applied sciences

MDPI

Article

Enhanced Speech Emotion Recognition Using Conditional-DCGAN-Based Data Augmentation

Kyung-Min Roh [1] and Seok-Pil Lee [2,*]

[1] Department of Computer Science, Graduate School, Sangmyung University, Seoul 03016, Republic of Korea; kungminno@naver.com
[2] Department of Intelligent IoT, Sangmyung University, Seoul 03016, Republic of Korea
* Correspondence: esprit@smu.ac.kr

Abstract: With the advancement of Artificial Intelligence (AI) and the Internet of Things (IoT), research in the field of emotion detection and recognition has been actively conducted worldwide in modern society. Among this research, speech emotion recognition has gained increasing importance in various areas of application such as personalized services, enhanced security, and the medical field. However, subjective emotional expressions in voice data can be perceived differently by individuals, and issues such as data imbalance and limited datasets fail to provide the diverse situations necessary for model training, thus limiting performance. To overcome these challenges, this paper proposes a novel data augmentation technique using Conditional-DCGAN, which combines CGAN and DCGAN. This study analyzes the temporal signal changes using Mel-spectrograms extracted from the Emo-DB dataset and applies a loss function calculation method borrowed from reinforcement learning to generate data that accurately reflects emotional characteristics. To validate the proposed method, experiments were conducted using a model combining CNN and Bi-LSTM. The results, including augmented data, achieved significant performance improvements, reaching WA 91.46% and UAR 91.61%, compared to using only the original data (WA 79.31%, UAR 78.16%). These results outperform similar previous studies, such as those reporting WA 84.49% and UAR 83.33%, demonstrating the positive effects of the proposed data augmentation technique. This study presents a new data augmentation method that enables effective learning even in situations with limited data, offering a progressive direction for research in speech emotion recognition.

Keywords: speech emotion recognition; data augmentation; CGAN; DCGAN; artificial intelligence; deep learning

check for updates

Citation: Roh, K.-M.; Lee, S.-P. Enhanced Speech Emotion Recognition Using Conditional-DCGAN-Based Data Augmentation. *Appl. Sci.* **2024**, *14*, 9890. https://doi.org/10.3390/app14219890

Academic Editors: Luigi Bibbò and J. Artur Serrano

Received: 27 September 2024
Revised: 22 October 2024
Accepted: 27 October 2024
Published: 29 October 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The advancement and integration of Internet of Things (IoT) and Artificial Intelligence (AI) technologies have increasingly highlighted the importance of speech recognition technology in modern society. In particular, emotion speech recognition is gaining importance in various fields of application, including personalized services, enhanced security, and the medical field. Research on emotion recognition is actively being conducted and developed alongside the growth of the global market.

Early research on emotion speech recognition primarily focused on the Hidden Markov Model (HMM) [1–4]. HMM was effectively used for recognizing patterns in sequence data by assuming hidden states and inferring them. More recently, with the advancement of deep learning, studies utilizing deep neural networks [5–11] have become mainstream. Furthermore, research employing graph-based theories to leverage the structural and statistical relationships between speech data has also emerged [12]. Graph-based methods can also extract key structural features from speech data, providing simpler and faster performance than traditional deep learning methods. These studies help in learning the complex patterns of speech data, allowing for more accurate differentiation of various

subtle emotional states. However, building speech emotion recognition systems still faces many challenges due to the diversity and complexity of speech data. Emotions can vary not only based on acoustic characteristics but also depending on the context, and in certain scenarios such as business conversations, cultural and linguistic differences can significantly influence emotional expression. Therefore, it is important for speech emotion recognition research to use data that minimize these differences and focuses on the auditory characteristics of emotions. This research mainly uses dialog data where actors focus on emotional expressions, which is effective for learning emotional patterns. Commonly used datasets include Emo-DB [13], RAVDESS [14], and IEMOCAP [15], though these datasets are significantly smaller in scale compared to large datasets used in other fields such as image research [16,17].

In this study, we propose a novel data augmentation technique using Conditional-DCGAN (Deep Convolutional Generative Adversarial Nets), which combines CGAN (Conditional Generative Adversarial Nets) [18] and DCGAN [19], to address these issues and improve the performance of emotion recognition models. In this study, the Mel-spectrogram, which analyzes the characteristics of the voice signal and visually represents the change in frequency of the signal over time, was used as an image. Additionally, instead of the traditional loss function calculation methods used by GAN (Generative Adversarial Nets) [20], we implemented a new loss function calculation method borrowed from reinforcement learning, enabling a more precise model for emotion analysis. Reinforcement learning [21] is a field of machine learning in which an agent learns optimal actions through trial and error. The agent interacts with its environment and receives rewards based on the outcomes of its actions. Using these rewards, the agent develops an optimal strategy for action. In this study, we applied the concept of reinforcement learning to explore a new method for improving the performance of adversarial generative models. Specifically, we designed a reward mechanism using an emotion recognition model to evaluate the generated images, and used this to adjust the training direction of the generative model.

This paper aims to demonstrate the impact of the augmented data on the performance improvement of emotion speech recognition models. For this purpose, we used a CNN + Bi-LSTM model capable of effectively analyzing the unique characteristics of Mel-spectrogram data as an emotion speech recognition model. We used WA (weighted accuracy) and UAR (weighted average recall) as performance metrics.

The structure of this paper is as follows. Section 2 introduces widely used emotion speech databases in emotion speech recognition research and explains studies on emotion recognition using AI algorithms. It also covers Convolutional Neural Networks and Bidirectional Long Short-Term Memory, which are frequently used in emotion speech recognition. Section 3 describes the methodology proposed in this study, including how data were preprocessed, what features were extracted, the structure of the models used, and how they were trained. Section 4 explains the experimental methods and results, discussing the impact of the proposed method on improving the performance of emotion speech recognition models. Finally, Section 5 concludes with a discussion of the significance of this study and future research directions.

