Lymphoma Classification
Data Augmentation Techniques to improve the dataset

Deep Learning Project[1]
Ferrari Davide[1], **Frighetto Eddy**[1], **Manuzzato Matteo**[1]
[1]Università di Padova
davide.ferrari.6@studenti.unipd.it, eddy.frighetto@studenti.unipd.it, matteo.manuzzato@studenti.unipd.it

Abstract

In this study, we address the common challenge in medical image classification posed by limited dataset sizes, particularly in the context of Lymphoma images. Our project aimed to develop and evaluate new data augmentation techniques to enhance the performance of neural network classifiers. We introduced three novel algorithms that apply distinct transformations to the images, thereby increasing the dataset size and diversity. These transformations involved modifying image properties and introducing structured noise, enhancing the datasets' overall quality. Our experimental results demonstrate that these new augmentation techniques improve classification accuracy, showcasing their potential in overcoming the limitations of small medical image datasets and advancing the field of medical image analysis.

Introduction

Lymphoma is the general term for cancer in the lymphatic system, the network of tissues, vessels and organs that help the body to fight infections. It's considered a blood cancer because the condition starts in the lymphocytes in the lymphatic system. (What is Lymphoma? 2023)
There are two main lymphomas, *Hodgkin Lymphomas* and *Non-Hodgkin Lymphomas*, and more than 70 subtypes.
Lymphoma happens when the white blood cells in the lymphatic system change into rapidly growing cancer cells that don't die. The majority of these genetic mutations that cause lymphoma happen without an identifiable cause.
Nowadays there are different technologies that help doctors to identify these types of cancers:

- Blood tests:
  - Complete Blood Count (CBC)
  - Erythrocyte Sedimentation Rate (ESR)
  - Lactate Dehydrogenase (LDH)
  - Liver and kidney function tests
- Imaging tests:
  - Computed tomography (CT) scan
  - Positron emissions tomography (PET) scans

The recent advancements in Deep Neural Networks (DNNs) have significantly enhanced the effectiveness of computer vision tasks. In the context of classifying different types of Lymphoma, having a sufficiently large dataset is essential for the optimal performance of DNNs.
In this project we used a dataset with 375 images of malignant Lymphoma subdivided into three different classes:

- Chronic Lymphocytic Leukemia (CLL)
- Follicular Lymphoma (FL)
- Mantle Cell Lymphoma (MCL)

However, the size of the dataset is too small to train a DNN, that usually needs large datasets to be trained.
We used some called **Data Augmentation techniques**, that allow us to create new (fake) images from the original ones, in order to increase the size and the robustness of the dataset.

Data Augmentation

Data Augmentation is a technique used to increase the amount of data by adding slightly modified copies of already existing training pattern or newly created synthetic data from existing patterns. It helps to reduce overfitting when training a machine learning model. Deep Neural Networks must be invariant to a wide variety of input variations. We want our model to be invariant to input variations, like pose, appearence, lighting and many other types. The best way towards better generalization is to train on more data, that often is limited.
The goal of data augmentation is to create fake images from the existing data, these one are generated randomly from each image in our batch at the beginning of every training iteration. There are many popular Data Augmentation algorithms like:

- Geometric transformations:
  - Image Cropping
  - Image Flipping
  - Affine Transformations
- Local Filters:
  - Gaussian Blur
  - Image Sharpening
  - Edge Detection

- Noise adding:
  - Gaussian Noise
  - Salt & Pepper Noise
- Color Transformations:
  - Contrast
  - Brightness

We used many of them during the testing of our project, usually combined with three algorithms developed by us, **RGBRotation**, **HSVRotation** and **GridColored**.
During the development of this project we exploited different data augmentation techniques:

- **DWTAverageFusion**: Applies Discrete Wavelet Transform (DWT) to fuse two images based on their wavelet coefficients.
- **RGBRotation**: Rotates the RGB channels around each pixel in a clockwise direction.
- **HSVRotation**: Rotates the Hue, Saturation, and Value channels of an image.
- **HSVSwap**: Randomly swaps the Hue and Saturation channels of an HSV image.
- **SaltAndPepper**: Adds salt and pepper noise to the image.
- **ShuffleSquares**: Shuffles square regions of the image.
- **RandomGeometricTransform**: Applies random geometric transformations like rotation and flipping.
- **Rotation**: Applies random rotation.
- **Flip**: Applies random flip.
- **GridColored**: Adds grid lines with random colors to the image.
- **RandomBrightness**: Applies random brightness adjustment.
- **RandomShifts**: Applies random shifts to the image.
- **ComboGeometricBrightness**: Combines random geometric transformations with random brightness adjustments.
- **ComboGeometricRGBRotation**: Combines random geometric transformations with RGB channel rotation.
- **ComboGeometricShift**: Combines random geometric transformations with random shifts.
- **ComboGeometricHSVRotation**: Combines random geometric transformations with HSV channel rotation.
- **ComboGeometricGridColored**: Combines random geometric transformations with grid coloring.
- **ComboGeometricShuffleSquares**: Combines random geometric transformations with shuffling squares.

