Hindawi
Scientific Programming
Volume 2021, Article ID 9916915, 12 pages
https://doi.org/10.1155/2021/9916915

Research Article

The Investigation of Different Loss Functions with Capsule Networks for Speech Emotion Recognition

Anfernee Joan B. Ng and Kun-Hong Liu

School of Informatics, Xiamen University, Xiamen, Fujian, China

Correspondence should be addressed to Kun-Hong Liu; lkhqz@xmu.edu.cn

Received 30 May 2021; Revised 1 August 2021; Accepted 8 August 2021; Published 18 August 2021

Academic Editor: Antonio J. Peña

Speech emotion recognition (SER) is an important research topic. Image features like spectrograms are one of the common ways of extracting information from speech. In the area of image recognition, a relatively novel type of network called capsule networks has shown good and promising results. This study aims to use capsule networks to encode spatial information from spectrograms and analyse its performance when paired with different loss functions. Experiments comparing the capsule network with models from previous works show that the capsule network performs better than them.

1. Introduction

The research field of speech emotion recognition (SER) has a wide range of applications that benefit areas such as human-computer interaction, customer service, and computer games [1]. The general motivation is to identify the emotional state to provide a more personalized and often better user experience. For example, customer service systems can use SER to determine whether a customer is angry or dissatisfied with the aid of their voice throughout the call [2].

In recent years, deep learning is a common framework that has been used in a variety of fields, including SER [3]. One main benefit of using deep learning models is their innate ability to learn new features from a given set of data. Convolutional neural networks (CNNs) are typically used as the basic framework, resulting in many improvements and variations for the CNN in SER [4, 5]. Similarly, recurrent neural networks (RNNs) take advantage of the time dimension in speech and can extract better features that consider temporal relationships between points in a speech sample. Among RNNs, variations like long short-term memory (LSTM) networks and gated recurrent units (GRUs) are also widely used as the main framework in SER research [6, 7].

Another deep learning framework that has been on the trend recently is the capsule network [8]. Its conception mainly addresses the shortcomings of CNNs, including their insensitivity to changes in orientation like rotation and translation. Capsule networks achieve this by using a structure composed of a group of neurons called a capsule. Rather than receiving scalar values from individual neurons on traditional deep neural networks (DNNs), output values are instead vectors whose length and direction describe the pose, orientation, and probability of the existence of the entity being predicted or classified. Like traditional DNNs, the capsule network can be divided into different levels or layers of capsules. The first layer usually handles primitive or roughly simple entities like lines, and further layers manage more complex objects like lines joining together to make an object. Low-level capsules would pass their vector outputs to higher-level capsules, which tend to agree or complement with their outputs. The agreement is analogous to a simple table composed of its individual parts like the legs and surface. The individual parts are situated on a lower layer (legs and surface), which look for capsules in a higher layer (whole table) that "agree" with them. This agreement is determined by applying dynamic routing or routing-by-agreement.

Table 1: Loss functions analysed in this paper. y is the true label encoded in one-hot form, $\widehat{y}$ is the true label in +1/−1 encoding, $\sigma(\cdot)$ denotes probability estimate.

Name	Formula
L1 loss	$\|y - q\|_1$
L2 loss	$\|y - q\|_2^2$
Chebyshev loss	$\max_k \lvert\sigma(q)_k - y_k\rvert$
Hinge loss	$\sum_k \max(0, (1/2) - \widehat{y}_k q_k)$
Squared hinge loss	$\sum_k \max(0, (1/2) - \widehat{y}_k q_k)^2$
Cubed hinge loss	$\sum_k \max(0, (1/2) - \widehat{y}_k q_k)^3$
Tanimoto loss	$\sum_k \widehat{y}_k \sigma(q)_k / \|\widehat{y}_k\|_2^2 + \|\sigma(q)_k\|_2^2 - \sum_k \widehat{y}_k \sigma(q)_k$
Cauchy–Schwarz loss	$-\log(\sum_k \widehat{y}_k \sigma(q)_k / \|\widehat{y}_k\|_2 \|\sigma(q)_k\|_2)$

