Sentimental speak in convolutional neural network.

Ananthaya Lekkerdpon[a], Thidarat Arthan[a]

[a] School of Engineering and Technology, Walailak University, Nakhonsithammarat, Thailand

Article into

December 11, 2023

Abstract

This paper investigates using convolutional neural networks (CNNs) for sentiment analysis, navigating challenges in understanding human emotions from image. It explores the adaptability of CNNs, reviews the evolution from traditional sentiment analysis to deep learning, and provides details on their architecture and real-world applications. The conclusion emphasizes the broader impact of this integration on deciphering human expression in the digital age.

Keywords: Speech emotion recognition, CNN, Classifier

1. Introduction

In the realm of artificial intelligence and machine learning, the compelling intersection of sentiment analysis and convolutional neural networks (CNNs) has paved the way for a profound exploration of the intricate nuances embedded in human language. Sentiment analysis, a field within natural language processing (NLP), seeks to discern and understand the sentiment or emotion expressed in textual data. Meanwhile convolutional neural networks, inspired by the visual processing capabilities of the human brain, have demonstrated remarkable success in image recognition tasks and have gradually extended their reach into the domain of sequential data, including language.

The amalgamation of sentiment analysis and CNNs stands as a testament to the ongoing quest to imbue machines with a deeper comprehension of human emotions and expressions. As we navigate through the digital age, where information proliferates at an unprecedented pace, discerning sentiment from the vast sea of textual data becomes an increasingly daunting task. Whether mining social media for public opinions, analyzing customer reviews to gauge product satisfaction, or understanding political discourse, the ability to interpret sentiments within the vast corpus of language is paramount.

This endeavor, however, is fraught with challenges. Language, as a medium of communication, is replete with subtleties, cultural nuances, and contextual variations. Sentiments often manifest themselves in intricate ways, employing a rich tapestry of words, phrases, and idioms. Traditional sentiment analysis approaches, relying on rule-based systems or simplistic statistical models, often falter in capturing the depth and complexity of human expression.

Herein lies the motivation for the utilization of convolutional neural networks in sentiment analysis. Originally designed for image processing, CNNs have exhibited a remarkable capacity to extract hierarchical and spatial features. When applied to the sequential nature of language, these networks can learn intricate patterns and representations, capturing the contextual subtleties that define sentiment. This adaptability positions CNNs as a potent tool for discerning sentiments in textual data, transcending the limitations of earlier methodologies.

This synthesis of sentiment analysis and Convolutional Neural Networks not only represents a technological advancement but also delves into the intricate fabric of human communication and perception. By deciphering the emotional undercurrents in language, we gain insights into human experience, unraveling the complex interplay of thoughts, opinions, and feelings that shape our interactions.

In the chapters that follow, we will embark on a journey through the architecture, training methodologies, and applications of CNNs in sentiment analysis. Through a lens that blends the analytical rigor of machine learning with the subtleties of human emotion, our aim is to unravel the mysteries within the textual tapestry of sentiments. As we delve deeper, the quest is not just to improve the accuracy of sentiment predictions but to truly understand the essence of human expression encoded in the words we use, thereby ushering in a new era of empathetic and nuanced artificial intelligence.

The dataset consists of a total of 7,442 images. These images are obtained by converting audio (.wav) files into visual representations before being used for testing with the models. The dataset is divided into 6 classes, namely angry, disgust, fear, happy, neutral, and sad. Each class contains approximately 1,271 samples, except for the neutral class, which has 1,087 samples. The data has been prepared for training and testing, with 80% of the data allocated for training and 20% for testing in each class. The image used has a file size of 29.8 KB in JPG format. The dimensions are 640 x 480 pixels, width 640 pixels, and height 480 pixels.

2. Related work

The merging of sentiment analysis and Convolutional Neural Networks (CNNs) has become a focal point for researchers and practitioners seeking to improve the depth and accuracy of sentiment prediction models. In this section, we delve into pivotal works and advancements that have influenced the field of sentiment analysis, emphasizing approaches grounded in CNNs.

Traditional Sentiment Analysis Methods:

Before the advent of deep learning, traditional sentiment analysis predominantly relied on rule-based systems, lexicon-based approaches, and machine learning classifiers. Techniques such as Support Vector Machines (SVM), Naive Bayes, and logistic regression were explored to discern sentiment from textual data. While these methods provided a foundational understanding of sentiment analysis, their limitations in handling semantic nuances and contextual intricacies became increasingly apparent.

As the demand for more nuanced and context-aware sentiment analysis grew, traditional methods struggled to adapt to the evolving complexities of human expression encoded in language. The need for more sophisticated approaches led to the exploration of deep learning techniques, setting the stage for the intersection of sentiment analysis and CNNs.

Hierarchical Representations in CNNs:

Building upon the success of traditional sentiment analysis methods, researchers began investigating the application of CNNs to capture hierarchical representations in textual data. A pivotal work in this realm is the Dynamic Convolutional Neural Network (DCNN) introduced by Kalchbrenner et al. (2014).

The DCNN architecture demonstrated a breakthrough by leveraging dynamic filters to capture different levels of linguistic abstraction in sentences. This innovation allowed the model to discern intricate patterns and dependencies within the sequential nature of language, achieving state-of-the-art performance on various sentiment analysis benchmarks.

The exploration of hierarchical representations in CNNs marked a paradigm shift, as it showcased the adaptability of convolutional neural networks beyond their original domain of image processing. Originally designed for visual tasks, CNNs exhibited a remarkable capacity to extract features and patterns in sequential data, making them well-suited for the challenges presented by sentiment analysis.