2. Related Works

Since emotions can be subjectively judged depending on the individual, the data used in research on speech emotion recognition significantly impact the results. Emotional speech data support various languages and emotions, and reliable data play a crucial role in enhancing the credibility of research. The following are some widely used databases in speech emotion recognition research.

2.1. Emotional Speech Database

2.1.1. Emo-DB

Emo-DB [13] is a German emotional speech database created by the Technical University of Berlin. It includes recordings from a total of 10 actors (five males and five females)

who performed in German, resulting in a total of 535 utterances. Emo-DB consists of seven emotional categories: anger, boredom, fear, happiness, sadness, disgust, and neutral.

2.1.2. RAVDESS

RAVDESS [14], developed by Ryerson University in Canada, is a multimodal database containing emotional speech and video. It includes recordings from 24 professional actors (12 males and 12 females) who expressed emotions through speech or song. Each actor's utterances expressing emotions are recorded in approximately 3-s clips, with a total of 1440 files. RAVDESS consists of eight emotional categories: neutral, calm, happy, sad, angry, fearful, disgust, and surprised. Each expression is represented with two levels of emotional intensity (normal, strong).

2.1.3. IEMOCAP

IEMOCAP [15] is a multimodal database developed by the Signal Analysis and Interpretation Laboratory (SAIL) at the University of Southern California. It includes audio, video, motion capture data, and text. A total of 10 actors (five males and five females) provided detailed motion capture information in dialogue sessions expressing emotions, with markers attached to their faces, heads, and hands. The dataset includes approximately 12 h of dyadic interaction data, featuring various emotions and scenarios in both scripted and improvised performances. The primary emotions provided are happiness, anger, sadness, frustration, and neutral, covering a total of five emotional categories.

2.2. Research on Speech Emotion Recognition

2.2.1. Generative Adversarial Nets

GAN (Generative Adversarial Nets) [20] is a framework where two neural networks, a generator and a discriminator, compete against each other while learning simultaneously. GANs are used to generate various types of data, especially images, music, and text, and several variants have been proposed to adapt them to different fields. For example, CGAN (Conditional GAN) [18] generates data based on specific conditions, while DCGAN (Deep Convolutional GAN) [19] enhances the generation quality by applying convolutional layers, preserving the spatial information of high-dimensional data like images.

GANs are also useful as an oversampling technique to address imbalanced data problems; in particular, GAN oversampling methods based on Markov Random Field (vMRF) have been shown to outperform traditional techniques such as SMOTE (Synthetic Minority Over-sampling Technique), which synthesizes samples from a small number of classes [22]. GANs have also been actively applied in speech emotion recognition research. BAEK et al. [23] applied DCGAN to augment emotion data and combined it with existing data to improve performance. The datasets used were RAVDESS [14] and Emo-DB [13], with Mel-spectrograms used as features. As a result, RAVDESS showed an improvement in WA from 64.8% to 72.3% and UAR from 64.2% to 72.3%, while Emo-DB showed an improvement in WA from 80.6% to 90.4% and UAR from 82.6% to 91.3%. Additionally, Shilandari et al. [24] addressed class imbalance in emotional data by applying Cycle-GAN (Cycle-Consistent Adversarial Networks) [25] and compared it with other techniques like adding random noise, SMOTE [26], and 3D CRNNs feature generation methods [27]. By selecting five emotion categories from the Emo-DB dataset, they achieved WA of 84.49% and UAR of 83.33%.

In many existing studies, separate GANs are applied to each emotion category, which imposes the burden of creating new models for each emotion. Therefore, this study proposes a method to improve speech emotion recognition performance by using Conditional-DCGAN, which combines CGAN and DCGAN to augment emotion data. CGAN enables the generation of data based on specific conditions, while DCGAN preserves the spatial information of high-dimensional data, making it suitable for generating emotion data in a flexible and scalable manner. A new loss function is also introduced to reflect emotion-specific characteristics, and its impact on emotion recognition performance is evaluated.

2.2.2. Convolutional Neural Network

Convolutional neural networks (CNNs) are models that can be trained while retaining the spatial information of an image and are used in various fields such as image classification, object detection, and image processing, and are also widely used in the field of emotion recognition, as shown in the work of Chauhan et al. [28]. They proposed a CNN architecture based on Log-Mel-spectrograms. The datasets used for the experiment were Emo-DB [13] and IEMOCAP [15], and the experiments were conducted in two settings: speaker-dependent and speaker-independent. In the speaker-dependent experiment, 80% of the data were used for training and 20% for testing, while in the speaker-independent experiment, data from eight speakers were used for training and two for testing. Additionally, data augmentation techniques such as speed, pitch, and shift adjustments were applied to Emo-DB to increase the data volume. The results showed that, in the speaker-dependent experiment, Emo-DB achieved WA of 84.61% and UAR of 82.27%, while IEMOCAP achieved WA of 58.99% and UAR of 59.32%. In the speaker-independent experiment, Emo-DB achieved WA of 72.02% and UAR of 75.35%, while IEMOCAP recorded WA of 59.33% and UAR of 55.22%.

2.2.3. Bidirectional Long Short-Term Memory

Bi-LSTM (Bidirectional Long Short-Term Memory) is an extension of LSTM, where two separate LSTMs process the sequence in both forward and backward directions, utilizing both past and future information to learn the bidirectional context of sequential data. This approach is particularly useful in applications where bidirectional information is important, such as natural language processing (NLP), speech recognition, and time-series prediction. Bi-LSTM has also been widely used in speech emotion recognition, as shown by the experiments conducted by Zhang et al. [29] and Andayani et al. [30].

Zhang et al. [29] extracted 72-dimensional features, including chroma and MFCC, from the Emo-DB [13], SAVEE [31], and CASIA [32] datasets and used two parallel Bi-LSTM models. The results showed that Emo-DB achieved WA of 86.44% and UAR of 84.65%, SAVEE achieved WA of 62.49% and UAR of 56.50%, and CASIA achieved WA of 58.25% and UAR of 79.67%. Additionally, Andayani et al. [30] proposed an LSTM-Transformer model using MFCC as input and tested it on the RAVDESS [14] and Emo-DB datasets. The model achieved WA of 77.33% and UAR of 75.62% for RAVDESS and WA of 87.72% and UAR of 85.55% for Emo-DB.