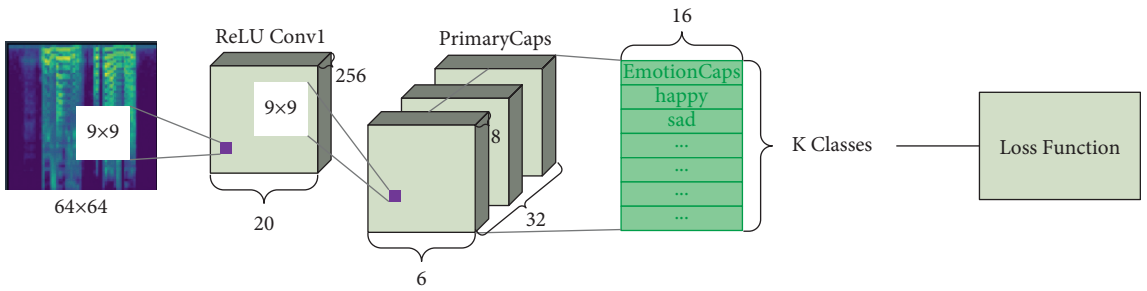

Figure 1: Capsule network architecture.

One important consideration in performing deep learning or machine learning in general is the choice of the loss function. Most of the capsule network implementations in other literature [9–11] use the original margin loss as described by Sabour et al. [8]. Only a few have attempted deviating from the original implementation and instead have employed other loss functions. Previous studies [12, 13] have designed custom loss functions but for a specific area or field. To the best of the authors' knowledge, no other existing literature has reported on the effect of different loss functions used in conjunction with a capsule network. Atmaja and Akagi [14] have published research papers on the analysis of loss functions in the field of SER, but they have not covered them with capsule networks. It is sufficient to say that the impacts of various loss functions on a capsule network are not well understood. If the effects of these loss functions are better understood, then the construction and design of future capsule networks will be more well-informed and easier. In addition, when the choice of a loss function is made easier, researchers can focus on other aspects of their deep learning capsule framework, thereby speeding up their research. As such, the main contribution of this paper is to explore the impacts of different loss functions with the use of a capsule network. Furthermore, this paper also provides insights on the usefulness of these loss functions on multiple SER data sets.

This paper aims to provide an experimental analysis of applying other kinds of loss functions to a capsule network. In a sense, this extends the work done by Janocha and Czarnecki [15], using some of the loss functions experimented there and applying them to a capsule network. The data sets in this paper also differ from the original literature; all of them are taken from the field of SER. In addition, a few baseline models from other papers are tested and compared with the capsule network. Results show that the capsule network architecture performs slightly better than these baselines.

The remaining contents of this paper are organized as follows. Chapter 2 lays the foundation and theoretical bases needed to understand the model and loss functions analysed in this paper. The same chapter also mentions and explores relevant literature. Chapter 3 explains the methods used in the experiments along with the data sets used. Finally, Chapter 4 provides results and discussion of the said experiments.

2. Relevant Theoretical Bases and Literature

2.1. Recent Advancements. Different techniques in SER classification have been constantly developed and improved over the years. Some have extracted novel types of features like adaptive time-frequency features [16] based on the fractional Fourier transformation and frequency modulation features [17] based on the amplitude modulation-frequency modulation model. In contrast to designing new kinds of features, Özseven [18] instead proposes a novel feature-selection method. The new method involves using multiple statistical measures that are then filtered through a threshold calculated from standard deviations and means between emotional classes.

Aside from features, several previous studies also made improvements on common deep learning models used in SER, such as CNNs and LSTMs. For instance, an ensemble combining DNNs, CNNs, and RNNs was used by Yao et al. [19] to provide different types of features. A confidence-based fusion strategy was also proposed to combine the outputs of these networks in classification. Zhao et al. [20] used different dimensions of CNNs to extract features of