RGB Rotation

RGBRotation is an algorithm that transforms images by rotating the Red (R), Green (G), and Blue (B) channels within the RGB color space. We implement this method taking idea from the data augmentation *RGBRotation* (Ozdemir, Dogan, and Kaya 2024). *RGB-Angle-Wheel* enhances image datasets by applying a rotation process to 3 × 3 patches of RGB images.
Specifically, in RGBRotation, the Red channel is rotated downward by one pixel, the Green channel is rotated rightward by one pixel, and the Blue channel is rotated upward by one pixel. This process introduces structured variations in color properties, resulting in a diverse set of augmented images that can improve the robustness and generalization of neural networks.

HSV Rotation

HSVRotation is an algorithm that transforms images by rotating the Hue (H), Saturation (S), and Value (V) channels within the HSV color space. This method is the HSV version of RGBRotation.
We explored both *HSVRotation* and *RGBRotation* to leverage different aspects of image properties and achieve more diverse data augmentation. RGB rotation modifies the primary color channels directly, introducing spatial variations in red, green, and blue intensities, helping the model learn from changes in color distribution and spatial relationships. In contrast, HSV rotation alters the Hue, Saturation, and Value channels, affecting the image's color tones, intensity, and brightness, thereby introducing variations in color properties not achievable through RGB rotation alone.

Grid Colored

GridColored is an algorithm that augments the input image by overlaying 10 random vertical and horizontal colored grid lines. Each grid line is drawn at a randomly selected row or column, and each pixel in the line is assigned a random RGB color. This method enhances the dataset's variety by adding organized noise, which can aid in boosting the robustness and generalization abilities of neural networks.

Combined data augmentation

In our project, we also tried to combine different data augmentation techniques to enrich the training set with different augmented images. In particular in each combo we generate with a specific probability images where are applied all or some augmentation to them. This approach diversifies the dataset and leads to a more robust neural network.

Data Splitting and Augmentation Strategy

To accurately estimate the performance of each data augmentation technique, we created augmented datasets for each technique and repeated the process over five iterations. For each iteration, we trained the model on the newly augmented dataset. This approach yielded five sets of results for each data augmentation technique. We then averaged these results to obtain a reliable performance measure for each technique.
Firstly, we divided the original dataset into three folders, each representing one class of Lymphoma images. For each iteration, we further split the original dataset into a training set (80%) and a test set (20%). In this way, we ensured that this split maintained the same proportional size for each class.
For each data augmentation technique, we generated 1000

augmented images for each class of Lymphoma from the original dataset. While the training set included both original and augmented images, the test set exclusively comprised original images. This approach ensured that during each iteration, the neural network was trained on varied datasets, as the training and test sets differed across iterations due to the random splitting of the data. This method allowed us to obtain better estimates of the average performance of each data augmentation technique.

AlexNet

CNNs achieved really interesting performance in 1998 on some small datasets for character recognition or low-resolution object recognition. In 2012 **AlexNet** was introduced by Alex Krizhevsky and Ilya Sutskever during the ImageNet Large Scale Visual Recognition Challenge, obtaining a top-5 error of 15.5%, 10.8% above the second classified.

AlexNet was much larger than previous CNNs used for Computer Vision tasks. It has 60 million parameters and 650,000 neurons. The architecture of AlexNet is shown below:

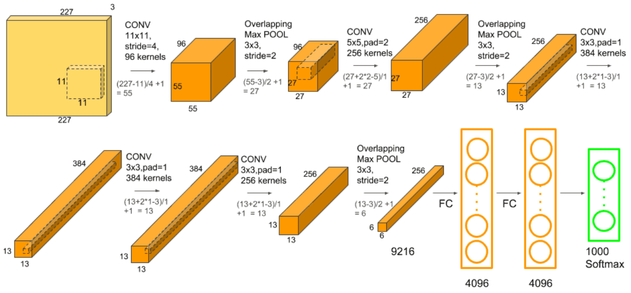

Figure 1: AlexNet Architecture.