3. Proposed Methodology

3.1. Data Preprocessing

3.1.1. Data Cleaning

This study is based on the emotional speech data from Emo-DB. The research was conducted on six emotions, excluding the disgust category. The number of voice data samples for each emotional state is summarized in Table 1.

Table 1. Number of samples per emotion in the Emo-DB database.

Emotion	Emo-DB
Anger	127
Boredom	81
Fear	69
Happiness	71
Sadness	62
Neutral	79
Total	489

Voice data often contain various noises or irregular silences at the beginning and end that are unnecessary for the experiment. These segments can hinder accurate data analysis;

hence, they must be removed through preprocessing. In this process, we used the 'trim' function from the librosa library [33]. This function effectively eliminates background noise and silent sections by detecting the significant start and end points in the voice data. This step is crucial as it extracts only the valid signal, excluding unnecessary parts, thereby increasing the accuracy of the analysis.

3.1.2. Mel-Spectrogram Extraction

To build an emotion speech recognition model, the key features extracted from voice data include the fundamental acoustic characteristics and variability of the speech. In this study, we utilized Mel-spectrograms among these features. A Mel-spectrogram is an effective tool for analyzing the characteristics of specific sounds, as it visually represents the changes in signal frequency over time.

Spectrograms are generated based on the Short Time Fourier Transform (STFT), and Mel-spectrograms are the result of applying the Mel-scale to these spectrograms to better reflect human hearing characteristics. The Mel-scale structures human hearing characteristics so that below 1 kHz the center frequency is evenly divided and above 1 kHz it follows a logarithm scale [34]. Equation (1) is the Mel-scale conversion function for frequency.

$$Mel(f) = 2595 * log_{10}\left(1 + \frac{f}{100}\right) \quad (1)$$

A Mel-filter bank is a set of filters designed based on the Mel-scale, which transforms the frequency spectrum of speech to match the Mel-scale and extracts the energy from each frequency band. These filters divide the frequency spectrum into multiple bands and calculate the energy in each band, which is then used to generate the Mel-spectrogram. The filters are narrower at lower frequencies and wider at higher frequencies, reflecting the characteristic of the Mel-scale that humans are more sensitive to lower frequencies and less sensitive to higher frequencies. The Mel-spectrogram accurately represents these auditory characteristics.

A Mel-spectrogram can be considered a three-dimensional image that visually represents the changes in frequency and amplitude over time by adding a time axis. Therefore, this study proposes a method for data augmentation by utilizing the Mel-spectrograms extracted from speech data as images. To achieve this, the cleaned data, which was preprocessed using the trim function, was converted into a Mel-spectrogram using the 'melspectrogram' function from the librosa library [33]. During this process, the power spectrum was converted into decibel units to compress the dynamic range of the data. This conversion is done using the decibel scale, and the formula is presented in Equation (2). *power* represents the original spectrum value, and ref is the maximum amplitude of the Mel-spectrogram. This transformation reduces the difference between very large and small values, ensuring that the intensity of the data is handled uniformly and that the various intensities of the speech signals are standardized. The data normalization process adjusts all input features to the same scale, setting the data values within the range [0, 1] based on the maximum and minimum values of each data point.

$$S_{db} = 10 * \log_{10}\left(\frac{power}{ref}\right) \quad (2)$$

After going through these preprocessing steps, the data converted into Mel-spectrogram form is used for experiments. Figure 1 shows an example of a Mel-spectrogram.

3.2. Proposed Model

In this study, we propose a new data augmentation technique that combines the traditional Conditional-DCGAN with an emotional speech recognition model. In the traditional Conditional-DCGAN method, a discriminator evaluates the authenticity of the Mel-spectrogram generated by the generator, which is in the range [0, 1]. This approach

requires time at the beginning of training until the discriminator is sufficiently trained, and effective training of the discriminator may be limited, especially with a small amount of data. In this study, we further introduce an emotional speech recognition model to evaluate how the generated Mel-spectrogram expresses emotion and implement a new loss optimization method by reflecting it in the loss calculation.

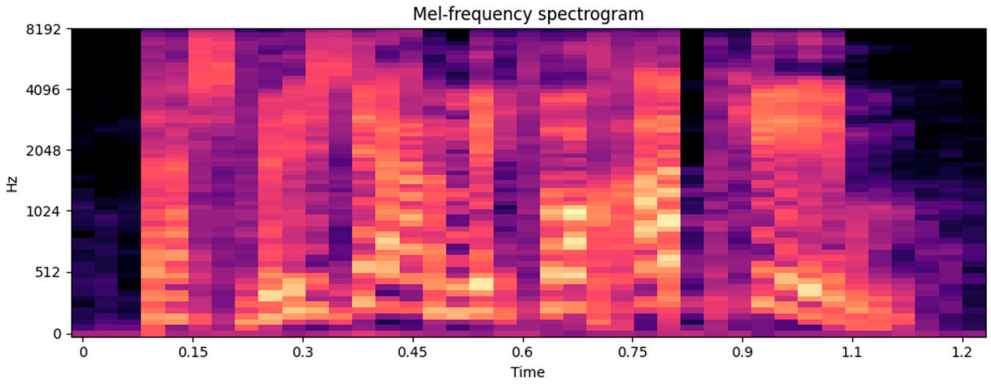


Figure 1. Mel-spectrogram.

The emotional speech recognition model is pre-trained with an accuracy of 0.86, and the Mel-spectrogram generated by the model uses a loss adjustment method where the better the emotion is expressed, the better the reward. While this process borrows the reward concept from reinforcement learning, it is somewhat different from the traditional policy gradient [35] method of reinforcement learning. In the policy gradient method, the agent learns the probability of choosing an action in such a way to maximize the reward. In this study, we learn by minimizing the loss value by evaluating how well the generated Mel-spectrogram expresses emotions. Table 2 summarizes how we applied the concept of reinforcement learning in this study.

