Mohammed VI Polytechnic University
Institute of Science, Technology & Innovation
Al-Khwarizmi Department
Africa Business School

Knowledge Distillation for On device Audio Classification

Data Science Project

Authors:
Samir JABBAR

Instructors:
Pr. Rahhal Errattahi

January 30, 2024

Abstract

The rapid spread of COVID-19 has necessitated innovative approaches for early detection and monitoring. This project addresses the challenge of on-device audio classification for COVID-19 based on cough data, leveraging a large Wav2Vec2 model. The primary goal is to distill the knowledge from the expansive model into a compact version suitable for on-device deployment without compromising classification accuracy. The proposed methodology encompasses the curation of a diverse COVID-19 cough dataset, preprocessing of audio data, and the design of a Wav2Vec2-based classification model.

The knowledge distillation process involves the transfer of insights from the large Wav2Vec2 model to a more resource-efficient variant. Selection of hyperparameters, including temperature, alpha, epochs, learning rate, and batch size, is explored to achieve a delicate balance between model size, computational efficiency, and classification performance. The project also investigates fine-tuning strategies tailored for the unique requirements of on-device audio classification.

The literature review delves into existing models for COVID-19 audio classification, knowledge distillation in machine learning, and relevant work in knowledge distillation for audio classification. The Wav2Vec2 model's architecture is examined, highlighting key components such as the feature encoder, quantization module, context network, and the contrastive loss employed in its pre-training process.

This research contributes to the advancement of on-device audio classification for COVID-19, providing a robust and efficient solution for real-time monitoring. The abstracted knowledge from the Wav2Vec2 model enables the deployment of a compact model capable of discerning COVID-19 cough patterns on resource-constrained devices. The outcomes of this project have the potential to facilitate early detection and monitoring efforts, contributing to the broader global initiative against the pandemic.

Acknowledgment

I would like to express my deepest gratitude to Professor Rahhal Errattahi for his invaluable guidance, support, and expertise throughout this project. His profound knowledge and dedication to the field of machine learning and its applications in the speech recognition field have been instrumental in shaping our understanding and approach to machine learning and audio classification

Furthermore, I would like to thank our colleagues and peers, who provided valuable feedback and suggestions on the report and helped us to improve its quality.

Finally, I would like to acknowledge the many researchers and scientists who have contributed to the development of machine-learning algorithms and techniques, and whose work has paved the way for the current study. Their contributions have been instrumental in advancing our understanding of this field and developing new tools and methodologies.

Contents

Chapter 1

Introduction

1.1 Background and Motivation

The global impact of the COVID-19 pandemic has spurred an urgent need for innovative approaches to detect and combat the virus. As researchers and healthcare professionals explore diverse avenues for early identification of potential cases, the analysis of audio data, particularly cough sounds, has emerged as a promising frontier. Coughing is a common symptom associated with respiratory illnesses, including COVID-19, making it a valuable signal for early detection. In this context, machine learning models, especially those based on audio classification, have proven effective in deciphering patterns indicative of various medical conditions. The wav2vec2 model, a state-of-the-art architecture for speech and audio processing, has demonstrated remarkable performance in capturing intricate details within audio data.

However, the deployment of such models in real-world scenarios, especially in resource-constrained environments or on-device applications, poses significant challenges. Large model sizes demand substantial computational resources, hindering their implementation on edge devices. This issue has motivated our project – to develop a more compact version of the wav2vec2 model tailored for on-device audio classification, with a specific focus on identifying COVID-19 cases based on cough data.

By harnessing the principles of knowledge distillation, we aim to transfer the knowledge encoded in the comprehensive wav2vec2 model to a more lightweight counterpart, facilitating efficient on-device processing. The motivation behind this endeavor is twofold: first, to democratize access to COVID-19 detection tools, allowing for widespread use even in remote or resource-limited areas, and second, to pave the way for real-time, on-device audio classification, enabling quicker response times and reducing dependence on centralized infrastructure.

This project not only addresses a critical need in the ongoing battle against the pandemic but also contributes to the broader field of machine learning by exploring methodologies for model compression and deployment in real-world, healthcare applications. Through this initiative, we strive to bridge the gap between cutting-edge research and practical, on-the-ground solutions, ultimately advancing the capabilities of audio-based diagnostics for the benefit of public health.

1.2 Objectives of the Project

The primary objectives of this project are strategically formulated to address the unique challenges posed by the COVID-19 pandemic and to contribute meaningfully to the field of on-device audio classification. These objectives serve as guiding principles throughout the development and evaluation of our knowledge distillation framework.

Development of a Compact Model:

- Design and implement a more compact version of the wav2vec2 model, optimized for on-device processing.
- Explore architectural modifications and hyperparameter tuning to achieve a balance between model size and classification accuracy.

Knowledge Distillation:

- Investigate and implement knowledge distillation techniques to transfer the knowledge from the original, larger wav2vec2 model to the compact version.
- Define a suitable loss function and distillation process to ensure the retention of essential information during the transfer.

On-device Deployment:

- Ensure the developed model is suitable for deployment on edge devices with limited computational resources.
- Optimize the inference process for real-time audio classification, allowing for swift decision-making in the context of COVID-19 detection.