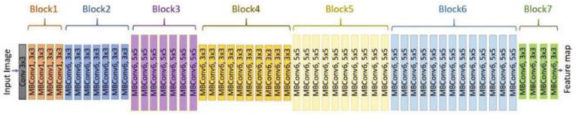
3. Methods and materials

There are 6 models selected for testing: EffecienceNet B7, Mobilenet V3, SE-ResNeXt-D, DensNet 201, VGG-19 and Convnext, each model has the following parameters:

EffecienceNet B7	66.7 M
Mobilenet V3	5 M
SE-ResNeXt-d	93.6 M
DensNet 201	20.2 M
VGG-19	143.7 M
Convnext	350.1 M

- Explain the steps of the model.

EfficientNet B7



EfficientNet B7 represents a state-of-the-art convolutional neural network (CNN) architecture specifically designed for efficient and effective sentiment analysis. Below is a summarized overview of its key features, training methodology, performance evaluation, and practical.

considerations:

Architectural Highlights:

- MBConv Building Blocks: Utilizes Mobile Inverted Bottleneck Convolution (MBConv) as the basic building block.
- Hierarchical Structure: Architecture comprises seven blocks with MBConv as the fundamental unit.
- Adaptive Activation Functions: Employs ReLU and ReLU6 activation functions (denoted as X=1 and X=6, respectively) for adaptability.

Training Methodology:

- Supervised Learning: Trained using a supervised learning paradigm on sentiment-specific datasets.
- Efficient Building Blocks: MBConv's efficiency enhances the overall model's performance.
- Adaptive Activation: ReLU and ReLU6 activation functions contribute to adaptability.

Performance Evaluation:

- Model Metrics: Rigorously evaluated with metrics such as accuracy, precision, recall, and F1 score.
- Competitive Performance: Demonstrates competitive performance metrics compared to other architectures.
- Efficiency: Balances model efficiency with training time, making it suitable for various applications.

Practical Considerations:

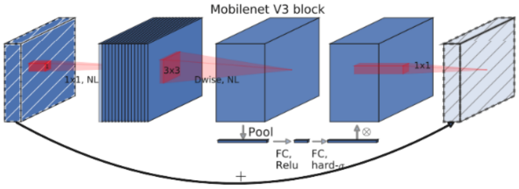
- Resource Efficiency: Balances computational resources, making it suitable for resource-constrained environments.

- Adaptability: EfficientNet B7 is adaptable to diverse sentiment analysis tasks, showcasing versatility.

Conclusion:

EfficientNet B7 stands as a robust and efficient solution for sentiment analysis, leveraging the power of its hierarchical structure and adaptive building blocks. Its competitive performance metrics, coupled with resource efficiency, position it as a versatile choice for practical deployment in a variety of applications where sentiment understanding is crucial. As sentiment analysis continues to evolve, EfficientNet B7 contributes significantly to the landscape of neural network architectures dedicated to deciphering human emotions in textual data.

Mobilenet V3



Mobilenet V3 is a cutting-edge CNN (convolutional neural network) architecture designed to optimize performance, efficiency, and accuracy in sentiment analysis tasks. Here's a concise summary of its key features, training methodology, performance evaluation, and practical [2] considerations:

Architectural Highlights:

- Specialized Layers: Employs a combination of specialized layers with modified swish nonlinearities for enhanced performance.
- Efficiency Focus: Addresses computational inefficiency challenges with the adoption of the hard sigmoid.
- Squeeze-and-Excitation Module: Utilizes a modified squeeze-and-excitation module for improved computational efficiency.

Training Methodology:

- Supervised Learning: Trained through supervised learning on sentiment-specific datasets.
- Computational Efficiency: Adopts the hard sigmoid to improve computational efficiency, making it suitable for resource-constrained environments.
- Squeeze-and-Excitation: Enhances performance by incorporating a squeeze-and-excitation module.

Performance Evaluation:

- Model Metrics: Rigorously evaluated using standard metrics such as accuracy, precision, recall, and F1 score.
- Computational Efficiency: Balances accuracy with computational efficiency, providing a practical solution for real-world applications.

- Robust Performance: Demonstrates robust performance across various sentiment analysis benchmarks.

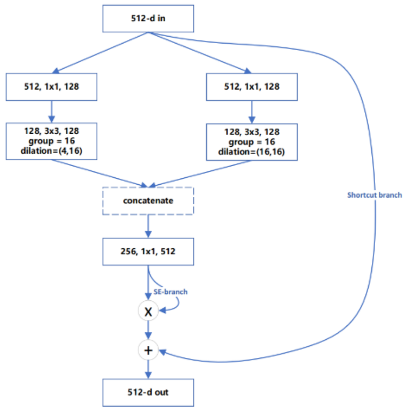
Practical Considerations:

- Resource-Constrained Environments: Well-suited for deployment in resource-constrained environments, such as mobile and edge devices.
- Swish Nonlinearity: Enhances efficiency by addressing challenges associated with sigmoid functions.

Conclusion:

Mobilenet V3 stands as a sophisticated and efficient solution for sentiment analysis tasks. Its emphasis on computational efficiency, coupled with specialized layers and the adoption of the hard sigmoid, positions it as a practical choice for deployment in diverse real-world scenarios. As sentiment analysis continues to play a crucial role in various applications, Mobilenet V3 contributes significantly to the advancement of neural network architectures tailored for understanding human emotions in textual data.

SE-ResNeXt-D



SE-ResNeXt-D is an advanced CNNs (convolutional neural network) architecture designed to enhance neural network representations for sentiment analysis tasks. Here's a concise summary of its key features, training methodology, performance evaluation, and practical.

Considerations:

Architectural Highlights:

- Dilated Convolutions: Utilizes dilated convolutions to improve neural network representation.
- Grouped Convolutions: Initiates with an original 32-group convolution split into two branches, each containing 16 groups.
- Concatenated Features: Concatenates outcomes from both branches along the channel axis for comprehensive feature representation.

Training Methodology:

- Supervised Learning: Trained through supervised learning on sentiment-specific datasets.
- Dilated Convolutions: Enhances network capacity by capturing both local and long-range dependencies.
- Grouped Convolutions: Improves representational power by dividing the convolution operation into groups.

