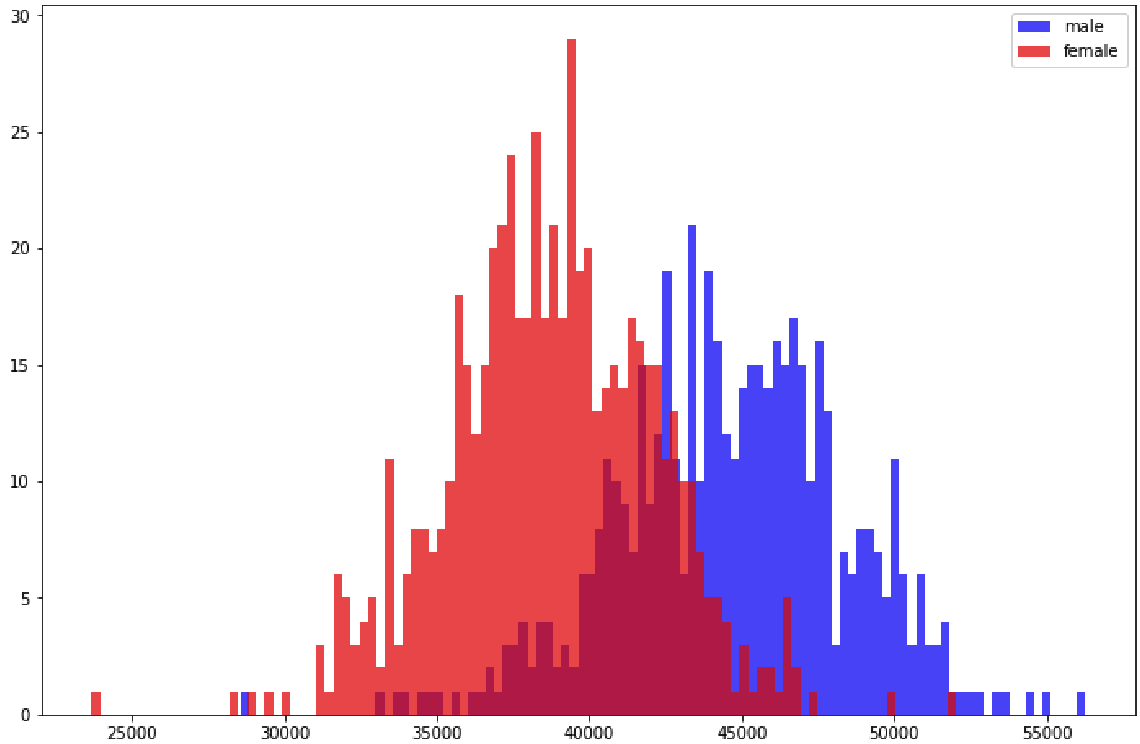
5. Results of variational autoencoders

In the process of training we were tracking both MSE and KL divergence separately. Considering constraints on embedding size and previous success, the architecture of 3 blocks was chosen to experiment with. Before measuring classification metrics, the training process showed that all models managed to achieve similar indicators of MSE and KL divergence. The best model was chosen to have 3 blocks with constant 32 channels in convolutions and means and variances vectors of size 3360 each.



As it is shown in the pictures, the images became a bit blurry but still pretty detailed. However, the classification accuracy was 0.542. This result was unexpected as the score is close to random prediction and it required further investigation. The first hypothesis of these results was that while approximating normal distribution, the distributions of male and female key features for classification were mixed and it is impossible to distinguish them. According to previous works, the key feature for sex classification was the brain volume. In order to prove our hypothesis we compared distributions of brain volumes of original and reconstructed images. However volume distribution was not corrupted by VAE as it is shown on pictures below. In fact, it was the same for the best normal AE.





The numerical comparison using p-value of these features also did not show any anomalies. At this point, there is no explanation for the low accuracy score of VAE classification.

6. Total results

Considering the results of all experiments we can present the following table of measurements:

Model	Channels	Blocks	MSE	Accuracy	KL
AE	16	3	0.01706	0.902	-
AE	32	3	0.01424	0.902	-
AE	16-32	3	0.017	0.909	-
VAE	32	3	0.020	0.542	0.018

Conclusion

Convolutional autoencoders are proven to be useful in computer vision and neuroimaging in particular. Considering the fact that MRI images data is usually very large and sensitive to sources that they are obtained from, it is common to use feature extraction and domain generalizing methods. In this study we have mentioned vanila convolutional autoencoders and variational convolutional autoencoders and considered and compared different architectures for these models. We have used them to obtain low-dimensional representations of MRI images, analyzed the obtained latent spaces and even applied them to actual dataset in order to classify demographic signs of gender. Apart from that, we explained how latent space depends on parameters of the autoencoder and presented a comparison of original and reconstructed data.

In future study, we would like to consider another usage of convolutional neural networks in neuroimaging which is generating data. To achieve that, generative adversarial networks should be studied and adapted for 3D brain MRI usage.

References

1. Ilse M., Tomczak J. M., Louizos C., Welling M.: DIVA: Domain Invariant Variational Autoencoders // 27 Mar 2019 (modified: 02 Jul 2019) ICLR 2019 Workshop DeepGenStruct Blind Submission
2. Hong S., Marinescu R., Dalca A. V., Bonkhoff A. K., Bretzner M., Rost N. S., Golland P.: 3D-StyleGAN: A Style-Based Generative Adversarial Network for Generative Modeling of Three-Dimensional Medical Images // September 2021 Deep Generative Models, and Data Augmentation, Labelling, and Imperfections (pp.24-34)
3. Yamaguchi H., Hashimoto Y., Sugihara G., Miyata J., Murai T., Takahashi H., Honda M., Hishimoto A., Yamashita Y. Three-Dimensional Convolutional Autoencoder Extracts Features of Structural Brain Images With a "Diagnostic Label-Free" Approach: Application to Schizophrenia Datasets // 07 July 2021 Front. Neurosci.
4. Pinaya W. H. L., Mechelli A., Sato J. R.: Using deep autoencoders to identify abnormal brain structural patterns in neuropsychiatric disorders: A large-scale multi-sample study // 2019 Feb 15;40(3):944-954. doi: 10.1002/hbm.24423. Epub 2018 Oct 11.
5. Arnaiz-Rodríguez A., KanberDeep B.: Convolutional autoencoders for reconstructing magnetic resonance images of the healthy brain // 19 Jan 2021 Universitat Oberta De Catalunya (UOC)
6. Kwon G., Han C., Kim D.: Generation of 3D Brain MRI Using Auto-Encoding Generative Adversarial Networks // 7 August 2019 MICCAI
7. Wolleb J., Sandkuhler R., Bieder F., Barakovic M., Papadopoulou A., Hadjikhani N., Yaldizli O., Kuhle J., Granziera C., Cattin P. C.: Learn to Ignore: Domain Adaptation for Multi-Site MRI Analysis // Pub Date : 2021-10-13, DOI: arxiv-2110.06803
8. Rocca J.: Understanding Variational Autoencoders (VAEs) // Sep 24, 2019 Towards Data Science
9. RElbers: Region Mutual Information loss // 30 March 2021 Github
10. Trencseni M.: Training a Pytorch Wasserstein MNIST GAN on Google Colab // 03 March 2021 Bytepawn
11. Khandelwal R.: How to improve image generation using Wasserstein GAN? // May 16, 2021 Medium / MLearning.ai
12. Barahlush: Generative models for 3D MRI scans // 27 Mar 2021 Github