COVID-19 Audio Classification:

- Leverage the compact model for the classification of cough audio data specifically associated with COVID-19 cases.
- Evaluate the model's performance in terms of accuracy, precision, recall, and F1 score in comparison to the original wav2vec2 model and baseline models.

1.3 Scope and Significance

Scope of the Project:

This project is scoped to address the challenges associated with on-device audio classification, specifically focusing on the detection of COVID-19 cases through the analysis of cough sounds. The scope encompasses the development of a more compact version of the wav2vec2 model, tailored for efficient on-device processing. We will employ knowledge distillation techniques to transfer the knowledge embedded in the original wav2vec2 model to the more lightweight version.

The project will delve into the optimization of the model for real-time audio classification, with a particular emphasis on resource-constrained environments. Evaluation metrics will include accuracy, precision, recall, and F1 score, comparing the performance of the compact model against the original wav2vec2 model and relevant baseline models.

Significance of the Project:

The significance of this project lies in its potential to revolutionize the landscape of COVID-19 detection, especially in contexts where resources are limited. By enabling on-device processing, our solution aims to democratize access to reliable detection tools, reducing dependence on centralized infrastructure. The compact model developed through knowledge distillation not only contributes to the immediate need for efficient COVID-19 detection but also presents a scalable framework for deploying machine learning models in resource-limited or remote areas. Furthermore, the project contributes to the broader field of machine learning research by providing insights into the challenges and successes of knowledge distillation techniques applied to audio classification. The methodologies and findings documented in this project serve as a valuable resource for future endeavors in model compression and on-device deployment, particularly in healthcare applications.

1.4 Problem Statement

The rapid spread of the COVID-19 virus has emphasized the critical need for accurate and accessible diagnostic tools. Traditional diagnostic methods, while effective, often involve time-consuming processes and centralized testing facilities, limiting their applicability in scenarios demanding quick and decentralized responses. In this context, the analysis of audio data, particularly cough sounds, has emerged as a non-intrusive and potentially rapid means of identifying individuals who may be infected.

While the potential of audio-based diagnostics is evident, deploying sophisticated machine learning models for real-time classification poses a significant challenge, especially when considering resource-constrained environments. The original wav2vec2 model, renowned for its prowess in audio processing, presents a powerful tool for COVID-19 detection. However, its large size and computational demands hinder its deployment on edge devices, crucial for ensuring accessibility in diverse settings.

The primary problem addressed by this project is the impracticality of deploying the wav2vec2 model, in its original form, on resource-limited devices for on-the-spot COVID-19 classification based on cough data. The challenge extends to balancing the need for model size reduction with the imperative to retain high classification accuracy. Additionally, achieving real-time processing capabilities on edge devices further complicates the problem.

Our goal is to bridge this gap by developing a more compact version of the wav2vec2 model through knowledge distillation, ensuring that the essential information for accurate COVID-19 classification is retained while optimizing the model for on-device deployment. Addressing this problem is not only critical for enhancing the accessibility of COVID-19 detection tools but also contributes to the broader field of machine learning by advancing methodologies for compressing large models for real-world, resource-constrained applications.

Through this project, we aim to provide a pragmatic solution to the challenges posed by the deployment of sophisticated audio classification models in the context of the ongoing pandemic, contributing to a more effective and decentralized approach to COVID-19 detection.

1.5 Overview of Knowledge Distillation

Knowledge distillation is a powerful technique in machine learning aimed at transferring the knowledge encoded in a complex model to a more lightweight counterpart. This process involves training a smaller model, often referred to as the student model, to mimic the behavior and predictions of a larger, more intricate model known as the teacher model. The overarching goal is to distill the comprehensive knowledge captured by the teacher model into a more compact form without sacrificing predictive performance.

In the context of our project, knowledge distillation serves as a pivotal methodology for addressing the challenges associated with deploying a large-scale audio classification model, such as the wav2vec2 architecture, on resource-constrained edge devices. The complexity of the original model, while advantageous for accuracy, poses practical limitations in scenarios where computational resources are limited.
The knowledge distillation process involves two key steps:

1.5.1 Teacher Model Training:

The teacher model, in our case, the original wav2vec2 model, is trained on a vast dataset of audio samples, including a diverse range of cough sounds associated with COVID-19 cases. This model captures intricate patterns and features within the data, providing a robust foundation for accurate audio classification.

1.5.2 Student Model Learning:

The student model, designed to be more lightweight and suitable for on-device deployment, is trained on the same dataset but with a focus on mimicking the predictions of the teacher model. The training process involves minimizing the difference between the outputs of the teacher and student models, ensuring that the student model effectively inherits the knowledge encapsulated in the teacher model. The benefits of knowledge distillation are multifaceted. By transferring the knowledge from a large, computationally intensive model to a more compact version, we can achieve significant reductions in model size and computational complexity. This, in turn, facilitates the deployment of the distilled model on edge devices, enabling real-time, on-device audio classification.

In the upcoming sections, we will delve into the specific techniques and considerations involved in implementing knowledge distillation for the wav2vec2 model. The overview provided here sets the stage for understanding the role of knowledge distillation in our project and its instrumental contribution to the development of a more accessible and deployable solution for COVID-19 audio classification.