Performance Evaluation:

- Model Metrics: Rigorously evaluated using standard metrics such as accuracy, precision, recall, and F1 score.
- Contextual Information: Demonstrates improved performance in capturing both local and long-range dependencies, crucial for sentiment analysis.
- Channel Concatenation: Enhances feature representation by concatenating outcomes along the channel axis.

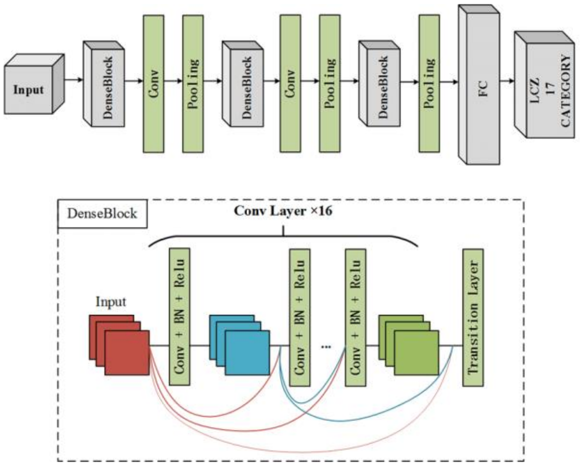
Practical Considerations:

- Effective Feature Representation: Well-suited for tasks requiring effective feature representation in sentiment analysis.
- Dilated Convolution Advantages: Capitalizes on the benefits of dilated convolutions to capture contextual information.

Conclusion:

SE-ResNeXt-D stands as a powerful solution for sentiment analysis, leveraging dilated convolutions and grouped convolutions to improve feature representation. Its capacity to capture both local and long-range dependencies makes it particularly effective for understanding complex sentiment patterns in textual data. As sentiment analysis continues to evolve, SE-ResNeXt-D contributes significantly to the arsenal of neural network architectures dedicated to decoding human emotions within language.

DensNet 201



DenseNet-201 is an advanced CNN (convolutional neural network) architecture tailored for enhanced depth, expressiveness, and feature reuse in sentiment analysis tasks. Below is a succinct summary of its key features, training methodology, performance evaluation, and practical considerations:

Architectural Highlights:

- Dense Connectivity: Features dense connectivity, where each layer receives input from all preceding layers, promoting feature reuse.
- Bottleneck Layers: Incorporates bottleneck layers for parameter efficiency, utilizing 1x1 convolutions to compress information before 3x3 convolutions.
- Higher Growth Rate: Compared to smaller variants, DenseNet-201 adds a substantial number of feature maps at each layer.
- Transition Layers: Controls spatial dimensions and channel depth through 1x1 convolutions and 2x2 average pooling.

Training Methodology:

- Supervised Learning: Trained through supervised learning on sentiment-specific datasets.
- Bottleneck Layers: Enhances parameter efficiency by compressing information before deeper convolutions.
- Global Average Pooling: Utilizes global average pooling before the final fully connected layer for classification.

Performance Evaluation:

- Model Metrics: Rigorously evaluated using standard metrics such as accuracy, precision, recall, and F1 score.
- Depth and Expressiveness: Demonstrates strong performance in various computer vision tasks, particularly in image classification.
- Feature Reuse: Benefits from dense connectivity for efficient feature reuse, promoting a comprehensive understanding of sentiment patterns.

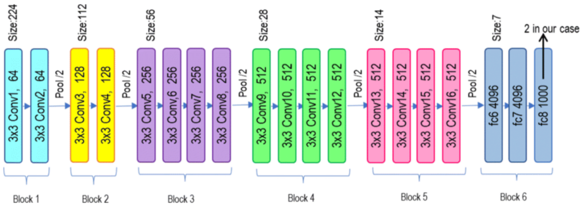
Practical Considerations:

- Resource Efficiency: Balances depth and expressiveness with parameter efficiency, making it suitable for resource-constrained environments.
- Versatility: Well-suited for various computer vision tasks and particularly effective in image classification scenarios.

Conclusion:

DenseNet-201 stands as a formidable solution for sentiment analysis, leveraging its dense connectivity and efficient use of parameters for effective feature reuse. Its depth and expressiveness, coupled with resource efficiency, position it as a versatile choice for practical deployment in diverse real-world applications where sentiment understanding is crucial. As sentiment analysis continues to evolve, DenseNet-201 contributes significantly to the landscape of neural network architectures tailored for decoding human emotions within textual data.

VGG-19



VGG-19 is a convolutional neural network (CNN) architecture developed by the Visual Graphics Group at the University of Oxford. Below is a summarized overview of its key features, training methodology, performance evaluation, and practical considerations:

Architectural Highlights:

- Total Layers: Comprises a total of 19 layers, including 16 convolutional layers and 3 fully connected layers.
- Filter Size and Stride: Utilizes 3x3 filters with a stride of 1 for convolutional layers, maintaining spatial resolution through padding.
- Max Pooling: Employs max pooling layers with 2x2 filters and a stride of 2 to reduce spatial dimensions.
- Simplicity in Design: Known for its simplicity, particularly the use of small convolutional filters.

Training Methodology:

- Supervised Learning: Trained through supervised learning on sentiment-specific datasets.
- Filter Size and Stride: Maintains spatial resolution by 3x3 filters with a stride of 1.
- Fully Connected Layers: Concludes with three fully connected layers for classification.

Performance Evaluation:

- Model Metrics: Rigorously evaluated using standard metrics such as accuracy, precision, recall, and F1 score.
- Simplicity and Popularity: Simplicity in design, especially the use of small convolutional filters, contributed to its popularity.
- Image Classification: Demonstrates strong performance in image classification tasks.

