Leveraging Modern Image Classification Algorithms for Medical Image Analysis and Diagnostics

Azamat Ilyassov

Date: 24/06/2024

Abstract

Image classification is a long-standing and one of the most researched problems of computer vision and a field of modern neural networks. Image classification is often associated with the introduction to the world of deep learning techniques and the field of modern artificial intelligence. Modern image classification models have become so powerful in handling images with 3 channels. In this paper ResNet1d was proposed as a method that uses principles of modern image classification approaches that are considered state-of-the-art.

Table of contents

Table of figures

1. Introduction

Image classification is a long standing problem and is probably one of the oldest computer vision problems. The problem of image recognition's history traces back to the middle of the 20th century, when first neural network structures based on perceptrons (Rosenblatt, 1958) applied to the task of classifying images. Modern solutions involve deep learning techniques based on convolution neural network layers (Lecun et al., 1998). Such solutions usually are mostly adapted to 3-channelled images. But when it comes to processing inputs with only one channel and applying transfer learning techniques, either amendments to a pre-trained structure should be applied or 1-channels input should be processed to 3-channel one when applying data preprocessing. This is usually encountered when trying to solve various medical imaging tasks, such as assisting in diagnosing based on MRI results.

In this paper various image classification approaches are explored. Their pros and cons are discussed and compared to each other. Aside from analysing current methodologies, a new approach is proposed - ResNet1d. This architecture is based on ResNet design (He et al., 2015) and adapted to 128x128 images with only one channel. Such approach employs classic principles of ResNet solution and applies its pros and cons to 1-channelled inputs without changing original architecture or amending source data, which may impact further fine-tuned model performance.

Proposed model was trained on an Alzheimer_MRI dataset published by Hugging Face (Salieh, 2023) through Kaggle (Kaggle, 2022) notebooks using cloud kernels with GPUs.

The outputs and conclusion of this article might give researchers a ground for further solutions regarding medical imaging and any other approaches to solve image classification tasks involving images with 1 channel.

2. Related literature

2.1 Image Classification before Convolutional Neural Network

When classic and basic perceptron structure was first introduced potential applications including pattern recognition tasks were discussed as well (Rosenblatt, 1958). (Minsky & Papert, 1969) highlighted the inability of perceptron to solve non-linear tasks, which led to the development of more complex structures in order to apply them to the problem of classifying images. However, due to limitations of neural networks and their inability to prove effectiveness to solve image recognition, the field started to lack popularity in society and no significant changes were made until the development of back-propagation algorithms were developed (Kushwaha, 2023).

(Rumelhart et al., 1986) constructed a back-propagation algorithm, which enhanced the performance of a multi-layer perceptron structure showing promising potential in further solution of image classification tasks. Main issues were disappearing gradients and over-fitting (Kushwaha, 2023).

2.2 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) rely on three basic principles: local receptive fields, shared weights and sub-sampling . Local receptive fields are kernels that follow the shared weights principle by sharing the same set of weights. Sub-sampling helps with decreasing dimensionality and averaging and decreasing the resolution of the map, since the lower the resolution the better the representation of features extracted from image input (Lecun et al., 1998). In this paper, authors not only introduced a novel approach to processing images, but also demonstrated an application of a given solution in a neural network named LeNet-5 (Figure 1).

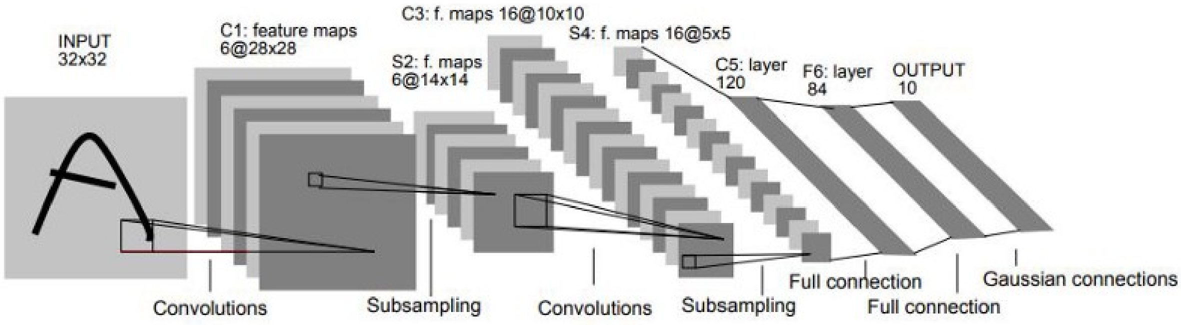


Figure 1. LeNet-5 design (Lecun et al., 1998)

2.3 AlexNet

AlexNet (Krizhevsky et al., 2012) is considered to be a pioneering and classic CNN-based approach representing image classification (Saxena, 2021). It simply consists of CNNs, pooling layers (Gholamalinezhad & Khosravi, 2020) and dense linear layers (Javid et al., 2020) shown in Figure 2.