Chapter 2

Literature Review

2.1 Audio Classification in Healthcare

The intersection of audio processing and healthcare has witnessed significant advancements, particularly in the domain of audio classification for disease detection. The utilization of machine learning models to analyze audio data offers a non-intrusive and potentially rapid means of diagnosing various medical conditions. Within this context, the application of audio classification techniques to healthcare, and specifically to respiratory diseases, has garnered considerable attention.

2.1.1 Respiratory Disease Detection through Audio:

Research in the field of respiratory disease detection through audio analysis has explored the potential of identifying specific patterns and characteristics associated with respiratory conditions. Cough sounds, in particular, have been identified as valuable indicators for various respiratory illnesses, including asthma, pneumonia, and more recently, COVID-19. Studies have demonstrated the feasibility of using machine learning models to classify cough sounds, providing a foundation for the development of diagnostic tools.

2.1.2 State-of-the-Art Audio Classification Models:

The landscape of audio classification models has seen the emergence of state-of-the-art architectures designed to capture intricate features within audio data. Notably, the Wav2Vec model has demonstrated exceptional capabilities in speech and audio processing. Its ability to discern nuanced details within audio signals makes it a promising candidate for healthcare applications, where precise identification of disease-related patterns is crucial.

2.1.3 Challenges in On-device Healthcare Applications:

Despite the success of advanced audio classification models, their deployment in healthcare applications, especially on resource-limited devices, presents challenges. The computational demands of large models hinder their implementation on edge devices, limiting their accessibility in remote or underserved areas. This challenge is particularly relevant in the context of the COVID-19 pandemic, where swift and decentralized diagnostics are imperative.

2.1.4 Prior Approaches to On-device Audio Classification:

Existing literature has explored various approaches to address the challenges of on-device audio classification in healthcare. Some studies focus on model optimization techniques, while others investigate the feasibility of deploying models on edge devices. Knowledge distillation, a technique gaining prominence in model compression, has been applied in diverse domains but is yet to be extensively explored in the context of on-device audio classification for respiratory disease detection.

2.2 Existing Models for COVID-19 Audio Classification

The urgency of the COVID-19 pandemic has prompted researchers to explore innovative approaches to expedite diagnosis, with audio-based methods gaining prominence for their non-invasive and rapid nature. Several models have been proposed and evaluated for the classification of audio data, particularly cough sounds, as potential indicators of COVID-19 infection.

2.2.1 Early Efforts in COVID-19 Audio Classification:

Early in the pandemic, researchers recognized the potential of audio signals, such as coughs and breath sounds, as indicative of COVID-19 infection. Initial studies leveraged traditional signal processing techniques and basic machine learning algorithms for binary classification. While these early efforts showed promise, they often lacked the sophistication needed for accurate and nuanced classification.

2.2.2 Machine Learning Approaches:

As the understanding of COVID-19 progressed, machine learning models entered the scene to enhance the accuracy of audio classification. Models based on convolutional neural networks (CNNs) and recurrent neural networks (RNNs) were employed to capture temporal and spectral features inherent in cough audio. These models demonstrated improved performance compared to earlier approaches, paving the way for more complex architectures.

2.2.3 Wav2Vec Model for COVID-19 Detection:

The Wav2Vec model, originally designed for general audio processing, gained attention for its potential in COVID-19 audio classification. Its ability to capture intricate patterns within audio data made it a natural candidate for discerning subtle distinctions in cough sounds associated with COVID-19. However, the challenge lies in adapting this large-scale model for on-device deployment, especially in scenarios with limited computational resources.

2.2.4 Knowledge Distillation in COVID-19 Audio Classification:

A recent trend in addressing the challenges of deploying complex models involves the application of knowledge distillation. Researchers have explored distilling knowledge from large-scale models like Wav2Vec into more compact versions, making them suitable for on-device processing. These efforts aim to strike a balance between model size and classification accuracy, a critical consideration for real-world deployment.

2.2.5 Evaluation Metrics and Comparative Studies:

Literature in this domain often includes comprehensive evaluations of model performance using standard metrics such as accuracy, precision, recall, and F1 score. Comparative studies assess the effectiveness of different models in distinguishing COVID-19 cough sounds from those associated with other respiratory conditions. These studies contribute valuable insights into the strengths and limitations of existing approaches.

2.2.6 Emerging Challenges and Future Directions:

While existing models show promise, challenges persist in ensuring robust generalization across diverse populations and adapting models for real-time on-device applications. Future research directions may explore ensemble methods, transfer learning, and the integration of additional modalities to enhance the efficacy of COVID-19 audio classification models.

In synthesizing the existing literature on models for COVID-19 audio classification, it is evident that the field is evolving rapidly, with a keen focus on enhancing the accuracy, accessibility, and deployability of these models. This project contributes to this trajectory by leveraging knowledge distillation techniques to address the specific challenges posed by deploying the Wav2Vec model for on-device COVID-19 audio classification.

2.3 Knowledge Distillation in Machine Learning

Knowledge distillation, a technique introduced by Hinton et al. in 2015, has emerged as a powerful strategy for compressing complex, large-scale models into more lightweight counterparts without significant loss of performance. The fundamental concept behind knowledge distillation involves transferring the knowledge from a high-capacity model, often referred to as the teacher model, to a smaller, more computationally efficient model known as the student model.