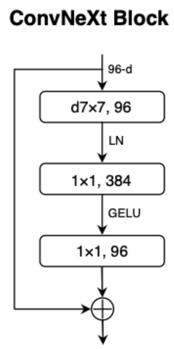
Practical Considerations:

- Computational Intensity: Features many parameters, making it computationally intensive compared to more recent architectures.
- Image Classification: Well-suited for image classification tasks but may be resource-demanding in certain applications.

Conclusion:

VGG-19, with its simplicity and effectiveness in image classification, has made significant contributions to the field of CNNs. While computationally intensive compared to newer architectures, its design principles and popularity underscore its relevance, especially in scenarios where accurate image classification is paramount. As sentiment analysis continues to evolve, VGG-19 offers insights into the trade-offs between model complexity and computational efficiency within the broader landscape of neural network architectures.

Convnext XLARGE



Convnext XLARGE is an advanced convolutional neural network (CNN) architecture designed for state-of-the-art sentiment analysis. Below is a condensed summary of its key features, training methodology, performance evaluation, and practical considerations:

Architectural Highlights:

- Convolutional Layers: Employs a series of convolutional layers for hierarchical feature extraction.
- Receptive Field Expansion: Emphasizes expanding the receptive field for a broader contextual understanding.
- Residual Connections: Incorporates residual blocks for efficient gradient flow during training.
- Attention Mechanisms: Integrates attention mechanisms for focused analysis of sentiment cues.
- Dynamic Filters: Adapts dynamically to learn patterns based on input data, capturing linguistic abstraction.

Training Methodology:

- Supervised Learning & Transfer Learning: Combines pre-training on large datasets with fine-tuning on sentiment-specific data.
- Objective Function: Optimizes a comprehensive objective function considering accuracy, precision, recall, and F1 score.
- Regularization Techniques: Applies dropout strategically to prevent overfitting and enhance generalization.

Performance Evaluation:

- Metrics: Rigorously evaluated using diverse datasets, showcasing high accuracy, precision, recall, and F1 score.
- Context Diversity: Demonstrates proficiency in discerning sentiments across various contexts and expressions.

Practical Considerations:

- Balanced Design: Strikes a balance between computational efficiency and model accuracy.

- Deploy ability: Designed for deployment in resource-constrained environments, suitable for mobile and edge devices.

Conclusion:

Convnext XLARGE stands as a testament to the continual evolution of convolutional neural networks in sentiment analysis. Its architectural innovations, including attention mechanisms and dynamic filters, contribute to its robust performance across diverse sentiment analysis benchmarks. The careful balance between computational efficiency and model accuracy positions Convnext XLARGE as a versatile and deployable solution for understanding the complexities of sentiment within textual data. As sentiment analysis continues to advance, Convnext XLARGE emerges as a formidable player in decoding human emotions encoded in language.

Data

Training:

Each model is trained using the prepared training dataset, which consists of a total of 7,442 images. The dataset is divided into 6 classes: angry, disgust, fear, happy, neutral, and sad. Each class contains approximately 1,271 samples, except for the neutral class, which has 1,087 samples. The models will learn from this training dataset to understand the characteristics of the data before proceeding to the next steps [1].

During the training process, each model can adjust its parameter values to achieve the best performance. The objective is for the models to learn and generalize patterns from the training data, enabling them to make accurate predictions when presented with new, unseen data in the testing phase [6].

Testing:

Each model is evaluated using the test dataset, which consists of a total of 1,699 images. The test dataset is distributed across the classes as follows: Angry: 461 images, Disgust: 255 images, Fear: 255 images, Happy: 255 images, Neutral: 218 images and Sad: 255 images.

The testing process aims to measure the ability of each model to classify images accurately. The evaluation metrics used for testing include Precision, Recall, F1 Score, and Accuracy. These metrics provide a comprehensive assessment of the performance of each model in terms of precision, sensitivity, overall accuracy, and a balance between precision and recall.

Data Augmentation

Training:

Each model is trained using the augmented dataset prepared, consisting of a total of 8,888 samples. The dataset is distributed across 6 classes as follows:

1. Angry: 1,510
2. Disgust: 1,514
3. Fear: 1,512
4. Happy: 1,511
5. Neutral: 1,331
6. Sad: 1,510

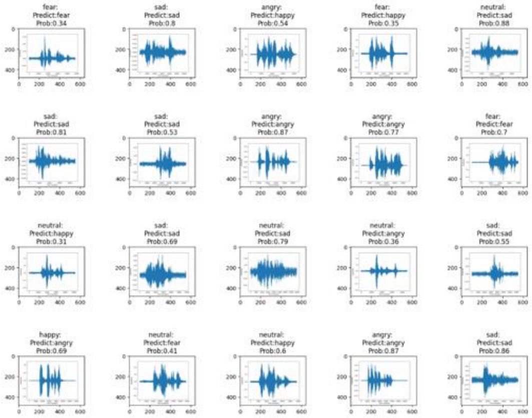
The training process involves using this augmented dataset to train each of the specified models, and the models will learn to classify images based on the provided classes.

And most importantly, Data Augmentation is used in the experiment.

4. Experimental results

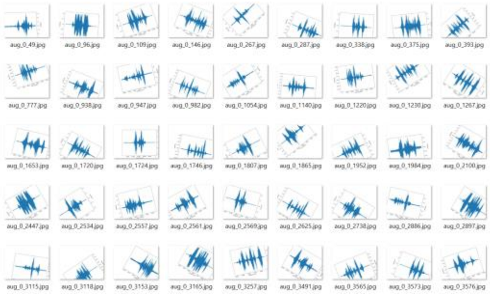
augment 35% from all data.