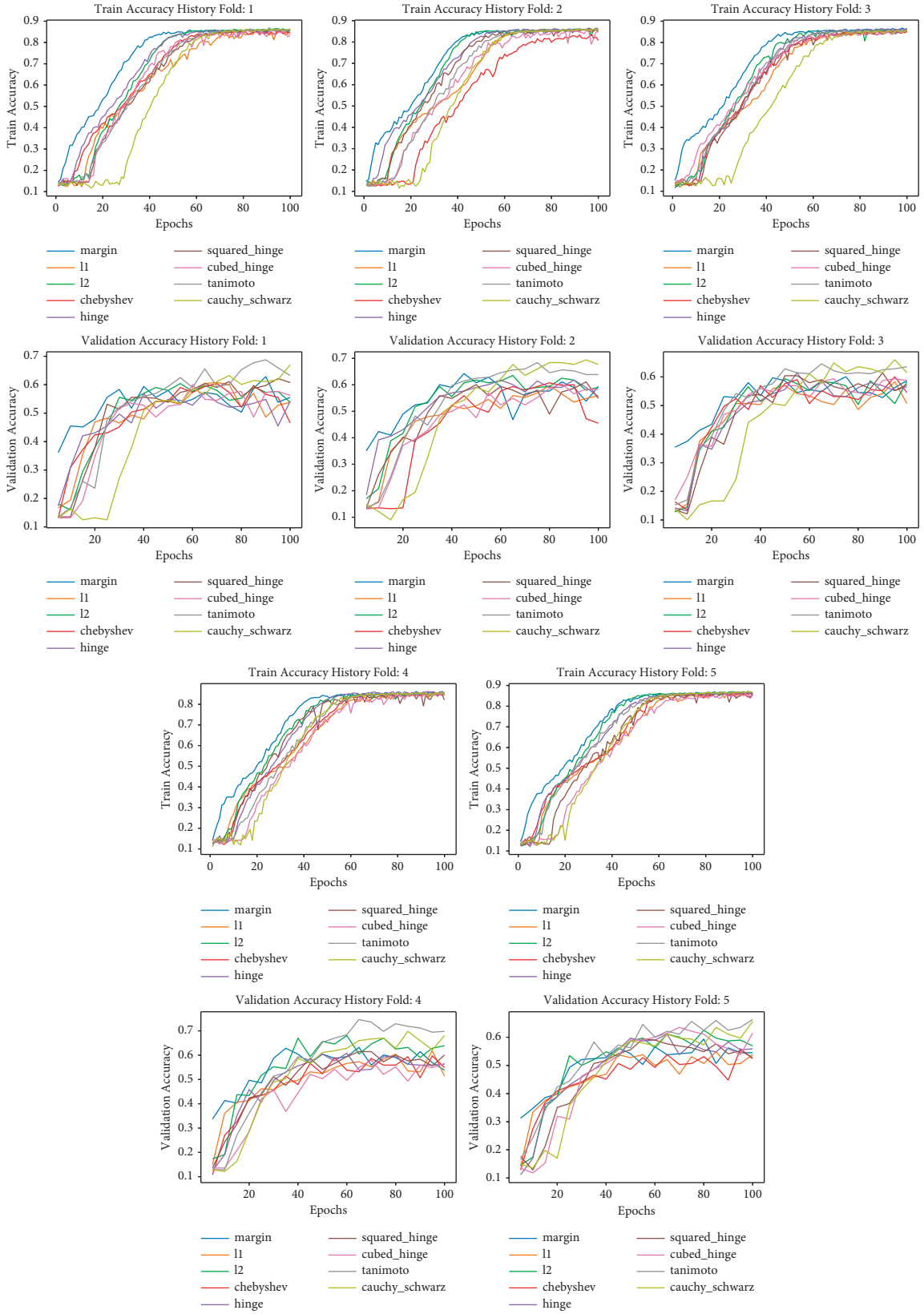

Figure 2: RAVDESS train and validation accuracy history for capsule model and different loss functions.

Table 2: RAVDESS class F1 scores and accuracies.

Loss type	Neutral (F1)	Calm (F1)	Happy (F1)	Sad (F1)	Angry (F1)	Fearful (F1)	Disgust (F1)	Surprise (F1)	Overall (F1)	Overall (Acc.)
Margin loss	51.83	70.68	54.09	44.38	69.97	67.07	71.40	**74.87**	63.04	64.38
L1 loss	55.54	66.40	49.92	39.41	67.54	57.46	63.96	67.23	58.43	58.26
L2 loss	58.93	67.03	59.04	46.36	73.16	65.32	64.47	73.04	63.42	64.31
Chebyshev loss	53.36	67.87	49.43	36.30	68.89	56.89	64.52	68.27	58.19	58.89
Hinge loss	53.97	67.67	49.30	39.18	68.50	61.52	62.51	67.91	58.82	59.37
Square hinge loss	51.76	69.78	55.98	49.58	75.56	67.70	69.62	67.91	63.48	64.51
Cubed hinge loss	52.40	73.83	53.09	46.35	69.67	64.16	66.85	70.12	62.06	63.40
Tanimoto loss	**66.55**	76.75	**59.72**	55.09	**77.12**	**69.41**	73.67	69.05	**68.42**	**69.03**
Cauchy–Schwarz loss	62.25	**77.72**	58.19	**57.49**	75.80	67.78	**75.24**	71.91	68.30	68.96

varying granularity, which are then passed to an LSTM network. The role of the LSTM network is to learn global contextual information from the CNN's resulting features. The researchers discovered that the 2D CNN LSTM network performed better.