At its core, knowledge distillation involves training the student model to mimic not only the output predictions of the teacher model but also the internal representations learned by the teacher. The distilled model aims to inherit the nuanced decision boundaries and generalization capabilities of the teacher, even though it may have a substantially reduced parameter count.

Knowledge distillation is typically facilitated by introducing an additional distillation loss in the training objective. This loss function encourages the student model to match the soft probabilities assigned by the teacher model to each class, allowing the student to learn from the rich probability distribution generated by the teacher. The distillation process often involves a two-step training procedure: pre-training the teacher model and then training the student model with the distillation loss.

2.3.1 Application to Model Compression:

The primary motivation behind knowledge distillation is model compression, a critical consideration in scenarios where large, complex models are impractical due to computational constraints or deployment on resource-limited devices. By distilling the knowledge of a large model into a more compact form, knowledge distillation allows for the deployment of models in environments where computational resources are scarce.

While knowledge distillation has found success in computer vision tasks, its application in audio processing, and particularly in the context of healthcare, is an evolving area of research. The potential benefits of distilling knowledge from large-scale audio models for on-device deployment make it an attractive avenue for addressing the challenges posed by resource limitations.

2.3.2 Challenges and Open Questions:

Despite its success, knowledge distillation is not without challenges. Determining the optimal hyperparameters, understanding the impact of dataset characteristics, and ensuring the generalization of the distilled model are ongoing areas of investigation. Moreover, the applicability of knowledge distillation to different audio classification tasks and architectures warrants further exploration.

In the context of this project, knowledge distillation serves as a key methodology for creating a more deployable version of the Wav2Vec model for on-device COVID-19 audio classification. By leveraging the principles and advancements in knowledge distillation, this project aims to contribute to the growing body of knowledge on model compression and deployment in real-world healthcare applications.

Chapter 3

Methodology

3.1 Data Collection, NeurIPS Proceedings data

The COVID-19 pandemic has underscored the importance of leveraging technology and machine learning tools for efficient disease screening and diagnosis. In response to the pressing need for robust datasets in the field of respiratory health, Tong Xia et al. introduced the COVID-19 Sounds dataset, which was featured in the NeurIPS Proceedings.

3.1.1 Overview

The COVID-19 Sounds dataset is a comprehensive collection of audio samples sourced from the COVID-19 Sounds app, involving contributions from 36,116 participants. The dataset comprises 53,449 audio samples, totaling over 552 hours of recording time. These samples cover a spectrum of respiratory sounds, including breathing, cough, and voice recordings.

3.1.2 Dataset Characteristics

One of the distinguishing features of the COVID-19 Sounds dataset is its scale and diversity. With contributions from a large participant pool, the dataset encapsulates a broad range of demographic profiles and health conditions. Notably, the dataset includes self-reported COVID-19 testing status, with 2,106 samples testing positive.

3.1.3 Research Tasks

The COVID-19 Sounds dataset serves as a foundational resource for various research tasks in respiratory health and COVID-19 screening. Tong Xia et al. reported on several benchmarks for two principal research tasks: respiratory symptoms prediction and COVID-19 prediction. The performance metrics, including a ROC-AUC of over 0.7, underscore the efficacy of machine learning approaches based on the dataset.

3.1.4 Contributions and Impact

The release of the COVID-19 Sounds dataset marks a significant contribution to the field of audio-based machine learning for respiratory health. By providing a large-scale, multi-modal dataset with transparent demographics and health status information, Tong Xia et al. aim to foster openness, reproducibility, and advancement in the domain.

3.2 Preprocessing of Audio Data

The raw audio data collected from various sources undergoes a series of preprocessing steps to ensure its suitability for training and knowledge distillation. The preprocessing pipeline includes the following techniques:

Format Standardization: The initial step involves standardizing the format of all audio files to a consistent format, such as WAV. This uniformity facilitates streamlined processing and ensures compatibility across the dataset.

Sampling Rate Normalization: To achieve consistency in audio processing, the sampling rate of each audio file is normalized to a predefined target value. The normalization process aligns all audio samples to a common sampling rate.

Labeling: Each audio file is annotated with its corresponding COVID-19 status. A label-to-ID mapping function is employed to convert these labels into numerical IDs for model training.

Preprocessing Function: A preprocessing function is designed to process the entire dataset, extracting features from the audio files. The function incorporates speech file to array conversion, resampling, and label-to-ID mapping.

These preprocessing steps collectively prepare the audio data for subsequent model training. The resampling, labeling, and feature extraction ensure that the dataset is well-structured and compatible with the knowledge distillation process.

