Report on Final Project - Knee Osteoarthritis Classification

ICT Department - USTH
GROUP 1

Dao Xuan Bach
Nghiem Phu Khang
Le Duc Anh

***Abstract*—Knee osteoarthritis (KOA) is a common degenerative joint disease that causes pain and disability. Early and accurate classification of knee osteoarthritis is important for effective treatment. In this study, we propose a deep learning approach to classify knee X-ray images into five Kellgren-Lawrence (KL) grades using Convolutional Neural Networks (CNN). A baseline model was initially trained, achieving moderate performance in terms of accuracy, precision, and recall. The model was then fine-tuned in an attempt to improve performance. Although some improvements were observed, challenges such as class imbalance and suboptimal recall remain, indicating that further optimization and refinement are needed. This study demonstrates the potential of deep learning for KOA classification, while also highlighting the need for additional techniques, such as data augmentation or class balancing, to enhance the model's performance.**

I Introduction

Knee osteoarthritis (KOA) is a common degenerative joint disease, leading to pain and disability, especially in older adults. Accurate and early diagnosis is crucial for effective treatment. Traditionally, KOA severity is assessed using knee X-rays and the Kellgren-Lawrence (KL) grading system. However, manual interpretation of X-rays is time-consuming and prone to errors.

Recent advancements in deep learning, particularly Convolutional Neural Networks (CNN), have shown great potential in automating medical image analysis. This study aims to develop a deep learning model to classify knee X-ray images into five KL grades of KOA. We train a baseline CNN model and fine-tune it to improve classification accuracy, demonstrating the potential of CNNs for early KOA diagnosis.

II Knee Osteoarthritis Dataset

The **Knee Osteoarthritis (KOA)** [1] dataset contains X-ray images of knee joints, which are used to classify the severity of osteoarthritis based on the Kellgren-Lawrence (KL) grading scale, ranging from 0 to 4. This dataset is commonly used for developing machine learning and deep learning models to detect and classify knee osteoarthritis from X-ray images.

Number of images: The dataset includes over 5,700 knee X-ray images.

Classification labels:

- **KL 0:** No osteoarthritis.
- **KL 1:** Mild osteoarthritis.
- **KL 2:** Moderate osteoarthritis.
- **KL 3:** Severe osteoarthritis.
- **KL 4:** Very severe osteoarthritis.

This dataset is ideal for training image classification models to predict the severity of osteoarthritis, aiding in the diagnosis and treatment planning for knee arthritis.

III Methodology

This section outlines the methodology employed in our study, covering data preprocessing, model architecture, and training procedures.

A Dataset Preprocessing

In this study, we used the Knee Osteoarthritis dataset, which consists of knee X-ray images with various levels of osteoarthritis severity. The dataset is imbalanced, with a significant number of images belonging to the less severe categories. To address this imbalance, we decided not to use data augmentation, as it might introduce unrealistic transformations for medical images. Instead, we focused on resizing the images to a fixed input size of 299x299 pixels, which is compatible with the Xception model. This resizing ensures that all input images have the same dimension, making them suitable for model training.

B Model Selection and Architecture

For the classification task, we chose the Xception model, a deep convolutional neural network known for its efficiency in handling image classification problems. Xception uses depthwise separable convolutions, which reduce the computational cost while maintaining high performance, making it an ideal choice for medical image classification.

Initially, we used the pre-trained weights of Xception, which were trained on the ImageNet dataset. This helps leverage features learned from a large-scale dataset, enhancing the model's ability to generalize to new data. We then fine-tuned the model by unfreezing the last 20 layers to adapt the pre-trained features to the knee X-ray images in our dataset.

C Training Procedure

The baseline model was trained using categorical cross-entropy as the loss function, which is suitable for multi-class classification tasks. After training the baseline model, we performed fine-tuning by unlocking the last 20 layers and using the Focal Loss function to address the class imbalance issue. Focal Loss places more emphasis on hard-to-classify examples, allowing the model to focus on learning from the minority classes. This approach helps the model adjust its weights gradually, thereby improving its performance in classifying knee X-ray images.

The training process involved first training the model with the pre-trained layers frozen. Afterward, the last 20 layers were unfrozen, and fine-tuning was performed using the Focal Loss function. This approach allows the model to adjust its weights gradually and achieve better performance on the knee X-ray images.

D Model Evaluation

The model's performance was evaluated using standard classification metrics, including accuracy, precision, recall, and F1-score. These metrics provide a comprehensive assessment of the model's ability to classify knee X-ray images correctly across different osteoarthritis grades.

IV Experiment

A Baseline Model

The baseline model was trained for 5 epochs on the training dataset. The performance metrics were evaluated on both the training and validation datasets. The results of the training process for the baseline model are presented in the table below:

TABLE I
TRAINING RESULTS OF BASELINE MODEL (5 EPOCHS)

Epoch	Loss	Accuracy	Precision	Recall	AUC	Val Loss
1	1.2878	0.4586	0.5731	0.1826	0.7696	1.1919
2	1.1897	0.4929	0.6073	0.2484	0.8055	1.1721
3	1.1510	0.5116	0.6162	0.2837	0.8195	1.1572
4	1.1306	0.5246	0.6345	0.3065	0.8269	1.1448
5	1.1034	0.5277	0.6413	0.3228	0.8355	1.1310

V Conclusion

A Fine-Tuned Model

The model was fine-tuned by unlocking the last 20 layers of the pretrained model, with the training set augmented using focal loss to handle the class imbalance. The results of the fine-tuned model after 5 epochs are shown in the table below:

The fine-tuned model showed significant improvement in performance after 8 epochs of training, with increased accuracy, precision, recall, and AUC scores. However, the model still shows room for improvement, as the test results indicate moderate accuracy and recall. This suggests that further adjustments, such as hyperparameter tuning or the application of additional data augmentation techniques, could be explored to enhance the model's performance.

TABLE II
TRAINING RESULTS OF FINE-TUNED MODEL (5 EPOCHS)

Epoch	Loss	Accuracy	Precision	Recall	AUC	Val Loss
1	0.1666	0.4785	0.6378	0.1198	0.7957	0.2749
2	0.1269	0.5635	0.6928	0.2435	0.8574	0.1276
3	0.1049	0.6146	0.7391	0.3496	0.8903	0.1204
4	0.0871	0.6687	0.7807	0.4553	0.9163	0.1200
5	0.0749	0.6990	0.8145	0.5128	0.9324	0.1240
6	0.0644	0.7323	0.8262	0.5668	0.9450	0.1346
7	0.0552	0.7593	0.8453	0.6135	0.9564	0.1346
8	0.0463	0.7864	0.8576	0.6565	0.9654	0.1421

The evaluation on the test set with 1,656 images across 5 classes resulted in the following performance metrics:

TABLE III
TEST SET EVALUATION RESULTS

Metric	**Value**
Test Loss	0.1402
Test Accuracy	0.5501
Test Precision	0.6592
Test Recall	0.4076
Test AUC	0.8610

References

[1] *Knee Osteoarthritis Dataset with KL Grading - 2018*