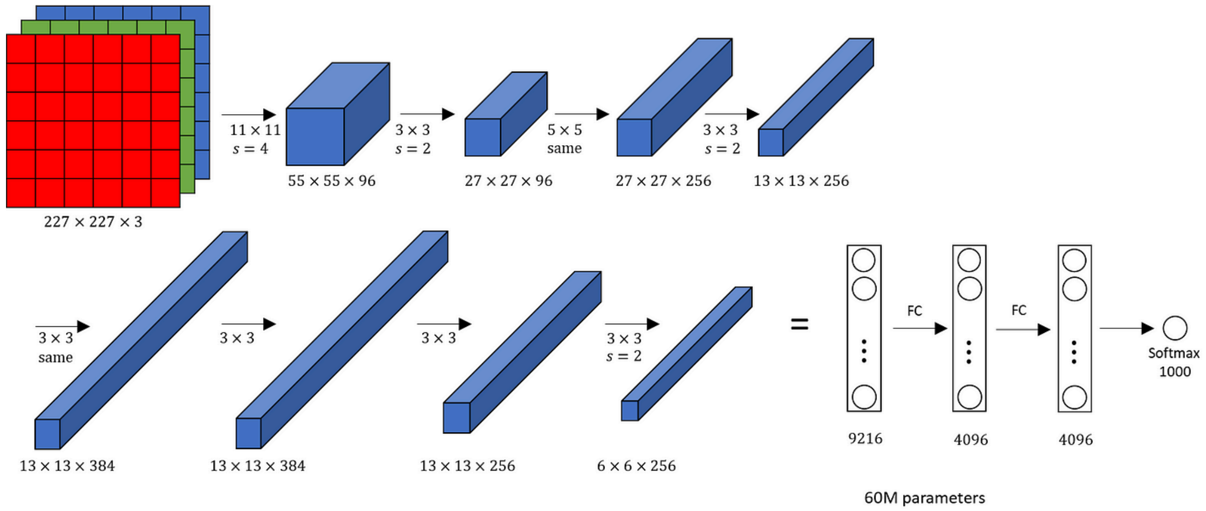


Figure 2. AlexNet architecture (Krizhevsky et al., 2012)

AlexNet structure is simple and easy to understand, hence it is thought of as an introductory and basic classic image classification approach. Even though the structure is easy to understand and implement, trained AlexNet is heavy, requires high computational cost and has slow inference time. But this paper opened a way for more advanced techniques to be developed.

2.4 Visual Geometry Group

(Simonyan & Zisserman, 2015) developed a better approach that managed to set up a new standard for image classification solutions and is considered to be SOTA. The architecture was named Visual Geometry Grouped Network (VGGNet). This approach used tiny convolutional filters tied up with a rectified linear unit (ReLU) (Abien Fred Agarap, 2018) and pooling layers (Gholamalinezhad & Khosravi, 2020) in one feature extraction block. After features are extracted, they are then used to tune Linear layers (Javid et al., 2020) parameters. Classic VGG-16 architecture is portrayed in Figure 3.

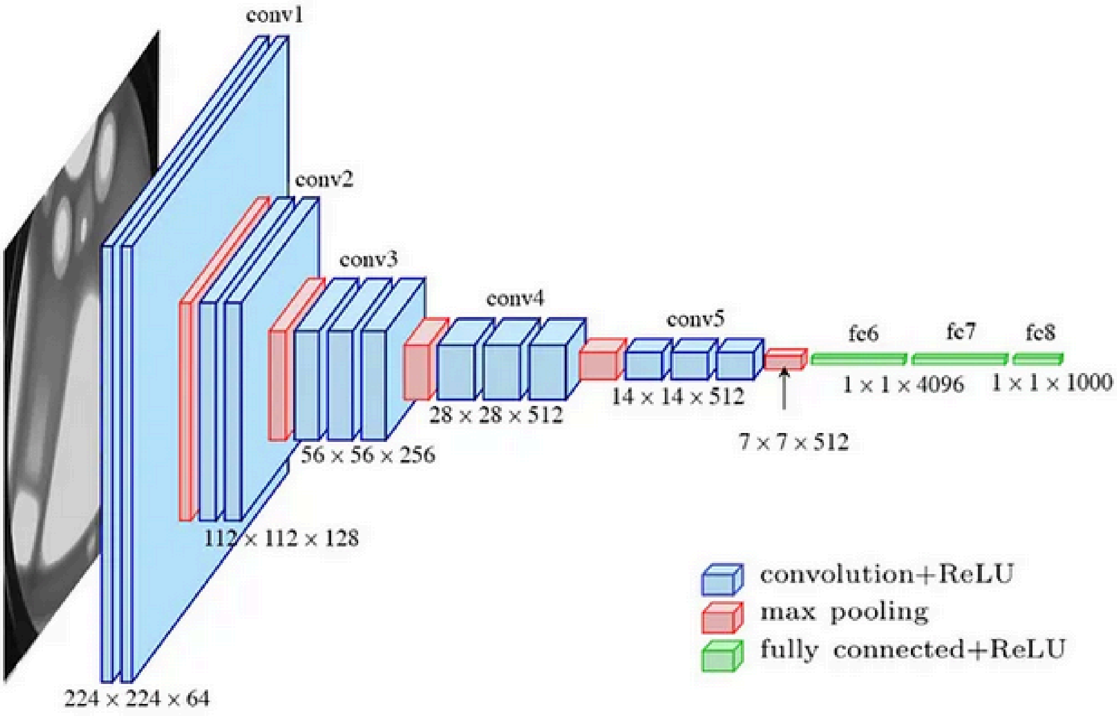


Figure 3. VGG-16 architecture (Simonyan & Zisserman, 2015)