2.2. Capsule Network. The basic unit for computation in a capsule network is the namesake itself—"capsule," which is simply a group of neurons. Unlike regular neurons, capsules output vectors whose length and direction can describe an entity or an object. The length of the vector would represent the probability of the object's existence in the scene, while the direction or instantiation parameters would provide information on the position, orientation, size, and other properties. A typical network comprises few layers of capsules, with each layer responsible for checking objects of different size or complexity. The first layer is tasked to check for simple or small objects, while the subsequent layers build upon the existence of these primitive objects to compose larger ones. Higher-level capsules do this by receiving activations from lower-level capsules, which are so-called "components" of the more complex object it is trying to predict.

The network determines these lower-to-higher-level capsule relationships using an iterative dynamic routing mechanism. In a nutshell, the dot product is calculated from the "prediction vectors" taken from the previous and the output vector of the current layer and then used to update coupling coefficients which can either strengthen or weaken the relationship between a capsule in the preceding and current layer. In mathematical terms, it can be formulated as

$$\begin{aligned} \mathbf{o_j} &= \sum_i c_{ij} \widehat{u}_{\mathbf{j|i}}, \\ \widehat{u}_{\mathbf{j|i}} &= \mathbf{W_{ij}} \mathbf{u_i}, \end{aligned} \tag{1}$$

where c_{ij} are the coupling coefficients which are updated at each routing iteration, $\widehat{u}_{\mathbf{j|i}}$ is the prediction vector of the previous layer produced by multiplying weight matrix $\mathbf{W_{ij}}$ and output vector $\mathbf{u_i}$ of the previous layer, and $\mathbf{o_j}$ is the preactivation vector for the next layer. This activation function is the squash which ensures that $\mathbf{o_j}$ shrinks to a vector with a length from 0 to 1. The function also has the effect of producing vectors with length close to 0 for short vectors while producing vectors with length close to 1 for long vectors.

$$\mathbf{v_j} = \text{squash}\left(\mathbf{o_j}\right) = \frac{\left\|\mathbf{o_j}\right\|^2}{1 + \left\|\mathbf{o_j}\right\|^2} \frac{\mathbf{o_j}}{\left\|\mathbf{o_j}\right\|}. \tag{2}$$

Furthermore, the coupling coefficients c_{ij} are calculated from initial logits b_{ij} which are the log prior probabilities that capsule i should be paired with capsule j. The calculations are designed in such a way that c_{ij} from one specific capsule i all sum up to unity, termed "routing softmax":

$$c_{ij} = \frac{\exp\left(b_{ij}\right)}{\sum_k \exp\left(b_{ik}\right)}. \tag{3}$$

Finally, b_{ij} is updated (thereby updating c_{ij} as well) by adding the scalar product $\mathbf{v_j} \cdot \widehat{u}_{\mathbf{j|i}}$, which represents the agreement measure of capsule i and capsule j. Along with $\mathbf{W_{ij}}$, this process dictates the network's learning through every iteration. The output vectors $\mathbf{v}_k$, $1 \le k \le K$ (where K is the number of classes) from the last layer will have their magnitudes calculated, afterwards the highest length vector would correspond to the predicted class.

The loss function to be used as a baseline in this paper is from Sabour et al.'s study [8]—the margin loss function:

$$L_k = T_k \max\left(0, m^+ - \left\|\mathbf{v}_k\right\|\right)^2 + \lambda\left(1 - T_k\right)\max\left(0, \left\|\mathbf{v}_k\right\| - m^-\right)^2, \tag{4}$$

where $T_k = 1$ if the corresponding class k is present, $m^+ = 0.9$, $m^- = 0.1$, and the down-weighting parameter λ for the absent class is 0.5. In addition, L_k will be added onto a reconstruction loss scaled by a factor of 0.0005.

Within the past few years, other studies in the field of speech processing have incorporated the use of capsule-inspired networks. For instance, Lee et al. [21] made use of a CapsNet-only architecture for a sequence-to-sequence speech recognition task. The input sequence was sliced into windows then classified through the same dynamic routing mechanism. The margin loss was replaced by the computation of connectionist temporal classification (CTC). In another paper, Poncelet et al. [10] used capsule networks with recurrent neural networks, additionally encoding time information—an essential property present in speech. They applied this approach in the field of spoken language understanding (SLU). The main focus of this paper, speech emotion recognition, has also received some developments