Table 2. Application of Reinforcement Learning Concepts in This Study.

Concept	Application in the Proposed Method
Agent	Generator
Environment	Pre-trained emotion speech recognition model
Action	The process of generating Mel-spectrograms by the generator
State	The information currently held by the generator (Loss value)
Reward	The probability value assigned by the emotion recognition model to the generated Mel-spectrogram
Policy	The generator's network that determines how to generate Mel-spectrogram

In this study, the probability values obtained from the emotion recognition model are used as rewards to optimize the generator's network. The more accurately the Mel-spectrogram expresses emotions, the higher the probability values, and these values are used to adjust the loss so that the generator is guided to produce data that better represent emotional expressions. The model is designed such that the higher the reward, the lower the loss value, reinforcing the generator's learning as emotional expression improves.

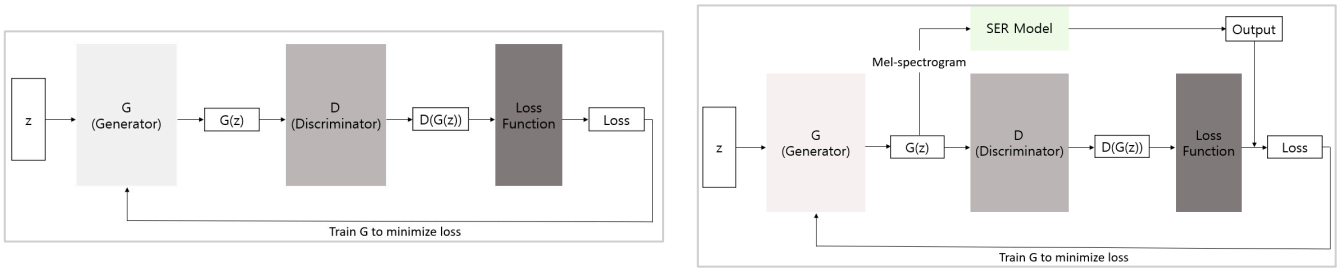
Although the concept of rewards from reinforcement learning is applied, this approach differs from the traditional policy gradient method. While policy gradient methods directly optimize the probability of selecting actions, our approach focuses on minimizing loss based on the probability values provided by the emotion recognition model. Thus, the goal

of this study is not to maximize rewards but to minimize loss based on the accuracy of the emotion recognition model. The equation used in this process is shown in Equation (3).

$$\nabla_\theta J(\theta) = E_{\tau\sim\pi_\theta}[\nabla_\theta log\pi_\theta(a_t|s_t)]R \quad (3)$$

R is a value that quantifies the emotion expression accuracy of the generated Mel-spectrogram, and in this study, it was converted to (1-emotion recognition probability value) to minimize the loss. $log\pi_\theta(a_t|s_t)$ denotes the logarithmic probability of selecting action a_t given state s_t at time t, where a higher value means that the probability of choosing the correct behavior in that state increases. This indicates that the generated Mel-spectrogram is an accurate representation of the emotion.

The optimization method proposed in this study combines the traditional Conditional-DCGAN loss function with the new loss function based on emotion recognition. The gradients for each loss are calculated, normalized, and summed, and the generator learns in the direction of minimizing these combined gradients. By incorporating emotion recognition accuracy into the traditional loss minimization approach, this method contributes to generating Mel-spectrograms that more accurately and precisely reflect emotional expressions. Figure 2 compares the traditional and the newly proposed Conditional-DCGAN.



(**a**) (**b**)

Figure 2. (**a**) Traditional Conditional-DCGAN; (**b**) Newly Proposed Conditional-DCGAN.

3.2.1. Conditional-DCGAN Architecture

Conditional-DCGAN is a model that combines GAN and DCGAN, designed to effectively learn the complex patterns of emotional speech data and generate high-resolution Mel-spectrograms that reflect emotional features. This model uses Mel-spectrograms extracted from voice data as training data and was implemented using the PyTorch 2.2.0 framework. The loss function applied is Binary Cross Entropy, and to enhance training efficiency, the AdamW optimizer—a variant of the traditional Adam optimizer with a weight decay strategy—was used along with the Cosine Annealing Warm Restarts scheduler. In particular, the Cosine Annealing Warm Restarts scheduler helps generate a wide variety of samples by dynamically adjusting the learning rate. This can mitigate the mode collapse that often occurs in GANs by periodically fluctuating the learning rate, forcing the generator to learn different patterns rather than getting stuck on a particular pattern.

The generator is responsible for creating Mel-spectrograms by receiving a latent vector generated from random noise and labels as inputs. The latent vector, combined with the embedded labels, gradually increases the resolution through five deconvolutions [36], eventually generating a Mel-spectrogram of size (1, 64, 64). After each deconvolution, batch normalization [37] is performed to prevent gradient vanishing and improve training stability. The ReLU function [38] is used as the activation function, and the output layer uses the Tanh function to adjust the pixel values of the output Mel-spectrogram within the range of $[-1, 1]$.

The discriminator receives Mel-spectrogram and label data as inputs and determines the authenticity of the input data and the consistency of the labels. The label data is embedded and matched in size with the Mel-spectrogram for combined use. The discriminator consists of five strided convolutions [39], utilizing batch normalization and the LeakyReLU

activation function [40]. The final output uses the sigmoid function to produce a probability value within the range of [0, 1].

The computational cost of conditional-DCGAN grows approximately $O(N^2)$ with the resolution of the image (e.g., 64 × 64) N. The detailed architecture of the model can be found in Table 3.

Table 3. Architecture of the generator and discriminator in Conditional-DCGAN for Mel-Spectro gram augmentation.