VGG manages to achieve competitive accuracy and is still being held in high regard when it comes to image classification and other computer vision problems, such as style transfer (Johnson et al., 2016) or object detection (Boesch, 2021). However, VGG lacks computational efficiency with a relatively slow inference time due to high number of parameters and hence is very expensive.

2.5 ResNet

(He et al., 2015) decided to incorporate shortcuts as means of implementing a residual learning strategy in order to boost image classification performance and accuracy. Residual learning addresses the degradation problem and instead of hoping that stacked layers will fit into the desired mapping, residual mappings are applied. Residual learning has the following notation: F(x) + x. Such a framework had the following design for building blocks:

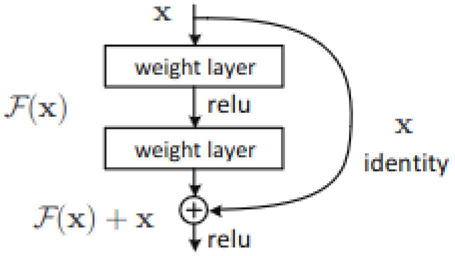
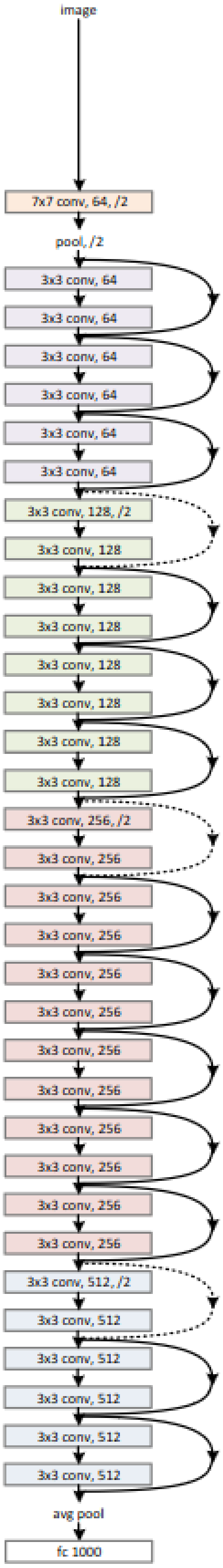

x
weight layer
relu
weight layer
$\mathcal{F}(\mathbf{x})$
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Figure 4. Residual block (He et al., 2015)

ResNet architecture has the following architecture:



34-layer residual

Figure 5. ResNet architecture (He et al., 2015)

ResNet is now referred to as the SOTA model and is usually used for transfer learning, due to its adaptability, low computational cost and high inference speed. Such a solution is also used for such tasks as object detection (Suherman et al., 2023) and style transfer (Wang et al., 2021). One of its drawbacks is complexity and susceptibility to over-fitting problems.

2.6 Vision Transformer

(Dosovitskiy et al., 2020) introduced a fusion of natural language processing (NLP) and computer vision techniques dedicated to image processing. Architecture proposed is named Vision Transformer or ViTand has the following design (Figure 6):

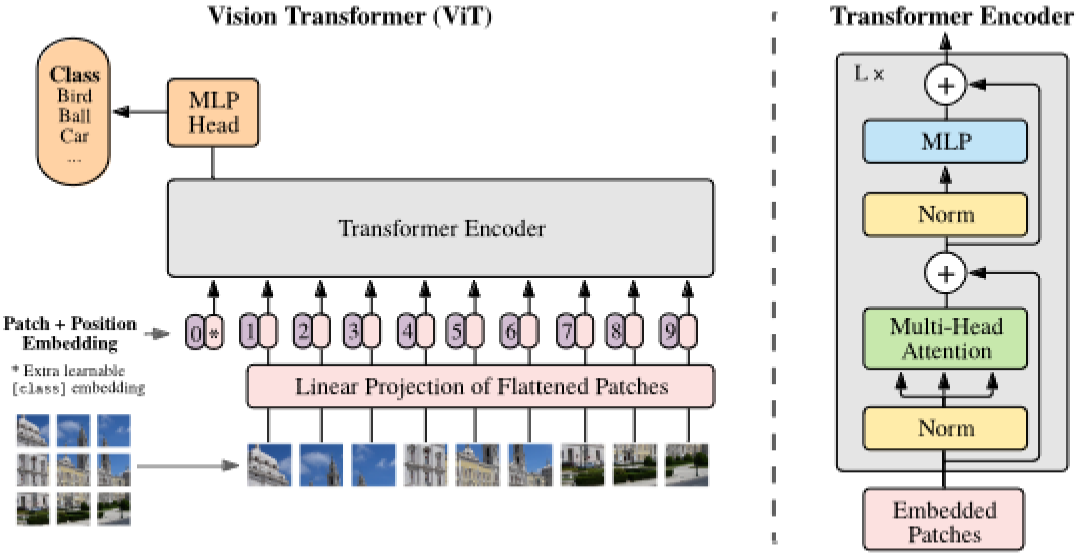


Figure 6. ViT design (Dosovitskiy et al., 2020)