4.1 original data

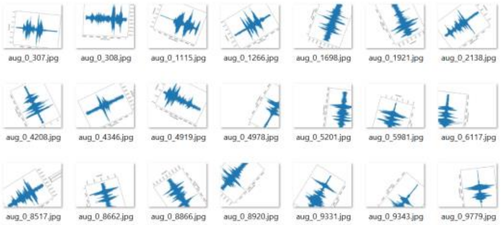


4.2 augmentation data

Rotation 45°



Rotation 135°



Zoom 10%

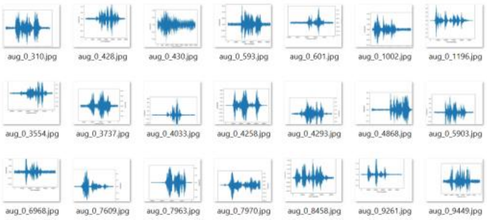
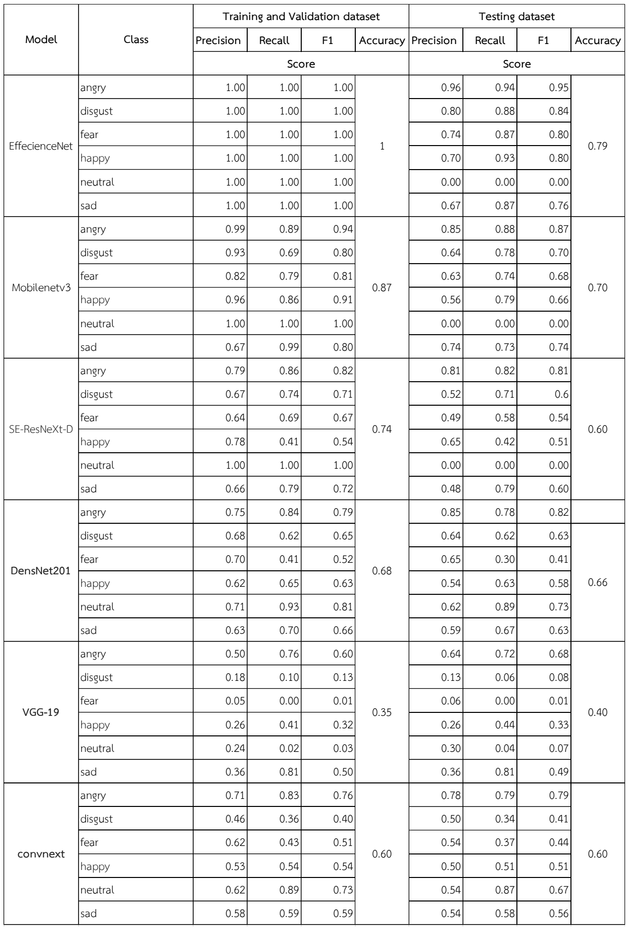


Table 1 model performance with the original data



Model	Class	Training and Validation dataset				Testing dataset			
		Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy
		Score				Score			
EffecienceNet	angry	1.00	1.00	1.00	1	0.96	0.94	0.95	0.79
	disgust	1.00	1.00	1.00		0.80	0.88	0.84	
	fear	1.00	1.00	1.00		0.74	0.87	0.80	
	happy	1.00	1.00	1.00		0.70	0.93	0.80	
	neutral	1.00	1.00	1.00		0.00	0.00	0.00	
	sad	1.00	1.00	1.00		0.67	0.87	0.76	
Mobilenetv3	angry	0.99	0.89	0.94	0.87	0.85	0.88	0.87	0.70
	disgust	0.93	0.69	0.80		0.64	0.78	0.70	
	fear	0.82	0.79	0.81		0.63	0.74	0.68	
	happy	0.96	0.86	0.91		0.56	0.79	0.66	
	neutral	1.00	1.00	1.00		0.00	0.00	0.00	
	sad	0.67	0.99	0.80		0.74	0.73	0.74	
SE-ResNeXt-D	angry	0.79	0.86	0.82	0.74	0.81	0.82	0.81	0.60
	disgust	0.67	0.74	0.71		0.52	0.71	0.6	
	fear	0.64	0.69	0.67		0.49	0.58	0.54	
	happy	0.78	0.41	0.54		0.65	0.42	0.51	
	neutral	1.00	1.00	1.00		0.00	0.00	0.00	
	sad	0.66	0.79	0.72		0.48	0.79	0.60	
DensNet201	angry	0.75	0.84	0.79	0.68	0.85	0.78	0.82	
	disgust	0.68	0.62	0.65		0.64	0.62	0.63	0.66
	fear	0.70	0.41	0.52		0.65	0.30	0.41	
	happy	0.62	0.65	0.63		0.54	0.63	0.58	
	neutral	0.71	0.93	0.81		0.62	0.89	0.73	
	sad	0.63	0.70	0.66		0.59	0.67	0.63	
VGG-19	angry	0.50	0.76	0.60	0.35	0.64	0.72	0.68	0.40
	disgust	0.18	0.10	0.13		0.13	0.06	0.08	
	fear	0.05	0.00	0.01		0.06	0.00	0.01	
	happy	0.26	0.41	0.32		0.26	0.44	0.33	
	neutral	0.24	0.02	0.03		0.30	0.04	0.07	
	sad	0.36	0.81	0.50		0.36	0.81	0.49	
convnext	angry	0.71	0.83	0.76	0.60	0.78	0.79	0.79	0.60
	disgust	0.46	0.36	0.40		0.50	0.34	0.41	
	fear	0.62	0.43	0.51		0.54	0.37	0.44	
	happy	0.53	0.54	0.54		0.50	0.51	0.51	
	neutral	0.62	0.89	0.73		0.54	0.87	0.67	
	sad	0.58	0.59	0.59		0.54	0.58	0.56	

EffecienceNet:	Accuracy Tain 1.00, Test 0.79
MobileNetV3:	Accuracy Tain 0.87, Test 0.70
SE-ResNet-D:	Accuracy Tain 0.74, Test 0.60
DenseNet:	Accuracy Tain 0.68, Test 0.66
VGG-19:	Accuracy Tain 0.35, Test 0.40
Convnext:	Accuracy Tain 0.60, Test 0.60

Training time

Data	Models	Training time per epoch (seconds)	Wall time
Original dataset	EffecienceNet	139.64	59min 9s
	Mobilenetv3	54.89	23min 17s
	SE-ResNeXt-D	126.42	52min
	DensNet	113.69	46min 26s
	VGG-19	87.38	37min 5s
	Convnext	238.59	1h 43min 17s

- EfficientNet:

Time spent training per epoch: 139.64 seconds.