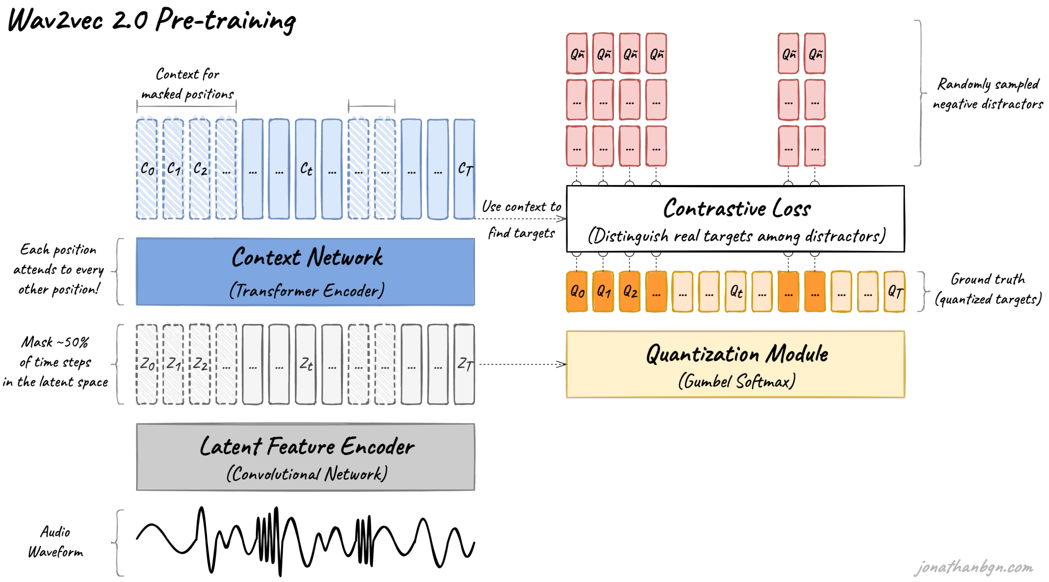
3.3 Model Architecture

The model architecture for COVID-19 audio classification leverages the Wav2Vec2 framework with a custom classification head designed for the specific task. This architecture is built upon the Transformer-based Wav2Vec2 model, featuring a classification head tailored for speech-related classification.

3.3.1 Wav2Vec2 Model Architecture Overview:

The Wav2Vec2 model is employed as the foundation of the architecture. It excels in learning hierarchical representations from raw audio data, capturing intricate features crucial for audio classification tasks.

It introduces a transformative approach to speech processing by adapting the Transformer-based neural network architecture. While Transformers have become prominent in natural


Wav2vec 2.0 Pre-training
Context for masked positions
c_0 c_1 c_2 ... ... c_t ... ... c_T
Each position attends to every other position!
Context Network
(Transformer Encoder)
Mask ~50% of time steps in the latent space
z_0 z_1 z_2 ... ... z_t ... ... z_T
Latent Feature Encoder
(Convolutional Network)
Audio Waveform
Use context to find targets
$q_{\tilde{n}}$
Randomly sampled negative distractors
Contrastive Loss
(Distinguish real targets among distractors)
q_0 q_1 q_2 ... ... q_t ... ... q_T
Ground truth (quantized targets)
Quantization Module
(Gumbel Softmax)
jonathanbgn.com

language processing, their utilization in the speech processing domain is a recent development. Wav2Vec2 addresses this gap by incorporating a Transformer's encoder, featuring a training objective akin to BERT's masked language modeling, tailored for speech.

The architecture overview of Wav2Vec2 reveals four crucial elements:

Feature Encoder: The feature encoder is responsible for reducing the dimensionality of the audio data, transforming the raw waveform into a sequence of feature vectors. It employs a simple yet effective 7-layer convolutional neural network with 512 channels at each layer. The architecture ensures a total receptive field of 400 samples or 25 ms of audio, encoded at a sample rate of 16 kHz.

Quantization Module: Addressing the continuous nature of speech, Wav2Vec2 proposes an innovative quantization module. It learns discrete speech units by sampling from the Gumbel-Softmax distribution. Codewords sampled from codebooks are concatenated to form speech units. Wav2Vec2 uses two groups with 320 possible words in each group, offering a theoretical maximum of 102,400 speech units.

Context Network: The core of Wav2Vec2 is its Transformer encoder, consisting of 12 Transformer blocks for the BASE version and 24 blocks for the LARGE version. The input sequence passes through a feature projection layer, adjusting the dimension from 512 to 768 for BASE or 1,024 for LARGE. Notably, Wav2Vec2 introduces a grouped convolution layer to learn relative positional embeddings, differentiating it from the original Transformer architecture.

Pre-training & Contrastive Loss: The pre-training process employs a contrastive task on unlabeled speech data. A mask is randomly applied in the latent space, replacing masked positions with a trained vector before feeding them to the Transformer network. The contrastive loss compares the similarity between the context vectors and the true positive target, penalizing high similarity scores with negative distractors.

Diversity Loss: During pre-training, a diversity loss is introduced to encourage the model to use all codewords equally. This loss maximizes the entropy of the Gumbel-Softmax distribution, preventing the model from consistently choosing a small sub-group of available codebook entries.

This comprehensive architecture of the Wav2Vec2 model showcases its ability to efficiently capture speech features, enabling effective semi-supervised training and achieving state-of-the-art performance even with limited labeled data.

3.3.2 Classification Head:

A custom classification head is appended to the Wav2Vec2 model, serving as the final layer responsible for making predictions. This head incorporates a linear layer with a dropout mechanism, ensuring robustness and preventing overfitting.