(Vaswani et al., 2017) presented a novel approach for NLP tasks. ViT has such useful features that stands out to CNN -based approach as adaptability to various inputs and ability to capture long range dependencies which is vital in managing images with a vast range of quality, considering datasets dedicated to transfer learning tasks may contain corrupted images or when used for real world images - some inputs may contrast to the ones models were used to be trained on.

ViT is considered to be SOTA and has more applications to computer vision field than image classification such as image deblurring (Liu et al., 2024) and semantic segmentation (Zhang et al., 2023). Despite its ability to learn complex features and adaptability, it is very expensive to train, due to the fact that self-attention layers complexity increases quadratically when the resolution of images is high.

3. Methodology

The main task is to develop a solution that would be able to assist in dealing with the problem of diagnosing based on single channelled medical images. In order to deal with the main problem, the classic approach would be used, which consists of: getting the data, preprocessing, solution architecture, training and evaluation. Such an description better reminds the trace that of CRISP-DM methodology (Wirth & Hipp, 2000):

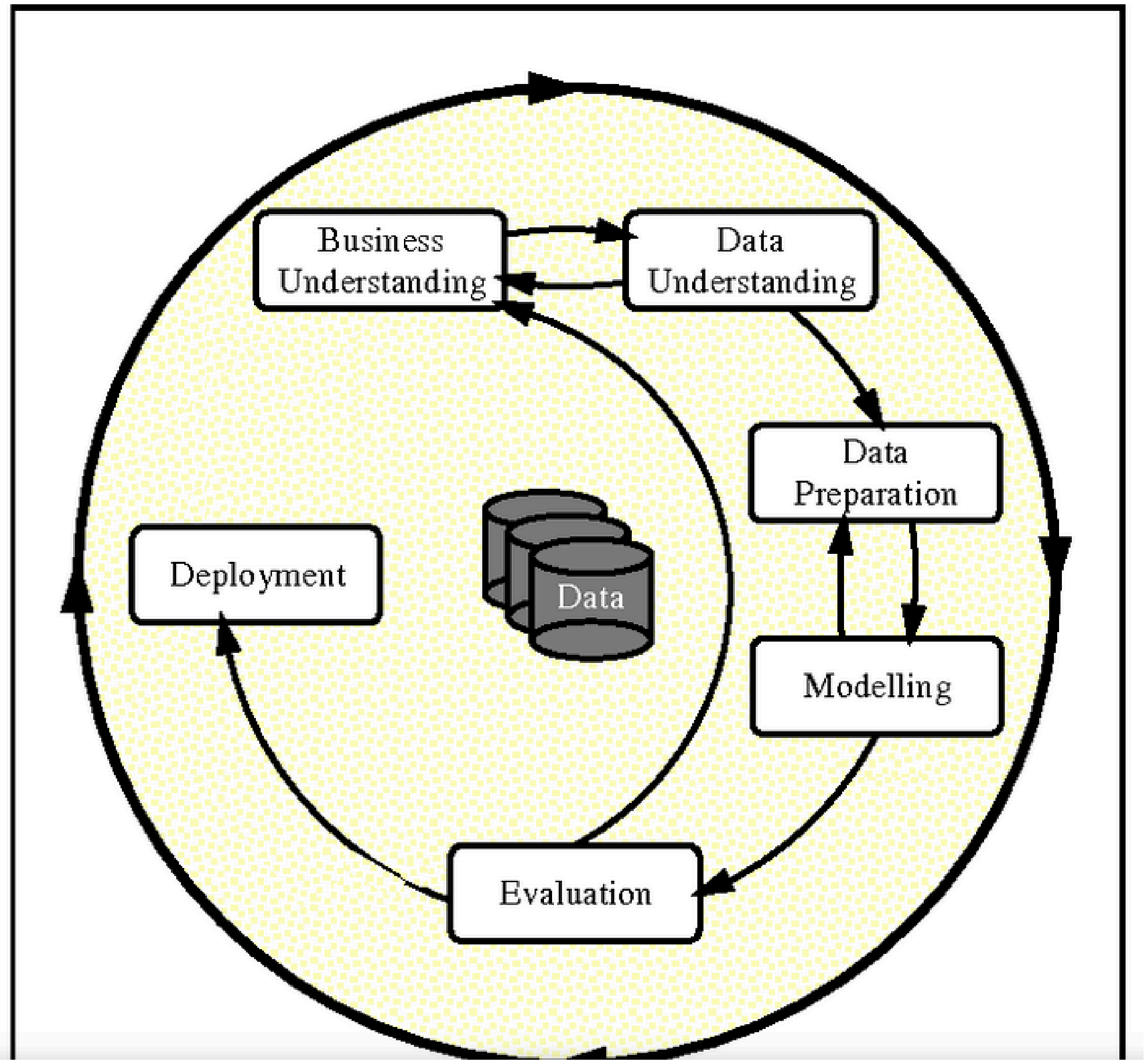


Figure 7. CRISP-DM approach originally designed for solving Data Mining tasks (Wirth & Hipp, 2000)