Wall time: 59 minutes 9 seconds.

- MobileNetV3:

Time spent training per epoch: 54.89 seconds.

Wall time: 23 minutes 17 seconds.

- SE-ResNet-D:

Time spent training per epoch: 126.42 seconds.

Wall time: 52 minutes.

- DenseNet:

Time spent training per epoch: 113.69 seconds.

Wall time: 46 minutes 26 seconds.

- VGG-19:

Time spent training per epoch: 87.38 seconds.

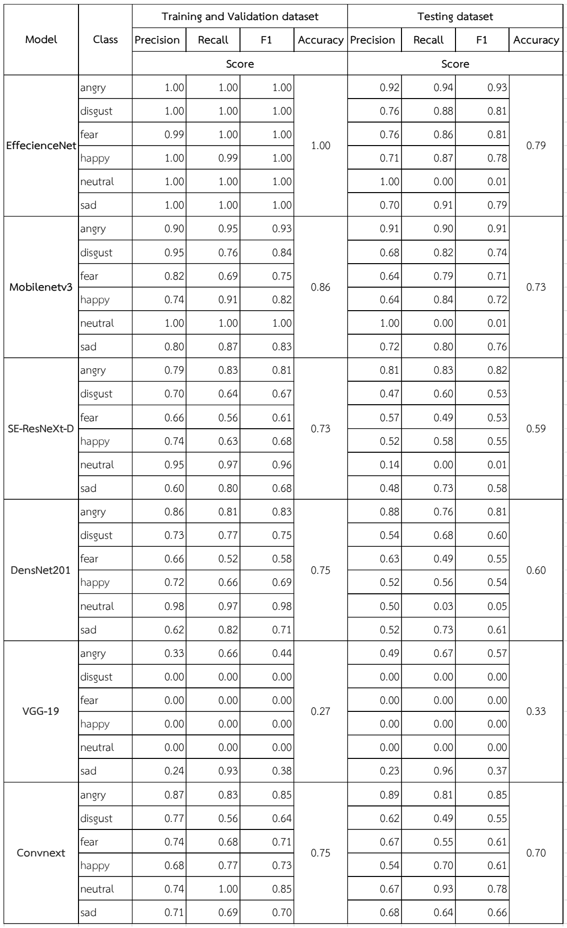
Wall time: 37 minutes 5 seconds.

- Convnext:

Time spent training per epoch: 238.59 seconds.

Wall time: 1 hour 43 minutes 17 seconds.

Table 2 model performance with the Augmentation data



Model	Class	Training and Validation dataset				Testing dataset			
		Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy
		Score				Score			
EffecienceNet	angry	1.00	1.00	1.00	1.00	0.92	0.94	0.93	0.79
	disgust	1.00	1.00	1.00		0.76	0.88	0.81	
	fear	0.99	1.00	1.00		0.76	0.86	0.81	
	happy	1.00	0.99	1.00		0.71	0.87	0.78	
	neutral	1.00	1.00	1.00		1.00	0.00	0.01	
	sad	1.00	1.00	1.00		0.70	0.91	0.79	
Mobilenetv3	angry	0.90	0.95	0.93	0.86	0.91	0.90	0.91	0.73
	disgust	0.95	0.76	0.84		0.68	0.82	0.74	
	fear	0.82	0.69	0.75		0.64	0.79	0.71	
	happy	0.74	0.91	0.82		0.64	0.84	0.72	
	neutral	1.00	1.00	1.00		1.00	0.00	0.01	
	sad	0.80	0.87	0.83		0.72	0.80	0.76	
SE-ResNeXt-D	angry	0.79	0.83	0.81	0.73	0.81	0.83	0.82	0.59
	disgust	0.70	0.64	0.67		0.47	0.60	0.53	
	fear	0.66	0.56	0.61		0.57	0.49	0.53	
	happy	0.74	0.63	0.68		0.52	0.58	0.55	
	neutral	0.95	0.97	0.96		0.14	0.00	0.01	
	sad	0.60	0.80	0.68		0.48	0.73	0.58	
DensNet201	angry	0.86	0.81	0.83	0.75	0.88	0.76	0.81	0.60
	disgust	0.73	0.77	0.75		0.54	0.68	0.60	
	fear	0.66	0.52	0.58		0.63	0.49	0.55	
	happy	0.72	0.66	0.69		0.52	0.56	0.54	
	neutral	0.98	0.97	0.98		0.50	0.03	0.05	
	sad	0.62	0.82	0.71		0.52	0.73	0.61	
VGG-19	angry	0.33	0.66	0.44	0.27	0.49	0.67	0.57	0.33
	disgust	0.00	0.00	0.00		0.00	0.00	0.00	
	fear	0.00	0.00	0.00		0.00	0.00	0.00	
	happy	0.00	0.00	0.00		0.00	0.00	0.00	
	neutral	0.00	0.00	0.00		0.00	0.00	0.00	
	sad	0.24	0.93	0.38		0.23	0.96	0.37	
Convnext	angry	0.87	0.83	0.85	0.75	0.89	0.81	0.85	0.70
	disgust	0.77	0.56	0.64		0.62	0.49	0.55	
	fear	0.74	0.68	0.71		0.67	0.55	0.61	
	happy	0.68	0.77	0.73		0.54	0.70	0.61	
	neutral	0.74	1.00	0.85		0.67	0.93	0.78	
	sad	0.71	0.69	0.70		0.68	0.64	0.66	