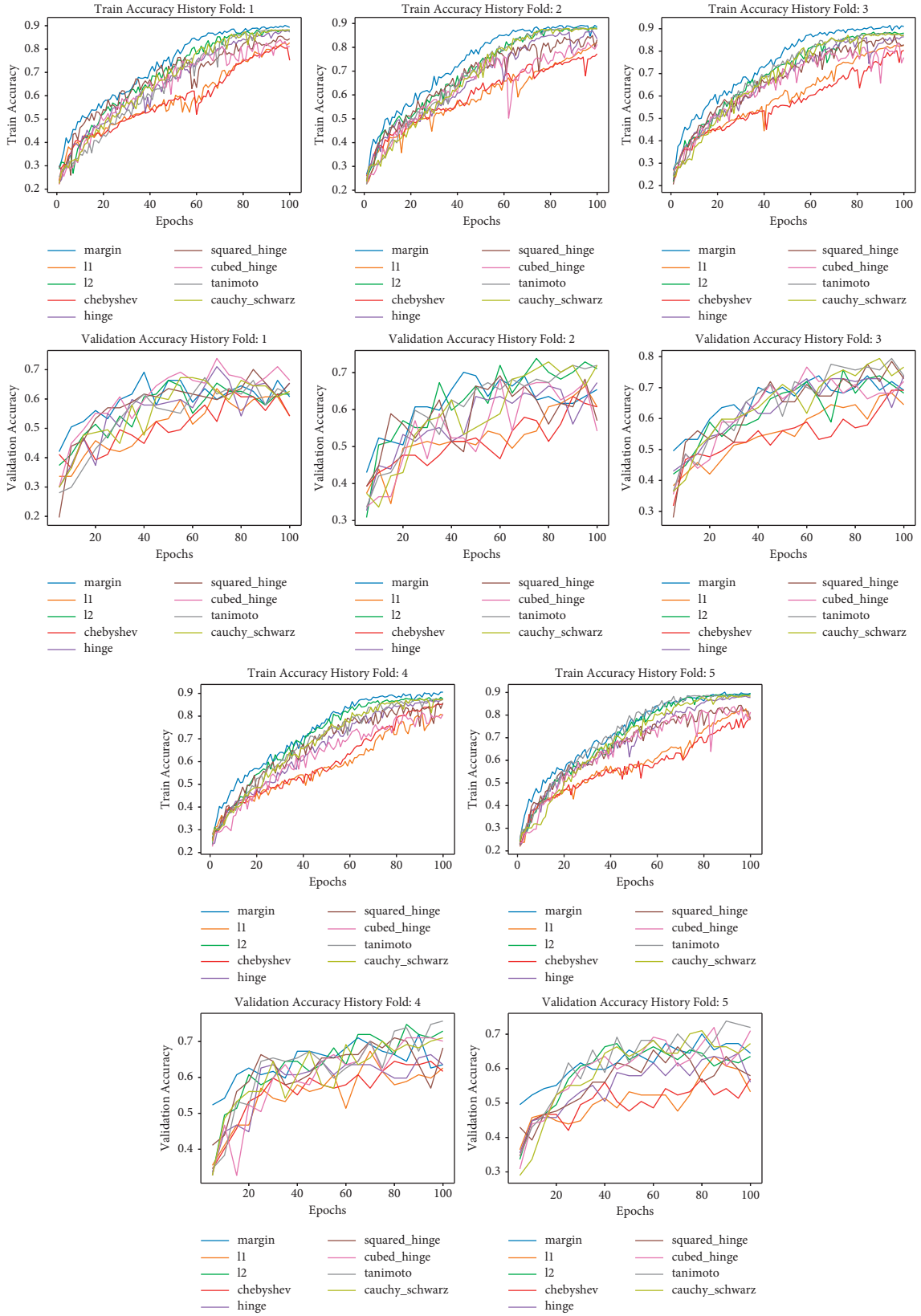

Figure 3: EMODB train and validation accuracy history for capsule model and different loss functions.

Table 3: EMODB class F1 scores and accuracies.