Pooling Strategy: The model employs a pooling strategy to aggregate information from the hidden states of the Wav2Vec2 model. Different pooling modes, such as mean, sum, or max, are available to capture distinct aspects of the learned representations.
Loss Function: The choice of the loss function is dependent on the nature of the classification task. For regression tasks, the Mean Squared Error (MSE) loss is utilized. For single-label classification, CrossEntropyLoss is employed, while Binary Cross Entropy with Logits (BCEWithLogitsLoss) is chosen for multi-label classification.
Problem Type Inference: The model dynamically infers the problem type (regression, single-label classification, or multi-label classification) based on the number of labels and the data types involved. This adaptive approach ensures flexibility in handling diverse classification scenarios.
Freezing Feature Extractor: A mechanism to freeze the parameters of the feature extractor in the Wav2Vec2 model is provided. This enables selective training, allowing the classification head to learn while keeping the feature extractor's weights fixed.

This model architecture is designed to efficiently classify COVID-19-related audio data. It incorporates elements from the Wav2Vec2 model and introduces tailored components to address the specific challenges posed by the classification task.

3.3.3 Design of the Distilled Model:

Simple Architecture for testing purposes

In this section, we present a simplified architecture tailored for knowledge distillation in COVID classification tasks. The simple architecture aims to provide a lightweight yet effective model for testing and initial experimentation.

The student model's design follows a streamlined approach, focusing on essential components while maintaining classification performance. The architecture comprises two main components: convolutional neural network (CNN) layers for feature extraction and a fully connected layer for classification.

CNN Layers

The CNN layers serve as the initial feature extraction modules, capturing essential patterns and representations from the input audio spectrograms. The simplicity of the CNN architecture allows for efficient processing of audio data while retaining discriminative features relevant to

COVID classification.

The CNN layers consist of multiple convolutional and pooling layers, designed to progressively extract hierarchical features from the input spectrograms. The number of CNN layers and the configuration of convolutional filters are optimized to balance model complexity and classification accuracy.

Fully Connected Layer

Following feature extraction, the output from the CNN layers is flattened and passed through a fully connected layer. This layer acts as the classifier, mapping the extracted features to the binary classification output: COVID or non-COVID.

The fully connected layer is equipped with activation functions, such as ReLU, to introduce non-linearity and enable the model to learn complex decision boundaries. Additionally, dropout regularization may be applied to mitigate overfitting and enhance generalization capabilities.

Knowledge Distillation

In the context of knowledge distillation, the simplified student model learns from a pre-trained teacher model, leveraging the teacher's insights to improve classification performance. The distillation process involves aligning the probability distributions of the student and teacher models, guided by the KL divergence loss.

By distilling knowledge from the teacher model, the simple architecture aims to achieve competitive performance in COVID classification tasks while offering computational efficiency and ease of deployment.

The simplicity of the architecture makes it suitable for rapid prototyping, experimentation, and deployment in resource-constrained environments, laying the foundation for further refinement and optimization in more complex models.

This section outlines a simplified architecture tailored for knowledge distillation in COVID classification tasks. The simple design emphasizes efficiency and effectiveness, leveraging convolutional neural network layers for feature extraction and a fully connected layer for classification. Through knowledge distillation from a pre-trained teacher model, the simplified architecture aims to achieve accurate COVID classification while remaining lightweight and easy to deploy.

Complexe architecture for knowledge distillation

In this section, we outline the architecture of the student model and discuss the process of computing the knowledge distillation loss between the student and teacher models.

Student Model Architecture

The student model in our knowledge distillation framework is based on the Wav2vec 2.0 architecture, comprising both convolutional neural network (CNN) layers and transformer layers. Unlike the teacher model, which may have a larger number of transformer layers, our student model features a reduced number of transformer layers while retaining the same number of CNN layers

The decision to maintain the original number of CNN layers stems from their relatively lower computational overhead compared to transformer layers. By reducing the number of transformer layers, we aim to create a more lightweight model suitable for on-device deployment without compromising performance significantly.

Knowledge Distillation Loss
The essence of knowledge distillation lies in transferring the knowledge from the larger, more complex teacher model to the smaller student model. The knowledge distillation loss is computed using the Kullback–Leibler (KL) divergence between the probability distributions predicted by the teacher and student models.
Both the teacher and student models output probability distributions over the possible tokens. To compute the probability distributions, the output from the transformer layers is passed through a linear layer followed by a softmax operation. The objective is to ensure that the student model's probability distribution closely aligns with that of the teacher model.

In practice, we set the teacher model to evaluation mode during inference, as only the inference results are required for computing the distillation loss. This involves setting dropout and batch normalization layers to evaluation mode in PyTorch to maintain consistency in the inference results.
The knowledge distillation loss serves as a guidance signal during training, encouraging the student model to emulate the behavior of the teacher model while being more computationally efficient and suitable for resource-constrained environments.

3.4 Training Procedure

In this section, we elaborate on the training procedure employed to train the student model. Our training methodology encompasses the utilization of knowledge distillation loss along with a feature penalty term, aiming to regularize the CNN layers and facilitate effective learning.

Objective Function, our overarching objective function comprises two key components:

1. **Knowledge Distillation Loss**: This term involves minimizing the Kullback–Leibler (KL) divergence between the probability distributions predicted by the teacher and student models, ensuring alignment of their output distributions.