Network	Layer	Input Shape	Kernel Shape	Output Shape
Generator	Embedding		-	[6]
	Linear	[106]	-	[16,384]
	ReLU	[16,384]	-	[16,384]
	ConvTranspose2d	[1024, 4, 4]	[4, 4]	[512, 8, 8]
	BatchNorm2d	[512, 8, 8]	-	[512, 8, 8]
	ReLU	[512, 8, 8]	-	[512, 8, 8]
	ConvTranspose2d	[512, 8, 8]	[4, 4]	[256, 16, 16]
	BatchNorm2d	[256, 16, 16]	-	[256, 16, 16]
	ReLU	[256, 16, 16]	-	[256, 16, 16]
	ConvTranspose2d	[256, 16, 16]	[4, 4]	[128, 32, 32]
	BatchNorm2d	[128, 32, 32]	-	[128, 32, 32]
	ReLU	[128, 32, 32]	-	[128, 32, 32]
	ConvTranspose2d	[128, 32, 32]	[4, 4]	[64, 64, 64]
	BatchNorm2d	[64, 64, 64]	-	[64, 64, 64]
	ReLU	[64, 64, 64]	-	[64, 64, 64]
	ConvTranspose2d	[64, 64, 64]	[3, 3]	[1, 64, 64]
	Tanh	[1, 64, 64]	-	[1, 64, 64]
Discriminator	Embedding	[1]	-	[4096]
	Conv2d	[2, 64, 64]	[3, 3]	[64, 64, 64]
	LeakyReLU	[64, 64, 64]	-	[64, 64, 64]
	Conv2d	[64, 64, 64]	[4, 4]	[128, 32, 32]
	BatchNorm2d	[128, 32, 32]	-	[128, 32, 32]
	LeakyReLU	[128, 32, 32]	-	[128, 32, 32]
	Conv2d	[128, 32, 32]	[4, 4]	[256, 16, 16]
	BatchNorm2d	[128, 32, 32]	-	[256, 16, 16]
	LeakyReLU	[256, 16, 16]	-	[256, 16, 16]
	Conv2d	[256, 16, 16]	[4, 4]	[512, 8, 8]
	BatchNorm2d	[256, 16, 16]	-	[512, 8, 8]
	LeakyReLU	[512, 8, 8]	-	[512, 8, 8]
	Conv2d	[512, 8, 8]	[4, 4]	[1, 4, 4]
	AdaptiveAvgPool2d	[1, 4, 4]	-	[1, 1, 1]
	Sigmoid	[1]	-	[1]

The Mel-spectrogram generated by this process was used in experiments as an input to the speech emotion recognition model in combination with the original speech data to show its impact on the performance improvement of the speech emotion recognition model. The model was trained on an NVIDIA GeForce RTX 3090 GPU for about 36 min, and an example of the data generated by the generator can be seen in Figure 3.

3.2.2. CNN + Bi-LSTM Architecture

This study evaluated the impact of the proposed data augmentation technique on the performance of the emotion speech recognition model by using Mel-spectrograms extracted from the original voice data and those generated by the Conditional-DCGAN. The emotion speech recognition model combines CNN and Bi-LSTM. Mel-spectrograms are based on the time axis (X-axis) and frequency axis (Y-axis), representing temporal and frequency changes in the speech signal, while the color (Z-axis) visually displays decibel information. CNN is effective at analyzing the spatial patterns of Mel-spectrograms, enabling it to learn specific frequency bands and temporal features related to emotions in the speech

signal. This reflects how emotions are often tied to specific frequency and volume changes. Bi-LSTM analyzes these temporal changes, recognizing long-term patterns of emotional variations in the speech signal. By combining CNN's ability to learn the spatial features of Mel-spectrograms with Bi-LSTM's ability to capture temporal changes, this model enables more precise and accurate analysis of emotion speech recognition.

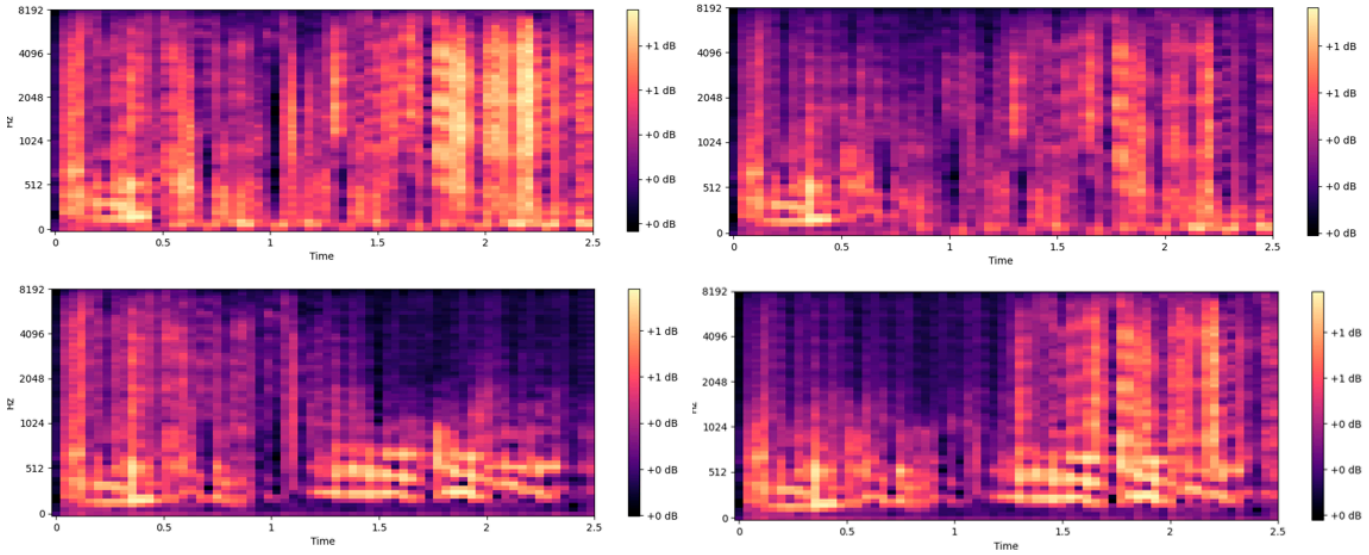


Figure 3. Generated Mel-spectrograms.