Loss type	Anger (F1)	Boredom (F1)	Disgust (F1)	Fear (F1)	Happiness (F1)	Sadness (F1)	Neutral (F1)	Overall (F1)	Overall (Acc.)
Margin loss	**83.85**	**72.49**	**61.04**	71.28	58.08	**84.32**	67.40	**71.21**	**73.46**
L1 loss	76.80	67.11	51.92	69.34	39.66	74.50	64.41	63.39	67.10
L2 loss	81.58	69.67	58.68	**71.61**	54.58	81.11	67.07	69.18	71.59
Chebyshev loss	74.74	65.47	45.73	64.80	34.40	68.74	60.60	59.21	64.30
Hinge loss	81.23	65.45	52.78	64.31	55.55	73.64	64.45	65.35	68.04
Square hinge loss	79.75	66.35	58.90	66.56	57.79	75.01	62.23	66.66	68.60
Cubed hinge loss	78.20	68.00	55.66	70.64	57.23	73.90	64.19	66.83	68.78
Tanimoto loss	81.09	64.60	58.19	67.44	**59.22**	79.94	64.35	67.83	69.91
Cauchy–Schwarz loss	82.48	66.51	57.09	69.00	57.71	80.98	**69.05**	68.97	71.61

with the use of capsule networks. These researches mainly use time-frequency spectrograms as their features. Wu et al. [22] improved the capsule network's performance by adding recurrent connections that can provide the network better feature modelling in the temporal dimension. Wu et al. [22] instead opted for MFCC features as the input for their capsule-based architecture. The capsule network used in this paper is identical to the one proposed by Sabour et al. [8]. Wu et al. [22] and Jain [23] also chose this configuration as well; however, they have added some modifications such as LSTMs and GRUs, further bolstering the feature extraction for the capsule network. This paper instead focuses on the impact of loss functions with the use of a capsule network.

2.3. Loss Functions. The loss functions to be compared in conjunction with the capsule network are listed in Table 1. Also worth noting is that output vectors $\mathbf{v}_k$ have to go through an extra step in order to be more suitable for these loss functions. The output q is calculated from equation (5).

$$q_k = \frac{\|\mathbf{v}_k\|}{\sum_n \|\mathbf{v_n}\|} \ . \tag{5}$$

L1 and L2 losses are primarily used in regression tasks. Both of these losses are used to complement the primary loss in other classification tasks as a form of regularization. Theoretically speaking, L1 loss is less sensitive to outliers than L2 loss.

The Chebyshev loss is characterized by taking the maximum absolute distance of one of the components between two vectors. Using Chebyshev loss this way would mean that in some cases, even if the model correctly classifies a sample, it may still be heavily penalized if even one component dramatically differs.

Also known as "maximum-margin" loss, hinge loss attempts to maximize the decision boundary between the groups being discriminated against. This type of loss has its origins in support vector machines (SVMs). The squared and cubed variants make the graph smoother and overgrow when the loss gets too big while making errors closer to zero weigh less on optimization.

Tanimoto and Cauchy–Schwarz divergence losses are relatively rarely used in deep learning tasks. The former is similar to Jaccard distance. It measures dissimilarity between two sampled sets by taking the ratio of the intersection over union among the individual values in the compared vectors. The latter also measures the distance between two random vectors and is an approximation to the Kullback–Leibler divergence [15].

3. Experimental Setup

Four data sets were used to perform the comparison experiments. The first data set is the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [24]. Only the 1,440 speech samples were used in this experiment, spanning across eight emotional classes: calm, happy, sad, angry, fearful, surprise, disgust, and neutral expressions. Each class is equally represented in the database except for the neutral class, which has 96 samples. The rest of the classes each have 192 samples. The database consists of 24 professional actors speaking in a neutral North American accent.

The second data set is the Berlin Emotional Database (EMODB) [25]. It has 535 utterances produced by ten actors (five female and five male) across seven different emotions: neutral, anger, fear, joy, sadness, disgust, and boredom. This data set is quite imbalanced as the difference between the number of samples of the largest and smallest classes is 81, which is alarmingly large for a small data set. The largest class is anger, while disgust was the smallest class.

The third data set is the Canadian French Emotional (CAFE) [26] speech data set with 936 utterances. The data set contains six different sentences, pronounced by 12 actors between two genders. Six basic emotions plus one neutral emotion are represented in the data set. Each class is equally represented except for the neutral emotion, half of one of the other emotions in the data set. The represented emotions are anger, disgust, happiness, fear, surprise, sadness, and a neutral state.

The last data set is the Sharif Emotional Speech Database (SHEMO) [27]. It contains 3000 Persian seminatural utterances extracted from online radio plays. Five emotions plus an extra neutral emotion are included in the data set. These emotions are anger, fear, happiness, sadness, surprise, and a neutral state. Similar to the second data set, a significant difference divides majority and minority classes of around 1000 samples. Anger and neutral emotions have over 1000 samples, while the other emotions have a few hundred samples.