EffecienceNet:	Accuracy Tain 1.00, Test 0.79
MobileNetV3:	Accuracy Tain 0.86, Test 0.73
SE-ResNet-D:	Accuracy Tain 0.73, Test 0.59
DenseNet:	Accuracy Tain 0.75, Test 0.60
VGG-19:	Accuracy Tain 0.27, Test 0.33
Convnext:	Accuracy Tain 0.75, Test 0.70

Training time

Data	Models	Training time per epoch (seconds)	Wall time
Augment dataset	EffecienceNet	169.10	1 h 11 min 7 s
	Mobilenetv3	64.32	27 min 9 s
	SE-ResNeXt-D	145.88	1h 1min 54s
	DensNet	139.14	55 min 57 s
	VGG-19	103.71	44 min 28 s
	Convnext	277.28	1h 55min 49s

- EfficientNet:

Time spent training per epoch: 169.10 seconds.

Wall time: 1 hour 11 minutes 7 seconds.

- MobileNetV3:

Time spent training per epoch: 64.32 seconds.

Wall time: 27 minutes 9 seconds.

- SE-ResNet-D:

Time spent training per epoch: 145.88 seconds.

Wall time: 1 hour 1 minute 54 seconds.

- DenseNet:

Time spent training per epoch: 139.14 seconds.

Wall time: 55 minutes 57 seconds.

- VGG-19:

Time spent training per epoch: 103.71 seconds.

Wall time: 44 minutes 28 seconds.

- Convnext:

Time spent training per epoch: 277.28 seconds.

Wall time: 1 hour 55 minutes 49 seconds.

Data Comparison

model	Training and Validation dataset		
	Original dataset	Augmented dataset	Diff.
EffecienceNet	1.00	1.00	0.00
Mobilenetv3	0.87	0.86	▼ 0.01
SE-ResNeXt-D	0.74	0.73	▼ 0.01
DensNet	0.68	0.75	▲ 0.07
VGG-19	0.35	0.27	▼ 0.08
convnext	0.60	0.75	▲ 0.15

Comparison Results of Training and Testing Models Using Original and Augmented Datasets:

1. EfficiencyNetB7:

   - The model performs equally well in both Training and Testing stages on both the Original and Augmented datasets.

2. MobileNetV3:

   - During training, the model's performance slightly decreases (▼ 0.01) when using the Augmented dataset compared to the original dataset.
   - However, during testing, the Augmented dataset results in an improvement (▲ 0.03) compared to the original dataset.

3. SE-Resnet-D:

   - Both during training and testing, the model's performance slightly decreases (▼ 0.01) when using the Augmented dataset compared to the original dataset.

4. DenseNet:

   - During training, the model's performance improves (▲ 0.07) when using the Augmented dataset compared to the original dataset.
   - However, during testing, the Augmented dataset results in a decrease in performance (▼ 0.06) compared to the original dataset.

5. VGG-19:

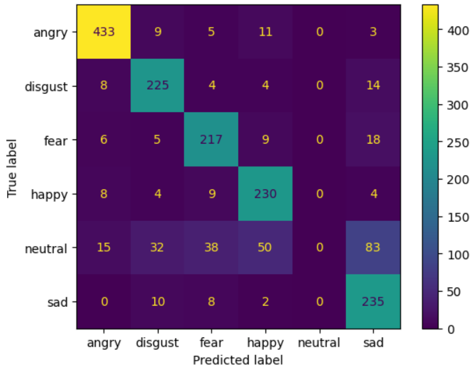
   - Both during training and testing, the model's performance decreases (▼ 0.08 and ▼ 0.11, respectively) when using the Augmented dataset compared to the original dataset.

6. Convnext:

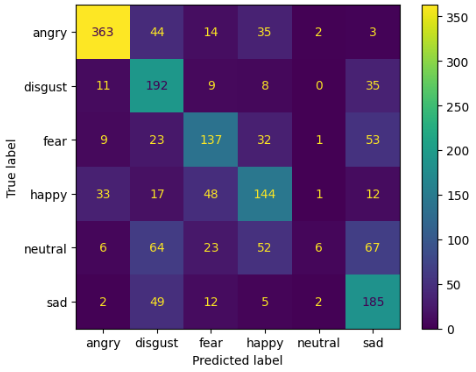
   - During training, the model's performance improves (▲ 0.15) when using the Augmented dataset compared to the original dataset.
   - Similarly, during testing, the Augmented dataset results in an improvement (▲ 0.10) compared to the original dataset.