AlexNet consists of 5 convolutional layers and 3 fully connected layers. Multiple Convolutional Kernels extract interesting features from the images. The first Convolutional Layer of AlexNet contains 96 kernels of size 11×11×3. The first two Convolutional Layers are followed by the Overlapping Max Pooling layers, the third, fourth and fifth Convolutional Layers are connected directly. The fifth Convolutional Layers is followed by another Overlapping Max Pooling layer, and its output goes into a series of two fully connected layers. The second fully connected layer feeds into a softmax classifier with 1000 class labels.

In our case we modified the last fully connected layer, due to the fact that we only have to classify 3 different classes (the 3 different categories of Lymphoma), and not 1000.

ReLU activation function has been applied after all the convolution and fully connected layers. The ReLU of the first and second convolution layers are followed by a local normalization step before doing pooling.

During the training of AlexNet, we used these parameters and starting hyperparameters:

- **Augmented Images**: 1000 per class
- **Batch size**: 8
- **Learning rate**: 0.001
- **Weight Decay**: 0.0001
- **Momentum**: 0.7

We used an Adaptive Learning Rate to dynamically adjust the learning rate during training based on the performance of the model. This technique allowed us to efficiently optimize the model's parameters while training our modified AlexNet architecture for classifying the three different categories of Lymphoma.

(Krizhevsky, Sutskever, and Hinton 2012) (Mallick 2018) (Alexnet 2012)

Results

After running the code over five iterations and training our model, we collected the results for each data augmentation technique. We computed the average test accuracy from the test sets and plotted the confusion matrix and classification report (including precision, recall, F1 score, support, and accuracy) for the final iteration.

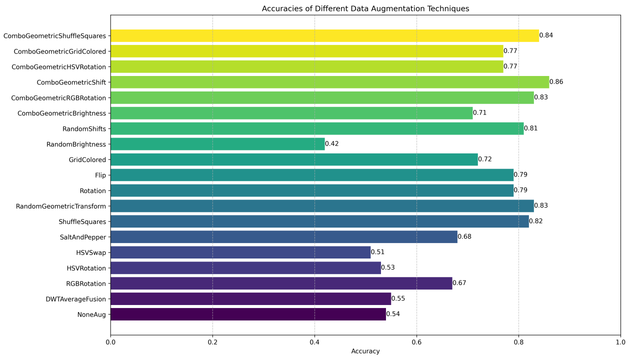

Figure 2: Data Augmentation Results.

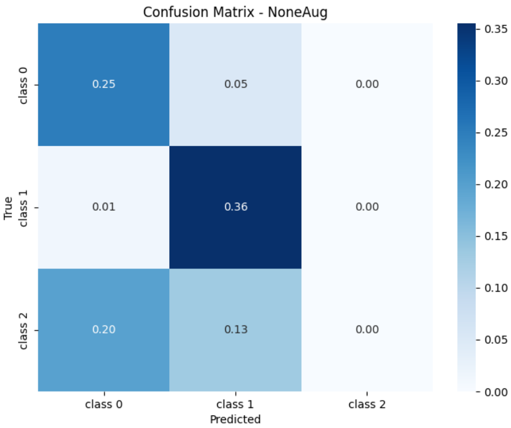

Figure 3: NoneAug Confusion Matrix.

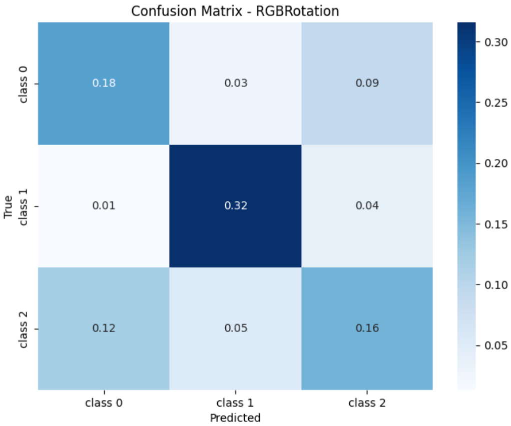

Figure 4: RGBRotation Confusion Matrix.

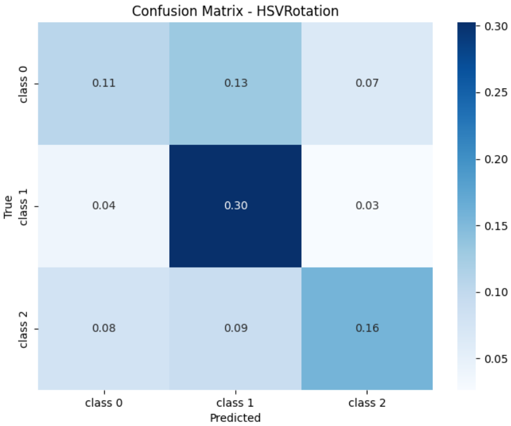

Figure 5: HSVRotation Confusion Matrix.