3.1 Dataset

In order to accomplish the given problem of classifying images with 1 channel, the MRI Alzheimer dataset was used (Salieh, 2023). This dataset was provided by Hugging Face as an open-source dataset. There are 5120 images for training and 1280 images for testing with 4 versions of Alzheimer disease diagnosis:

- 0 - Mild Demented
- 1 - Moderate Demented
- 2 - Non Demented
- 3 - Very Mild Demented

3.2 ResNet1d

In order to develop a proper solution the most classic methodology principles would be used. As was mentioned ResNet model (He et al., 2015) has become a SOTA model that is unusually light, fast and easy to use and fine-tune. Despite its drawbacks, it is still being used for tasks other than image recognition (Suherman et al., 2023), (Wang et al., 2021).

ResNet architecture was re-implemented using the original design (He et al., 2015) shown on Figure 5 adapted to processing smaller images with 1 channel with slight alterations to channels and kernel sizes as well as strides and paddings.

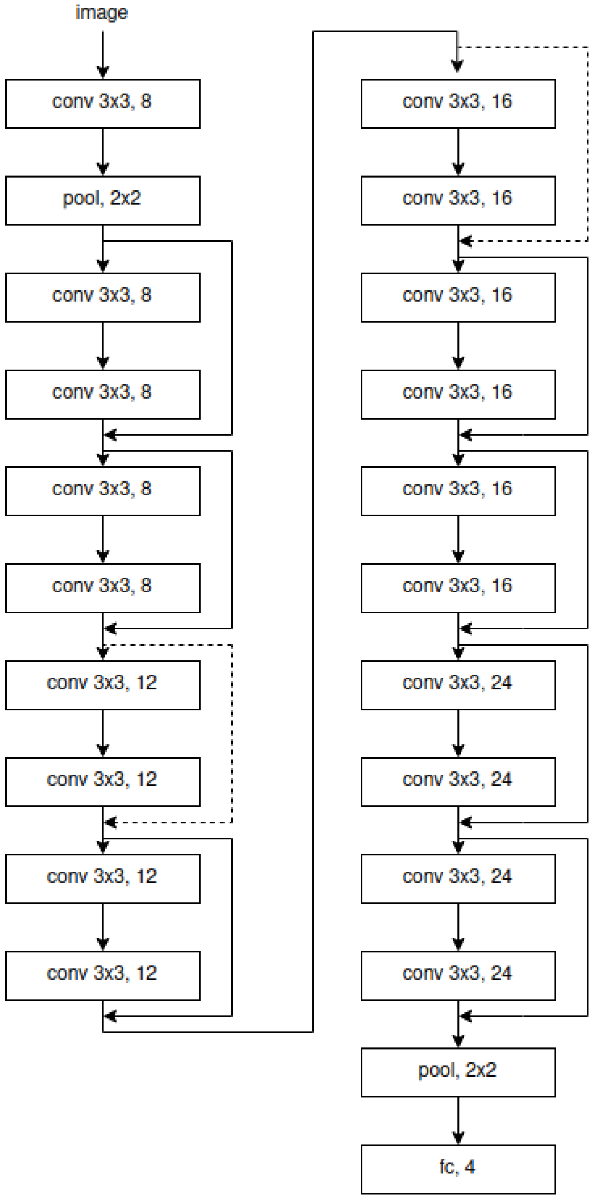


Figure 8. ResNet1d

3.3 ResNet50-1d

Another proposed solution that would be used as comparison is fine-tuning a slightly modified version of a pre-trained ResNet50 (He et al., 2015), that is capable of processing 1-channel images. The change that was made to the original ResNet50 structure is a single depthwise convolutional layer with 1x1 kernel that yields a tensor with 3 channels. This approach can be called ResNet50-1d. This solution is best explained through Figure 8.

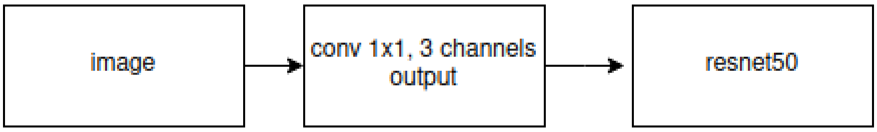


Figure 8. ResNet50-1d solution

4. Experiments and results

4.1 Training parameters and environment

ResNet1d and ResNet50-1d had very similar training parameters, except for the number of epochs required and hence the scheduler step size was different for each solution. Both had Stochastic Gradient Descent (Kiefer & Wolfowitz, 1952) as an optimisation algorithm. ResNet1d had 60 epochs to train with learning rate starting at 0.1 and update coefficient 0.1 each 20 epochs. ResNet50-1d had only 20 epochs with the same update coefficient each 10 epochs, since this model was already pre-trained on ImageNet dataset (Deng et al., 2009), hence less epochs required.