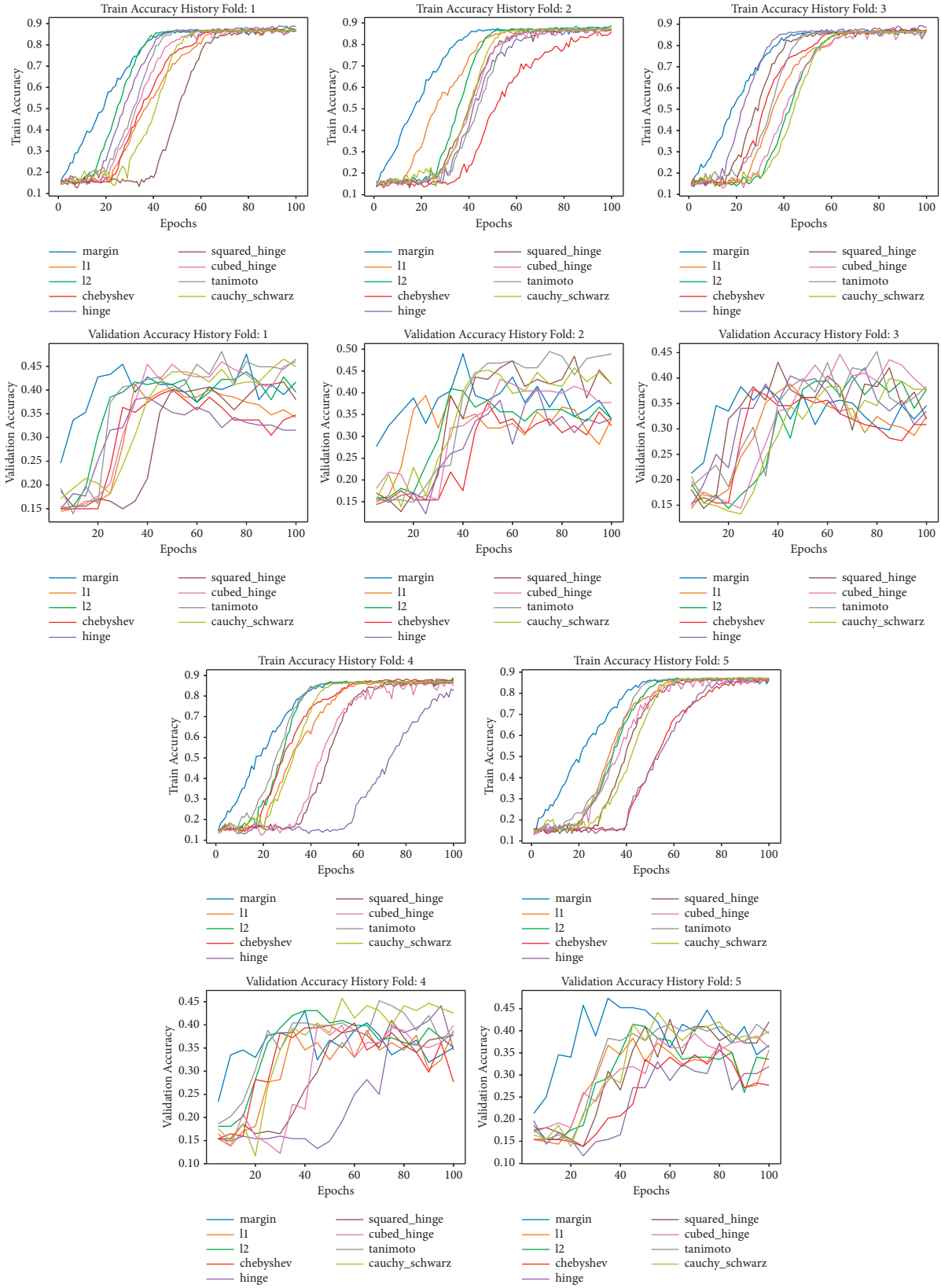

Figure 4: CAFE train and validation accuracy history for capsule model and different loss functions.

Table 4: CAFE class F1 scores and overall accuracies.