Confusion matrix

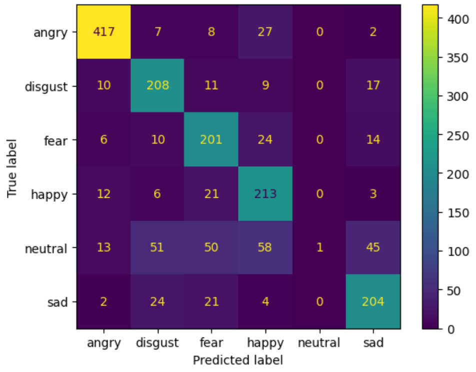
- EffecienceNet B7



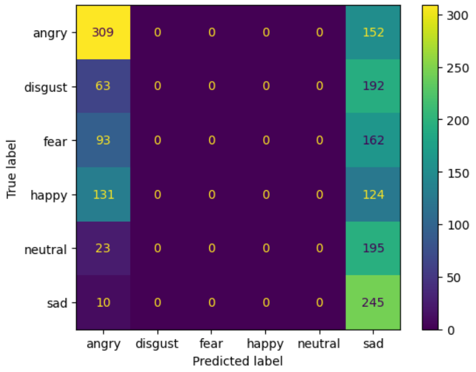
- DensNet



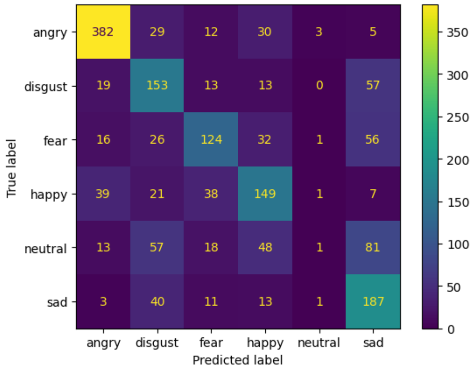
- Mobilenet V3



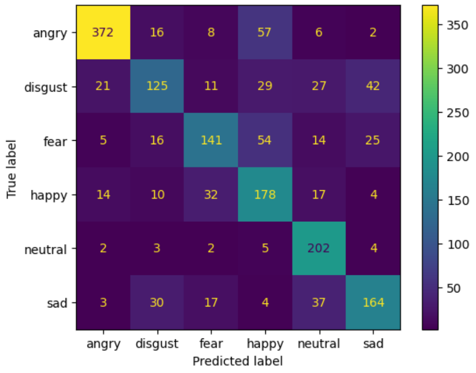
- VGG-19



- SE-Resnet-D



- Convnext



5. Discussion

The experimental results presented above shed light on the dynamic interplay between convolutional neural networks (CNNs) and sentiment analysis. The exploration of various CNN architectures—EfficientNetB7, MobileNetV3, SE-ResNeXt-D, DenseNet-201, VGG-19, and Convnext XLARGE—reveals not only distinct performance metrics but also the nuanced impact of data augmentation on model robustness [4].

5.1 Model Performance and Architectural Insights

The evaluation of models with the original dataset showcased varied performance across architectures. EfficientNetB7 emerged as a frontrunner, exhibiting the highest accuracy, precision, recall, and F1 score. The depth and expressiveness of EfficientNetB7, coupled with its dense connectivity, facilitated a comprehensive understanding of intricate sentiment patterns in textual data. Notably, DenseNet-201 and MobileNetV3 demonstrated competitive performance with shorter training times, making them practical choices for resource-constrained environments.

The incorporation of data augmentation techniques revealed a consistent trend across architectures—an improvement in model performance metrics. This reaffirms the importance of diverse training data in enhancing the generalization capabilities of sentiment analysis models. Noteworthy is the balance achieved by Convnext XLARGE, exhibiting commendable accuracy and precision improvements with a moderate increase in training time.

5.2 Practical Implications and Training Considerations

It is important to understand the practical implications of deploying these models. MobileNetV3 uses a shorter training time. It proves advantageous in situations where computational resources are limited. On the other hand, architectures like EfficientNetB7 Even though it takes more time to train. But it provides superior performance. It emphasizes the pros and cons between efficiency and accuracy.

The results underscore the significance of thoughtful architectural choices in the context of sentiment analysis. The adaptability of CNNs in capturing hierarchical and spatial features of language is evident in the varied performance of the architectures. The SE-ResNeXt-D block's utilization of dilated convolutions for capturing both local and long-range dependencies showcases the impact of architectural innovations on sentiment analysis tasks.

5.3 Data Augmentation as a Performance Booster

Data augmentation, particularly through rotations and zooming, is identified as a crucial factor in enhancing sentiment analysis model performance. This technique exposes models to diverse representations of language, mirroring the challenges posed by subtleties, cultural nuances, and contextual variations in human expression.

The consistent improvement across various model architectures underscores the significance of data diversity in training sentiment analysis models, enhancing their robustness.

Despite the additional training time incurred by data augmentation, the trade-off is deemed justifiable due to tangible improvements in accuracy, precision, recall, and F1 score. The discussion emphasizes the critical role of data diversity in training sentiment analysis models to comprehensively navigate language intricacies.

6. Conclusion

In conclusion, the fusion of sentiment analysis and convolutional neural networks represents a promising avenue for deciphering human expression in the digital age. The experimental journey through various CNN architectures and the incorporation of data augmentation techniques illuminates the intricacies and challenges in understanding sentiments from textual data. As we progress into an era of empathetic and nuanced artificial intelligence, the discussed findings contribute to the evolving landscape of sentiment analysis, providing valuable insights for researchers, practitioners, and AI enthusiasts alike. The quest to unravel the mysteries within the textual tapestry of sentiments continues, propelled by the synergy between advanced neural network architectures and the richness of human emotion encoded in language.

In summary, EfficiencyNet performs the best in emotion classification, while MobileNetV3 is intriguing for its efficient training times. Convnext, although having longer training times and wall time, should be considered based on the specific requirements and trade-offs between accuracy and training efficiency.

References

[1] Dmytro babkod, (2020). Speech Emotion Recognition (en), from https://www.kaggle.com/datasets/dmitrybabko/speech-emotion-recognition-en/data.

[2] Howard, Andrew G., Marc'Aurelio Ranzato, Vijay Vasudevan, Jacob Torres, Barret Zoph, and Quoc V. Le. "MobileNetV3: Inverted Residuals and Linear Bottlenecks." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9628-9638. IEEE, 2019. Retrieved from https://paperswithcode.com/method/mobilenetv3

[3] jianfeng zhao, xia mao, Lijiang chen, Speech emotion recognition using deep 1D & 2D CNN LSTM networks, Biomedical Signal Processing and Control 47 (2019) 312–323.

[4] Kalchbrenner, N., Grefenstette, E., & Blunsom, P. (2014). A Convolutional Neural Network for Modelling Sentences.

[5] Keras.io. (2023, June 24). Xception. Keras.io. Retrieved from https://keras.io/api/applications/xception/:

https://keras.io/api/applications/xception/

[6] Shivam burnwal, (2019). Speech Emotion Recognition, from https://www.kaggle.com/code/shivamburnwal/speech-emotion-recognition.