Both models were trained using Kaggle notebooks and cloud kernels with GPUs provided by Kaggle.

4.2 ResNet1d performance

ResNed1d performed reasonably well, considering the relatively small dataset and small number of epochs required to train the model from scratch. Loss and accuracy logs during training of the model are depicted on Figure 9.

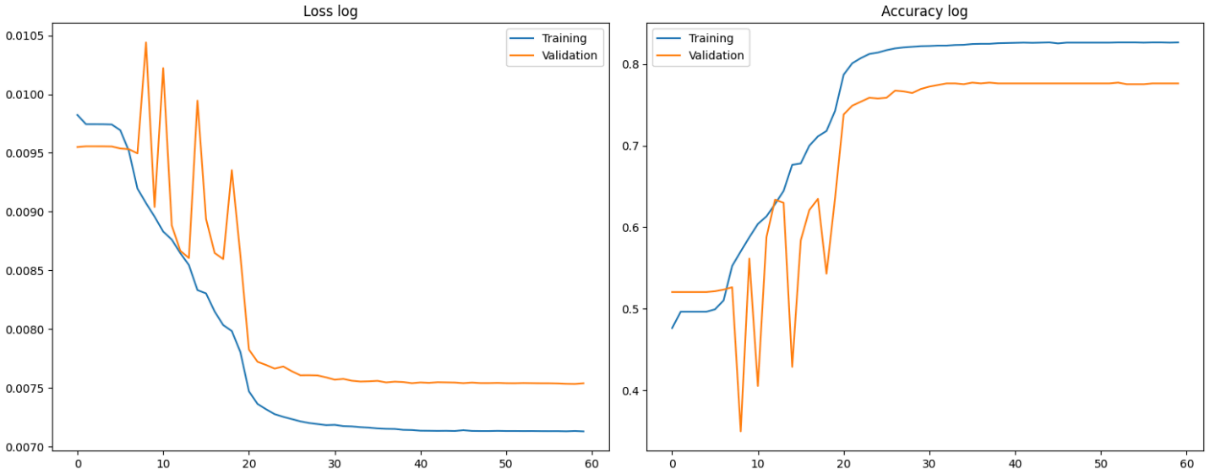
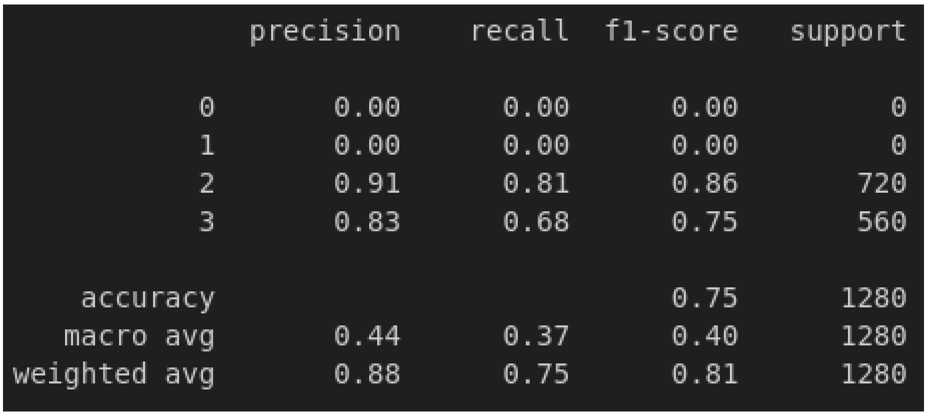


Figure 9. ResNet1d performance log

During the first 20 epochs the model had a very inconsistent track of validation loss and accuracy, but after the 20th epoch the learning rate is updated and becomes smaller, hence less fluctuations and more consistent learning curve. This indicates an efficient training approach and consistent learning curve growth.

However, when model was tested on a test set and the results were evaluated using through classification reporting techniques, we have the following outcome:



```
              precision    recall  f1-score   support

           0       0.00      0.00      0.00         0
           1       0.00      0.00      0.00         0
           2       0.91      0.81      0.86       720
           3       0.83      0.68      0.75       560

    accuracy                           0.75      1280
   macro avg       0.44      0.37      0.40      1280
weighted avg       0.88      0.75      0.81      1280
```

Figure 10. Classification report for ResNet1d

Overall accuracy score reached 75%, which is relatively good, however, the model is biassed towards labels 2 and 3, rather than 0 and 1.

4.2 ResNet50-1d performance

Figure 10 illustrates that ResNet50-1d had less fluctuations in loss and accuracy rates before the first learning rate update compared to ResNet1d. This implies that a pre-trained model performs better than a smaller one trained from scratch.