The CNN + Bi-LSTM model used in this study consists of three Convolution Layers, one Bi-LSTM Layer, and two Fully Connected Layers. The size of the Convolution Layers gradually increases as the layers deepen, followed by batch normalization and the application of the ELU activation function, with Max-Pooling used to reduce the dimensionality of the data. The extracted features are then flattened and passed to the Bi-LSTM, ultimately producing the final output using the Softmax function to yield class probabilities. This structure is designed to optimize the accuracy of emotion speech recognition, and Dropout was applied during the training process to randomly drop neurons in the hidden layers, preventing the model from overfitting by becoming too adapted to the training data. This visualizes how the proposed data augmentation technique contributes to the performance improvement of the emotion speech recognition model.

4. Experiments

4.1. Experimental Design

In this study, we evaluated the performance improvement of the emotion speech recognition model using both the original Emo-DB dataset and the Mel-spectrogram data generated by the proposed method. The experimental data consisted of Mel-spectrograms extracted from the original Emo-DB, with an additional 300 Mel-spectrograms for each emotional category. The data was split into training and test data in a ratio of 85:15, considering the balance of each dataset, and a 5-fold cross-validation method was applied to evaluate the generalization performance of the model. This method divides the training data into five equal parts and uses each part as a test set sequentially. For performance evaluation, a combined CNN and Bi-LSTM emotion speech recognition model was used. The Cosine Annealing Warm Restarts scheduler [41] was applied for model training, and a mini-batch approach was employed. The Cosine Annealing Warm Restarts scheduler dynamically adjusts the learning rate by gradually decreasing it and then resetting to the maximum learning rate at the end of a cycle. This approach improves the efficiency of the learning process. The optimizer used was AdamW [42], which applies weight decay during parameter updates. AdamW helps prevent overfitting and improves generalization performance by handling weight decay more effectively when the learning rate changes.

The initial learning rate was set to 0.001, and the first cycle was set to 15 epochs, with the cycle length doubling after each restart.

For comparative analysis in the experiments, the results using only the original Emo-DB data were compared with those obtained by combining the original data with the augmented data generated by the proposed method. Additionally, the results were compared with previous studies that used similar methods. The performance metrics used were WA (Weighted Accuracy) and UAR (Unweighted Average Recall), considering class imbalance. WA is calculated by weighting the accuracy of each class by the number of samples in that class, while UAR is the simple average of the recall rates across all classes. These metrics provide a comprehensive evaluation of each model's performance.

$$WA = \frac{\sum_{i=1}^{N} w_i \ * \ ACC_i}{\sum_{i=1}^{N} w_i} \tag{4}$$

$$UAR = \frac{\sum_{i=1}^{N} w_i \ * \ Recall_i}{N} \tag{5}$$

The formulas for calculating WA and UAR are given in Equations (4) and (5). Here, w_i represents the number of samples in class i, ACC_i represents the accuracy of class i, N is the number of classes, and $Recall_i$ represents the recall of class i.

4.2. Experimental Result

In this study, we compared the performance of the original Emo-DB data alone and the data augmented with Conditional-DCGAN. The performance evaluation was performed using 5-fold cross-validation, and the average performance and confidence interval (95%) of each model were used to confirm the consistency of the model performance. Confidence intervals provide a range for how much the average performance measured in an experiment might differ from the actual performance, and the narrower the confidence interval, the more consistent the model's performance, indicating that it will hold up with more data and in different situations. The 5-fold cross-validation results and confidence intervals are presented in Table 4, and the performance of the model using Conditional-DCGAN is (0.89, 0.92) with a 95% confidence interval. This means that the performance of the model is likely to remain stable within the range of 89% to 92%.

Table 4. 5-fold cross-validation and confidence interval results.

Fold	1	2	3	4	5
Accuracy	88.34%	91.47%	90.58%	91.74%	91.07%
Confidence Interval (95%)			(0.89, 0.92)		

Additionally, comparisons were made with other studies that used data augmentation techniques for the Emo-DB dataset or employed CNN or LSTM models. The performance metrics for each method are summarized in Table 5.

The original Emo-DB data achieved a WA of 79.31% and a UAR of 78.16%. In contrast, the model with the augmented data achieved a WA of 91.46% and a UAR of 91.61%. Moreover, the proposed method outperformed the results of studies that applied other data augmentation techniques, such as Shilandari et al. [24], which reported a WA of 84.49% and UAR of 83.33%, and Chauhan et al. [28], which reported a WA of 84.61% and UAR of 82.27%. The performance also exceeded other studies that used CNN or LSTM models, demonstrating that the data augmentation technique proposed in this study is more effective in enhancing the performance of emotion recognition models. In addition, the effects of original data and augmented data can be compared in more detail through performance analysis by emotion class, and these comparisons can be found in Tables 6 and 7.

Table 5. Comparison of Performance Metrics.

Author	Database	Features	Model	WA	UAR
This paper	Emo-DB	Mel-spectrogram	CNN + Bi-LSTM	79.31%	78.16%
This paper	Emo-DB + Augmented	Mel-spectrogram	CNN + Bi-LSTM	91.46%	91.61%
Shilandari et al. [24]	Emo-DB + Augmented	1582 features from OpenSMILE	DNN	84.49%	83.33%
Chauhan et al. [28]	Emo-DB + Augmented	Log Mel-spectrogram	Deep CNN	84.61%	82.27%
Zhang et al. [29]	Emo-DB	32D Low Level Descriptor + 40D High-Level Statistical Functions	CNN + Bi-LSTM	86.44%	84.65%
Andayani et al. [30]	Emo-DB	MFCC	LSTM-Transformer	87.72%	85.55%

Table 6. Confusion Matrix using original Emo-DB only.