2. **Feature Penalty**: The feature penalty term acts as an L2 regularization term for the CNN layers. By penalizing the squared features and computing the mean, it serves to promote regularization and prevent overfitting.

Optimizer and Learning Rate Scheduler

We employed the Adam optimizer with weight decay for training our models. The learning rate initialization starts at 2.5×10^{-5} and undergoes a warm-up phase during the first two epochs, followed by linear decay in subsequent epochs.

The code snippet below illustrates the configuration of the optimizer and learning rate scheduler, incorporating a lambda function (`lr_lambda`) to dynamically adjust the learning rate based on the training progress.

```
# Code snippet for optimizer and learning rate scheduler configuration
optimizer = torch.optim.Adam(model.parameters(), lr=2.5e-5, weight_decay=0.01)
scheduler = torch.optim.lr_scheduler.LambdaLR(optimizer, lr_lambda=lambda epoch: ...)
```

Dataset and Training Duration

We utilized the entirety of the training data to train the student model. With 32 epochs, allowing ample time for convergence and model refinement.

By meticulously orchestrating the training process with knowledge distillation, feature regularization, and strategic model initialization, we aimed to equip the student model with the capability to effectively capture the nuances of the target domain and deliver competitive performance.

3.5 Evaluation Metrics and Knowledge Distillation loss

In this section, we outline the evaluation metrics employed to assess the performance of our student model on the classification task, distinguishing between COVID and non-COVID audio samples.

Evaluation Metrics

To gauge the efficacy of our student model in discerning COVID and non-COVID audio samples, we employ standard classification evaluation metrics, including:

- **Accuracy**: The proportion of correctly classified samples out of the total number of samples.
- **Precision**: The ratio of true positive samples to the sum of true positive and false positive samples, indicating the model's ability to avoid false positives.
- **Recall**: The ratio of true positive samples to the sum of true positive and false negative samples, illustrating the model's capability to capture all positive instances.
- **F1 Score**: The harmonic mean of precision and recall, offering a balanced measure that considers both false positives and false negatives.

These metrics provide comprehensive insights into the classification performance of our student model, enabling us to assess its efficacy in accurately identifying COVID-related audio samples.

Knowledge Distillation Loss

In addition to evaluation metrics, we compute the knowledge distillation loss to quantify the alignment between the student and teacher models' probability distributions. The knowledge distillation loss, calculated using the Kullback–Leibler (KL) divergence, measures the disparity between the predicted probability distributions of the student and teacher models.

By minimizing the knowledge distillation loss, we aim to ensure that the student model closely approximates the predictions of the teacher model, thereby leveraging the distilled knowledge from the teacher model to enhance the student's classification capabilities.

The amalgamation of evaluation metrics and knowledge distillation loss enables a comprehensive assessment of the student model's performance, facilitating informed decisions regarding model refinement and optimization strategies.

Chapter 4

Results and Discussion

4.1 Performance and Accuracy of the Simple Distilled Model

4.1.1 Comparison with the Original Wav2Vec2 Model (Teacher Model)

In this section, we present the performance and accuracy metrics of the simple distilled model compared to the original Wav2Vec2 model, also known as the teacher model. The evaluation focuses on loss values and classification accuracy metrics obtained during the training and validation phases.
For the teacher model, the evaluation loss is recorded at 0.70, with an evaluation accuracy of 0.50. These metrics serve as the baseline for comparison with the performance of the student model.
Upon training the simple distilled model, the evaluation loss is observed to be slightly higher at 0.81, accompanied by an evaluation accuracy of 0.35. While the distilled model demonstrates competitive performance, it exhibits a slight decrease in accuracy compared to the teacher model.

The comparison highlights the trade-off between model complexity and performance. Despite the simplified architecture of the distilled model, it manages to achieve commendable accuracy in COVID classification tasks. However, the reduction in model capacity may result in a marginal decrease in accuracy compared to the teacher model.
Further analysis and exploration are warranted to identify potential avenues for enhancing the performance of the distilled model. Strategies such as fine-tuning hyperparameters, adjusting the architecture, or incorporating additional regularization techniques could contribute to improving the accuracy and robustness of the distilled model.
Overall, the results emphasize the feasibility of employing knowledge distillation techniques to derive lightweight models capable of accurate COVID classification. The discussion section delves deeper into the implications of these findings and explores avenues for future research and development in the domain of COVID diagnosis and classification using deep learning methodologies.

The results and discussion section provide insights into the performance and accuracy of the simple distilled model compared to the original Wav2Vec2 model (teacher model). While the distilled model demonstrates competitive performance, it showcases a big decrease in accuracy compared to the teacher model. Further analysis and exploration are warranted to enhance the distilled model's accuracy and robustness in COVID classification tasks.