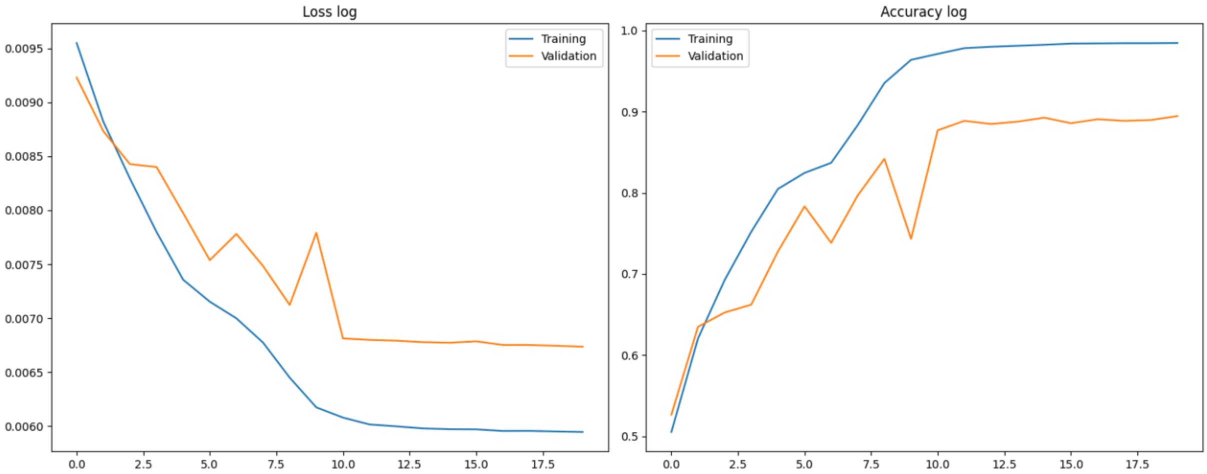
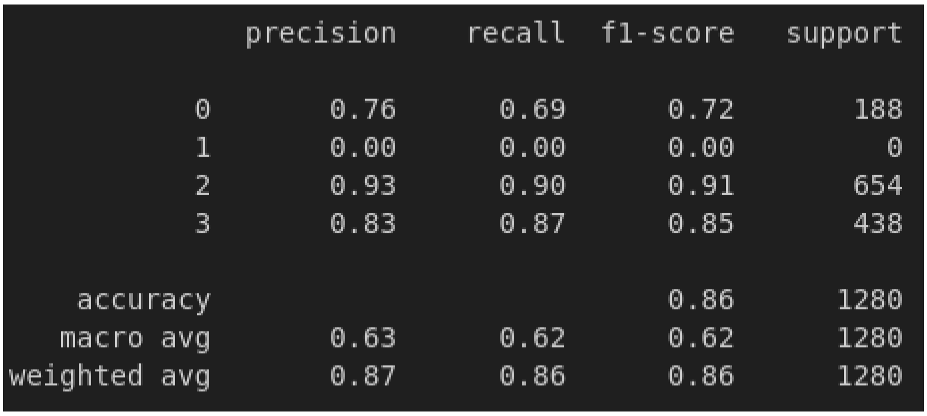


Figure 11. ResNet50-1d performance log

Thorough classification evaluation on Figure 12 shows less bias towards labels 2 and 3 and more strength to label 0, which leads to higher overall accuracy score. However, due to the lack of examples with label 1 or due to the lack and weakness of training approach and techniques used, label 1 was predicted:



```
              precision    recall  f1-score   support

           0       0.76      0.69      0.72       188
           1       0.00      0.00      0.00         0
           2       0.93      0.90      0.91       654
           3       0.83      0.87      0.85       438

    accuracy                           0.86      1280
   macro avg       0.63      0.62      0.62      1280
weighted avg       0.87      0.86      0.86      1280
```

Figure 12. ResNet50-1d classification report breakdown

4.3 Pros and cons

ResNet1d showed potential to become a proper model for medical imaging usage for diagnostic purposes. It is obviously smaller and faster, than original ResNet models (He et al., 2015), because it processes smaller and less requiring images with smaller channels and less features to be extracted. This is a simple and straightforward approach to processing medical images.

ResNet1d showed its weaknesses during training, by being very inconsistent in the beginning of the training with what turned out to be a high learning rate of 0.1, which implies a slight over-training problem. The ResNet50-1d solution was used as a comparison and as another more powerful approach to the current medical imaging problem. Compared to the main proposed solution, ResNet50-1d showed less bias and more consistent training and more potential, due to the model's pre-trained nature. This implies that ResNet1d requires a bigger dataset and more advanced implementation and training methodologies, but proved to have potential, since it has similarities during training performance.