		Predicted					
		Anger	**Boredom**	**Fear**	**Happiness**	**Sadness**	**Neutral**
Real	anger	0.85	0.00	0.08	0.07	0.00	0.00
	boredom	0.00	0.83	0.00	0.00	0.17	0.00
	fear	0.20	0.00	0.80	0.00	0.00	0.00
	happiness	0.30	0.00	0.00	0.70	0.00	0.00
	sadness	0.00	0.00	0.11	0.00	0.89	0.00
	neutral	0.00	0.38	0.00	0.00	0.00	0.62

Table 7. Confusion Matrix using original Emo-DB with augmented data.

		Predicted					
		Anger	**Boredom**	**Fear**	**Happiness**	**Sadness**	**Neutral**
Real	anger	0.85	0.02	0.05	0.03	0.02	0.03
	boredom	0.00	0.90	0.02	0.00	0.09	0.00
	fear	0.02	0.02	0.95	0.00	0.02	0.00
	happiness	0.08	0.00	0.00	0.92	0.00	0.00
	sadness	0.00	0.04	0.00	0.00	0.96	0.00
	neutral	0.00	0.06	0.00	0.00	0.02	0.92

5. Conclusions

As a new approach to improve the performance of emotional speech recognition, this study proposes a data augmentation technique using a generative model combining CGAN and DCGAN. This method is an attempt to solve the data sparsity problem faced by existing emotion recognition models and is specifically designed to enable generative models to learn the detailed speech patterns required for emotion recognition. The proposed model proposes a new loss function calculation method, which borrows the reward concept of reinforcement learning and uses the probability value of the emotion recognition model as the reward. This is designed to effectively manage the loss of the generator. Through this reward mechanism, by providing Mel-spectrograms with high probability values in the emotion recognition model as rewards, the generative model can generate more sophisticated emotion data, which has a positive impact on the model performance. The generator used Emo-DB data to generate Mel-spectrograms, and we compared the performance of the emotion speech recognition model using the Mel-spectrograms extracted from the original Emo-DB data and the augmented data. We used Weighted Accuracy (WA) and Unweighted Average Recall (UAR) as performance metrics for comparison. WA is the

accuracy calculated by considering the number of samples in each class as a weight and is a performance evaluation metric that takes into account class imbalance. UAR is the recall of all classes averaged equally, ensuring that all classes are evaluated fairly.

In our experiments, the WA and UAR values were 91.46% and 91.61%, respectively, when the augmented data was included, compared to 79.31% WA and 78.16% UAR when the original data was used alone. It also outperformed the methods of Shilandari et al. [24] and Chauhan et al. [28], which used different data augmentation techniques. This confirms that the proposed data augmentation technique is effective in improving the performance of emotion recognition models.

This study is particularly significant in that it demonstrates that effective learning is possible even when the amount of data is limited, and it is applicable not only to emotion speech data but also to various domains that suffer from data imbalance. However, there is a limitation that some emotional states suffered from mode collapse problem, which did not fully reflect the diversity of certain emotional states. Therefore, although the data augmentation technique proposed in this study is useful, it should be considered that the performance of the model may depend on the quality of the pre-trained emotion recognition model.

In future work, further experiments on various speech datasets and performance comparisons with state-of-the-art pre-trained models are needed to address these issues. Furthermore, it is expected that the accuracy of emotion recognition can be further improved through studies that validate its applicability in real-world applications.

Author Contributions: Conceptualization, K.-M.R. and S.-P.L.; methodology, K.-M.R.; investigation, K.-M.R.; writing—original draft preparation, K.-M.R.; writing—review and editing, S.-P.L.; project administration, S.-P.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the 2024 Research Grant from SangMyung University.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available in [Emo-DB] at [10.21437/Interspeech.2005-446], reference number [13].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Tokuda, K.; Zen, H.; Black, A.W. An HMM-based speech synthesis system applied to English. In Proceedings of the 2002 IEEE Workshop on Speech Synthesis, Santa Monica, CA, USA, 13 September 2002; IEEE: Piscataway, NJ, USA, 2002; pp. 227–230. [CrossRef]
2. Yamagishi, J.; Onishi, K.; Masuko, T.; Kobayashi, T. Modeling of various speaking styles and emotions for HMM-based speech synthesis. In Proceedings of the Eurospeech 2003 8th European Conference on Speech Communication and Technology, Geneva, Switzerland, 1–4 September 2003; pp. 2461–2464. [CrossRef]
3. Schuller, B.; Rigoll, G.; Lang, M. Hidden Markov model-based speech emotion recognition. In Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP'03), Baltimore, MD, USA, 6–9 July 2003; IEEE: Piscataway, NJ, USA, 2003; p. II-1. [CrossRef]
4. Lin, Y.-L.; Wei, G. Speech emotion recognition based on HMM and SVM. In Proceedings of the 2005 International Conference on Machine Learning and Cybernetics, Guangzhou, China, 18–21 August 2005; IEEE: Piscataway, NJ, USA, 2005; pp. 4898–4901. [CrossRef]
5. Lim, W.; Jang, D.; Lee, T. Speech emotion recognition using convolutional and recurrent neural networks. In Proceedings of the 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), Jeju, Korea, 13–16 December 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 1–4. [CrossRef]
6. Popova, A.S.; Rassadin, A.G.; Ponomarenko, A.A. Emotion recognition in sound. Advances in Neural Computation, Machine Learning, and Cognitive Research. In Proceedings of the XIX International Conference on Neuroinformatics, Moscow, Russia, 2–6 October 2017; Springer International Publishing: Cham, Switzerland, 2018; pp. 117–124. [CrossRef]
7. Tzirakis, P.; Zhang, J.; Schuller, B.W. End-to-end speech emotion recognition using deep neural networks. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 5089–5093. [CrossRef]