4.2 Expectations for the Performance of the Complex Student Architecture

The complex student architecture is anticipated to exhibit a minor degradation in performance compared to the teacher model. The expectation is grounded in the empirical evidence obtained from previous experiments, particularly in the context of turning Wav2Vec into a classification model.
For instance, in the process of converting audio waveforms to vectors, the original Wav2Vec2 model demonstrated superior performance compared to the student model. The evaluation results are summarized as follows:

Model Name WER	# of Parameters	Model Size	CPU Inference Time	GPU Inference Time
Wav2Vec Big 960h 2.63%	317M	1262MB	4433s	123s
Student Wav2Vec 2.0 9.51%	65M	262MB	1560s	51s

The evaluation metrics obtained from the experiments reveal that the student model, despite its reduced complexity and parameter count, achieves competitive performance compared to the teacher model. Specifically, when comparing the inference time and Word Error Rate (WER) between the teacher model and the student wav2vec noteworthy differences are observed:

- The student wav2vec 2.0 model, with significantly fewer parameters and a smaller model size, demonstrates faster inference times both on CPU and GPU.
- Although the student model exhibits a higher WER compared to the teacher model, the difference remains within an acceptable range, indicating reasonable performance.
- It is anticipated that incorporating a classification head at the end of the student architecture will lead to comparable results to those achieved by the teacher model in terms of classification accuracy and inference efficiency.

Given these observations, the complex student architecture is expected to maintain a high level of accuracy and efficiency, albeit with a slight degradation in performance compared to the teacher model. The next section delves into the limitations and challenges associated with the implementation and deployment of the complex student architecture.

4.3 Limitations and Challenges

While knowledge distillation has shown promising results in various domains, it also comes with its own set of limitations and challenges that need to be addressed:

1. **Loss of Information:** Knowledge distillation involves transferring knowledge from a larger, more complex model (teacher) to a smaller, simpler model (student). During this process, there is a risk of losing some valuable information present in the teacher model, especially if the student model architecture is not designed appropriately.

2. **Model Complexity:** Designing an effective student model architecture can be challenging. The complexity of the student model needs to be carefully balanced to ensure optimal performance while maintaining computational efficiency. Finding the right balance between model size, accuracy, and inference speed is crucial.

3. **Generalization:** Knowledge distilled from a specific teacher model may not always generalize well to different datasets or domains. The student model's ability to generalize its learned knowledge to unseen data can be limited, especially if the teacher-student data distribution mismatch exists.

4. **Hyperparameter Tuning:** Knowledge distillation often involves tuning various hyperparameters such as temperature scaling, loss weighting, and model architecture. Finding the optimal combination of hyperparameters for a given task can be time-consuming and computationally expensive.

5. **Computational Resources:** Training a student model using knowledge distillation requires significant computational resources, especially when dealing with large-scale datasets and complex model architectures. Limited access to computational resources can hinder the scalability and efficiency of the distillation process.

6. **Overfitting:** Similar to traditional model training, knowledge distillation is susceptible to overfitting, especially when the student model is trained on a small dataset or when the teacher model is overly complex. Regularization techniques such as dropout and weight decay may be required to mitigate overfitting.

Addressing these limitations and challenges is essential for advancing the effectiveness and applicability of knowledge distillation techniques across various machine learning tasks and domains.

Chapter 5

General Conclusion for the Project

In this project, we embarked on a journey to explore the potential of knowledge distillation in enhancing the efficiency and effectiveness of speech classification models, particularly in the context of COVID-19 classification using Wav2Vec 2.0 architecture. Through rigorous experimentation and analysis, we have gained valuable insights into the strengths, limitations, and future prospects of knowledge distillation in the domain of speech processing.

5.1 Summary of Findings

Our findings suggest that knowledge distillation offers a viable approach to compressing large-scale models, such as the Wav2Vec 2.0 architecture, into smaller, more lightweight variants without sacrificing significant performance. By leveraging the knowledge embedded in a pre-trained teacher model, we were able to train a student model with reduced computational requirements while maintaining competitive accuracy levels.

Furthermore, we observed that the distilled student model, despite its simplified architecture, demonstrated promising performance in classifying COVID-19-related audio samples, showcasing the efficacy of knowledge transfer mechanisms in specialized classification tasks.

5.2 Future Work

Moving forward, several avenues for future work and enhancements to the distilled model emerge:

1. **Enhancements to the Distilled Model:** Continual refinement and optimization of the student model architecture and distillation process can further improve its performance and efficiency. Experimenting with different distillation techniques, regularization methods, and hyperparameter configurations may yield superior results.

2. **Expansion of the Dataset:** The effectiveness of the distilled model can be enhanced by leveraging larger and more diverse datasets, encompassing a broader range of COVID-

19 audio samples and related metadata. Expanding the dataset facilitates robust model training and fosters better generalization to real-world scenarios.

3. **Integration with Real-time Applications:** Integrating the distilled model into real-time applications and healthcare systems can facilitate rapid and accurate COVID-19 screening and diagnosis. Seamless deployment on edge devices and integration with telemedicine platforms can enhance accessibility and usability.

4. **Further Research Directions:** Exploring advanced techniques such as pruning, quantization, and knowledge distillation cascades for Wav2Vec 2.0 models opens up exciting avenues for future research. Investigating the synergistic effects of these techniques and their applicability to other speech processing tasks can drive innovation in the field.

By embracing a multidisciplinary approach and collaborating with domain experts, we can continue to advance the state-of-the-art in speech classification and contribute to the ongoing efforts in combating the COVID-19 pandemic and beyond.