5. Conclusion

Modern image classification approaches are mostly used for processing images with 3 channels. However, not many known solutions are adapted for images with only one channel, which can be encountered in the medical imaging field for diagnostic procedures. ResNet1d applies principles used in (He et al., 2015) and is adapted to processing images with 1 channel. This approach showed great potential, despite having bias and a slight risk of over-training. This model yielded a pattern similar to the one used for alternative powerful approach ResNet50-1d a pre-trained ResNet50 with slight changes to its structure adapted to processing single channelled images.

MRI Alzheimer dataset (Salieh, 2023) was used for training and demonstration of the proposed solution potential strengths and risks.

Kaggle notebooks were used as training environment using their cloud kernels with GPUs.

References

[1] Kushwaha, N. (2023, August 30). A Brief History of the Evolution of Image Classification. Medium. https://python.plainenglish.io/a-brief-history-of-the-evolution-of-image-classification-402c63baf50

[2] Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review, 65*(6), 386–408. https://doi.org/10.1037/h0042519

[3] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278-2324

[4] He, K., Zhang, X., Ren, S., & Sun, J. (2015, December 10). *Deep Residual Learning for Image Recognition*. ArXiv.org. https://arxiv.org/abs/1512.03385

[5] Salieh, F. G. (2023). Hugging Face. https://huggingface.co/datasets/Falah/Alzheimer_MRI

[6] Kaggle. (2022). *Kaggle: Your Home for Data Science*. Kaggle.com. https://www.kaggle.com/

[7] Minsky, M., & Papert, S. (1969). An introduction to computational geometry. *Cambridge tiass., HIT*, *479*(480), 104.

[8] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, *323*(6088), 533-536.

[9] Saxena, S. (2021, March 19). *Alexnet Architecture | Introduction to Architecture of Alexnet*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/03/introduction-to-the-architecture-of-alexnet/

[10] Gholamalinezhad, H., & Khosravi, H. (2020). Pooling Methods in Deep Neural Networks, a Review. *ArXiv:2009.07485 [Cs]*. https://arxiv.org/abs/2009.07485

[11] Javid, A. M., Das, S., Skoglund, M., & Chatterjee, S. (2020, October 22). *A ReLU Dense Layer to Improve the Performance of Neural Networks*. ArXiv.org. https://doi.org/10.48550/arXiv.2010.13572

[12] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Communications of the ACM*, *60*(6), 84–90. https://doi.org/10.1145/3065386

[13] Simonyan, K., & Zisserman, A. (2015, April 10). *Very Deep Convolutional Networks for Large-Scale Image Recognition*. ArXiv.org. https://arxiv.org/abs/1409.1556

[14] Abien Fred Agarap. (2018). Deep Learning using Rectified Linear Units (ReLU). *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1803.08375

[14] Johnson, J., Alahi, A., & Fei-Fei, L. (2016). *Perceptual Losses for Real-Time Style Transfer and Super-Resolution*. ArXiv.org. https://arxiv.org/abs/1603.08155

[15] Boesch, G. (2021, October 6). *VGG Very Deep Convolutional Networks (VGGNet) - What you need to know*. Viso.ai. https://viso.ai/deep-learning/vgg-very-deep-convolutional-networks/

[16] He, K., Zhang, X., Ren, S., & Sun, J. (2015, December 10). *Deep Residual Learning for Image Recognition*. ArXiv.org. https://arxiv.org/abs/1512.03385

[17] Suherman, Endang & Rahman, Ben & Hindarto, Djarot & Santoso, Handri. (2023). Implementation of ResNet-50 on End-to-End Object Detection (DETR) on Objects. SinkrOn. 8. 1085-1096. 10.33395/sinkron.v8i2.12378.

[18] Wang, P., Li, Y., & Vasconcelos, N. (2021). Rethinking and Improving the Robustness of Image Style Transfer. *ArXiv:2104.05623 [Cs, Eess]*. https://arxiv.org/abs/2104.05623

[19] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ArXiv:2010.11929 [Cs]*. https://arxiv.org/abs/2010.11929

[20] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017, June 12). *Attention Is All You Need*. ArXiv.org. https://arxiv.org/abs/1706.03762

[21] Liu, H., Li, B., Liu, C., & Lu, M. (2024, March 19). *DeblurDiNAT: A Lightweight and Effective Transformer for Image Deblurring*. ArXiv.org. https://doi.org/10.48550/arXiv.2403.13163

[22] Zhang, B., Liu, L., Phan, M. H., Tian, Z., Shen, C., & Liu, Y. (2023, June 9). *SegViTv2: Exploring Efficient and Continual Semantic Segmentation with Plain Vision Transformers*. ArXiv.org. https://doi.org/10.48550/arXiv.2306.06289

[23] Wirth, R., & Hipp, J. (2000). *CRISP-DM: Towards a Standard Process Model for Data Mining*. Semantic Scholar. https://www.semanticscholar.org/paper/CRISP-DM%3A-Towards-a-Standard-Process-Model-for-Data-Wirth-Hipp/48b9293cfd4297f855867ca278f7069abc6a9c24

[24] Kiefer, J., & Wolfowitz, J. (1952). Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 462-466.

[25] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/cvpr.2009.5206848