8. Badshah, A.M.; Ahmad, J.; Rahim, N.; Baik, S.W. Speech emotion recognition from spectrograms with deep convolutional neural network. In Proceedings of the 2017 International Conference on Platform Technology and Service (PlatCon), Busan, Korea, 13–15 February 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1–5. [CrossRef]
9. Rehman, A.; Liu, Z.-T.; Wu, M.; Cao, W.-H.; Jiang, C.-S. Speech emotion recognition based on syllable-level feature extraction. *Appl. Acoust.* **2023**, *211*, 109444. [CrossRef]
10. Hema, C.; Marquez, F.P.G. Emotional speech recognition using cnn and deep learning techniques. *Appl. Acoust.* **2023**, *211*, 109492. [CrossRef]
11. Kim, S.; Lee, S.-P. A BiLSTM–Transformer and 2D CNN Architecture for Emotion Recognition from Speech. *Electronics* **2023**, *12*, 4034. [CrossRef]
12. Pentari, A.; Kafentzis, G.; Tsiknakis, M. Speech emotion recognition via graph-based representations. *Sci. Rep.* **2024**, *14*, 4484. [CrossRef] [PubMed]
13. Burkhardt, F.; Paeschke, A.; Rolfes, M.; Sendlmeier, W.F.; Weiss, B. A database of German emotional speech. In Proceedings of the Ninth European Conference on Speech Communication and Technology, Lisboa, Portugal, 4–8 September 2005; pp. 1517–1520. Available online: https://www.isca-archive.org/interspeech_2005/burkhardt05b_interspeech.html (accessed on 12 March 2024).
14. Livingstone, S.R.; Russo, F.A. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PloS ONE* **2018**, *13*, e0196391. [CrossRef] [PubMed]
15. Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J.N.; Lee, S.; Narayanan, S.S. IEMOCAP: Interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* **2008**, *42*, 335–359. [CrossRef]
16. Krizhevsky, A.; Hinton, G. *Learning Multiple Layers of Features from Tiny Images*; Technical report; University of Toronto: Toronto, ON, Canada, 2009.
17. Jia, D.; Wei, D.; Richard, S.; Li, L.-J.; Kai, L.; Li, F.-F. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; IEEE: Piscataway, NJ, USA, 2009; pp. 248–255. [CrossRef]
18. Mirza, M.; Osindero, S. Conditional generative adversarial nets. *arXiv* **2014**, arXiv:1411.1784.
19. Radford, A.; Metz, L.; Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv* **2015**, arXiv:1511.06434.
20. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 139–144. [CrossRef]
21. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*; MIT Press: Cambridge, MA, USA, 2018.
22. Salazar, A.; Vergara, L.; Safont, G. Generative Adversarial Networks and Markov Random Fields for oversampling very small training sets. *Expert Syst. Appl.* **2021**, *163*, 113819. [CrossRef]
23. Baek, J.-Y.; Lee, S.-P. Enhanced Speech Emotion Recognition Using DCGAN-Based Data Augmentation. *Electronics* **2023**, *12*, 3966. [CrossRef]
24. Shilandari, A.; Marvi, H.; Khosravi, H.; Wang, W. Speech emotion recognition using data augmentation method by cycle-generative adversarial networks. *Signal Image Video Process.* **2022**, *16*, 1955–1962. [CrossRef]
25. Zhu, J.-Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2223–2232. [CrossRef]
26. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
27. Chen, M.; He, X.; Yang, J.; Zhang, H. 3-D Convolutional Recurrent Neural Networks with Attention Model for Speech Emotion Recognition. *IEEE Signal Process. Lett.* **2018**, *25*, 1440–1444. [CrossRef]
28. Chauhan, K.; Sharma, K.K.; Varma, T. Speech emotion recognition using convolution neural networks. In Proceedings of the 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), Coimbatore, India, 25–27 March 2021; IEEE: Cambridge, MA, USA, 2021; pp. 1176–1181. [CrossRef]
29. Zhang, H.; Huang, H.; Han, H. A Novel Heterogeneous Parallel Convolution Bi-LSTM for Speech Emotion Recognition. *Appl. Sci.* **2021**, *11*, 9897. [CrossRef]
30. Andayani, F.; Theng, L.B.; Tsun, M.T.; Chua, C. Hybrid LSTM-Transformer Model for Emotion Recognition from Speech Audio Files. *IEEE Access* **2022**, *10*, 36018–36027. [CrossRef]
31. Jackson, P.; Haq, S. *Surrey Audio-Visual Expressed Emotion (Savee) Database*; University of Surrey: Guildford, UK, 2014.
32. Han, K.; Yu, D.; Tashev, I. Speech emotion recognition using deep neural network and extreme learning machine. In Proceedings of the Fifteenth Annual Conference of the International Speech Communication Association, Singapore, 14–18 September 2014. [CrossRef]
33. McFee, B.; Raffel, C.; Liang, D.; Ellis, D.P.; McVicar, M.; Battenberg, E.; Nieto, O. librosa: Audio and Music Signal Analysis in Python. In Proceedings of the 14th python in science conference 2015, Austin, TX, USA, 6–12 July 2015; pp. 18–24. [CrossRef]
34. Sim, Y.; Choi, J.; Im, S. Synthetic Speech Classification based on Cascade Connection of CNN and MKDE Models. *J. Inst. Electron. Inf. Eng.* **2023**, *60*, 94–101. [CrossRef]
35. Williams, R.J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.* **1992**, *8*, 229–256. [CrossRef]

36. Zeiler, M.D.; Taylor, G.W.; Fergus, R. Adaptive deconvolutional networks for mid and high level feature learning. In Proceedings of the 2011 IEEE International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 2018–2025. [CrossRef]
37. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the 32 nd International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 448–456.
38. Nair, V.; Hinton, G.E. Rectified linear units improve restricted Boltzmann machines. In Proceedings of the 27th International Conference on Machine Learning (ICML-10), Haifa, Israel, 21–24 June 2010; pp. 807–814.
39. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
40. Maas, A.L.; Hannun, A.Y.; Ng, A.Y. Rectifier nonlinearities improve neural network acoustic models. In Proceedings of the 30 th International Conference on Machine Learning, Atlanta, GA, USA, 16–21 June 2013; Volume 30, p. 3.
41. Loshchilov, I.; Hutter, F. SGDR: Stochastic Gradient Descent with Warm Restarts. In Proceedings of the International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017.
42. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv* **2017**, arXiv:1711.05101.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.