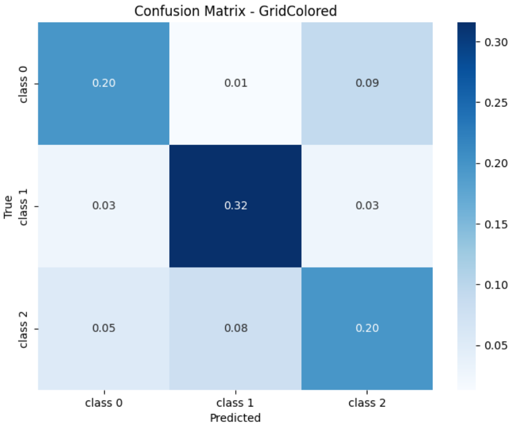

Figure 6: GridColored Confusion Matrix.

As seen in the earlier confusion matrices, AlexNet struggles to accurately classify images belonging to class 2. However, training AlexNet on the augmented dataset using techniques like *GridColored* has effectively addressed this issue, resulting in improved test accuracy.

From the test accuracy plot, it is evident that almost all the data augmentation techniques improved the performance of AlexNet. Notice that, our custom data augmentations yielded the following results:

- **NoAug:** Achieved 54% accuracy, however we can see from the confusion matrix (Figure 3) that class 2 has never been predicted.
- **RGBRotation:** Achieved 67% accuracy.
- **HSVRotation:** Achieved 53% accuracy, particularly improving the classification of class 2 compared to NoAug.
- **GridColored:** Achieved 72% accuracy.

However, the base data augmentation techniques that provided the best improvements were the geometric ones, such as **Rotation, Flip, RandomGeometricTransform, RandomShift**, and **ShuffleSquare**. This is because geometric transformations typically have a more substantial impact on enhancing model performance compared to other augmentation methods.

To leverage the advantages of geometric transformations, we opted to develop combined data augmentation techniques based on *Rotation* and *Flip*. The results indicate that all tested combinations outperformed the base augmentations. For example, **ComboGeometricRGBRotation** showed a 16% enhancement. Ultimately, the most effective data augmentation technique for this task was **ComboGeometricShuffleSquares**, which led to a 30% improvement over the original dataset.

Conclusions

In this project, we aimed to enhance the performance of neural network classifiers for Lymphoma image classification by employing various data augmentation techniques. Starting with a limited dataset of 375 malignant Lymphoma images categorized into three classes, we introduced new algorithms like RGBRotation, HSVRotation, and GridColored, with several combined geometric transformations.

Our experiments revealed that almost all data augmentation techniques improved AlexNet's performance. Specifically, RGBRotation and GridColored showed notable accuracy improvements. The geometric transformations, such as Rotation, Flip, and ShuffleSquares, consistently yielded the best results, underlining their efficacy in enhancing model performance.

By creating combined data augmentation techniques based on Rotation and Flip, we observed significant improvements. For instance, ComboGeometricRGBRotation improved accuracy by 16%, while ComboGeometricShuffleSquares achieved a remarkable 30% enhancement over the original dataset.

Overall, our data augmentation methods demonstrated substantial potential, highlighting the variability in their effectiveness depending on the task and dataset. Implementing data augmentation, especially in the realm of medical imaging, can greatly enhance the detection and classification of diseases such as Lymphoma, thereby boosting diagnostic precision and reliability.

References

Alexnet. 2012. https://pytorch.org/hub/pytorch_vision_alexnet/.

Davide Ferrari, Eddy Frighetto, Matteo Manuzzato. 2024. Lymphoma data augmentation project. https://github.com/CrocoPops/LymphomAug.

Dilmegani, C. 2024. Top data augmentation techniques. https://research.aimultiple.com/data-augmentation-techniques/.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, volume 25, 1097–1105.

Mallick, S. 2018. Understanding AlexNet. https://learnopencv.com/understanding-alexnet/.

Ozdemir, C.; Dogan, Y.; and Kaya, Y. 2024. RGB-Angle-Wheel: A new data augmentation method for deep learning models. *Knowledge-Based Systems*, 291: 111615.

What is Lymphoma? 2023. https://www.sciencedirect.com/science/article/pii/S0950705124002508.