Loss type	Anger (F1)	Disgust (F1)	Happiness (F1)	Neutral (F1)	Fear (F1)	Surprise (F1)	Sadness (F1)	Overall (F1)	Overall (Acc.)
Margin loss	45.07	42.57	**42.61**	14.44	47.72	51.38	**52.20**	42.28	45.84
L1 loss	37.88	41.68	33.92	16.83	36.82	50.57	46.28	37.71	40.78
L2 loss	39.80	39.82	39.96	29.49	40.64	49.84	43.27	40.40	41.67
Chebyshev loss	41.64	38.83	24.90	5.22	35.01	44.70	48.73	34.15	38.57
Hinge loss	41.83	37.42	30.89	22.62	41.92	49.03	43.38	38.16	40.92
Square hinge loss	46.85	**45.35**	37.52	24.89	44.75	49.22	47.08	42.24	44.66
Cubed hinge loss	45.54	42.32	34.47	31.82	40.81	45.09	47.09	41.02	42.20
Tanimoto loss	49.57	43.06	38.06	31.07	46.89	**54.83**	48.28	44.54	46.36
Cauchy–Schwarz loss	**49.58**	43.53	35.72	**42.03**	**48.79**	52.46	51.82	**46.28**	**47.01**

The configuration used for the capsule network used in this paper is exactly described by Sabour et al. [8] and is shown in Figure 1. An initial convolution layer with 256 filters of size 9 and stride 1 extracts features from the image inputs. After the initial CNN layer, a PrimaryCaps layer with 256 channels from 32 8-dimensional capsules of size 9 and stride 2 follows. The last layer will differ in the number of capsules based on the number of unique classes in the data set. Each capsule in this last layer has 16 dimensions. The Adam optimizer is used with a learning rate of 0.001 and betas equal to 0.9 and 0.999, respectively. A decoder is also used to add in a reconstruction loss as a regularization term. The baseline model uses the margin loss described in the original literature.

In contrast, the other comparative models will use the other loss functions, with the rest of the architecture staying the same. Since the capsule architecture works best with image inputs, the input sequence for the network are time-frequency spectrograms extracted from the speech samples. Each spectrogram is a 64×64 image, unlike the 28×28 images from the MNIST data set. The data sets were divided into a $2:1:1$ split with the larger split for the training set and the other two splits for the validation and test sets. The models were cross-validated on five-folds for 100 epochs with a validation step every five epochs. After the training stage in each fold, the highest validation accuracy model would be used for the test set. For the training sets, some data augmentation, such as noise injection and voice tract length perturbation (VTLP).

4. Results and Discussion

In each data set, the training and validation accuracies are logged and graphed in the course of 100 epochs. In addition, the F1 scores for each emotion class and overall accuracies are shown in the tables below.

4.1. The Analysis for Different Loss Functions. For the first data set RAVDESS, a few remarks can be observed from the data in Figure 2 and Table 2 regarding the loss functions. The original margin loss remains the fastest in learning among the loss functions reaching more than 80% train accuracy at around 40 epochs. L2 loss also seems to be a considerable choice for a faster learning speed but with a less significant difference from the following loss function. The Cauchy–Schwarz divergence loss function learns slowly but lessens overfitting as observed on the validation accuracy histories. The Cauchy–Schwarz divergence and Tanimoto losses are the top two loss functions on F1 and accuracy. Both loss functions greatly improved on the baseline for almost all the individual classes, including the minority neutral emotion. The reason for this might be that these two loss functions consider the similarity of the compared vectors from the perspective of set theory. Unsurprisingly, these same two loss functions also perform pretty well in Janocha and Czarnecki's study [15]. Also mentioned by Janocha and Czarnecki [15] is that Cauchy–Schwarz divergence performs as well as cross-entropy loss or log loss in terms of learning speed and final performance.

Two loss functions performed the worst in EMODB. As shown in Figure 3 and Table 3, they are the L1 loss and Chebyshev loss. For samples that have been classified as correct, the individual elements of the target and predicted vectors might still be considerably different, which will still lead to a massive penalty during optimization. The penalty is amplified even further when using Chebyshev loss as even a correct classification may still lead to a higher loss. Out of the four data sets, only EMODB produced results where the baseline, margin loss, remained the best. One major cause for this result is the lack of sufficient samples in EMODB. Even with data augmentation, the newly generated samples may still resemble the original audio sample.

As shown in Figure 4 and Table 4, margin loss remains the fastest among the loss functions on the CAFE data set. Owing to the values of m^+ and m^- being specifically chosen for the capsule network after rigorous experimentation by the original authors, it is not a surprise that the loss function would be highly optimized. The validation accuracy histories of the different loss functions present constant shifting, which means that model can no longer improve on the validation set. The constant shift can easily be an easy sign of overfitting and a signal for early stopping. In terms of accuracy, the two best loss functions are still Tanimoto and Cauchy–Schwarz, albeit with a less significant lead on the baseline. Among the maximum-margin based losses, only squared hinge was able to perform as well as the baseline. It also did the best on the minority class, which is disgust. Perhaps the order of this hinge loss function is just in the right spot to not amplify significant errors and minimize minor errors.

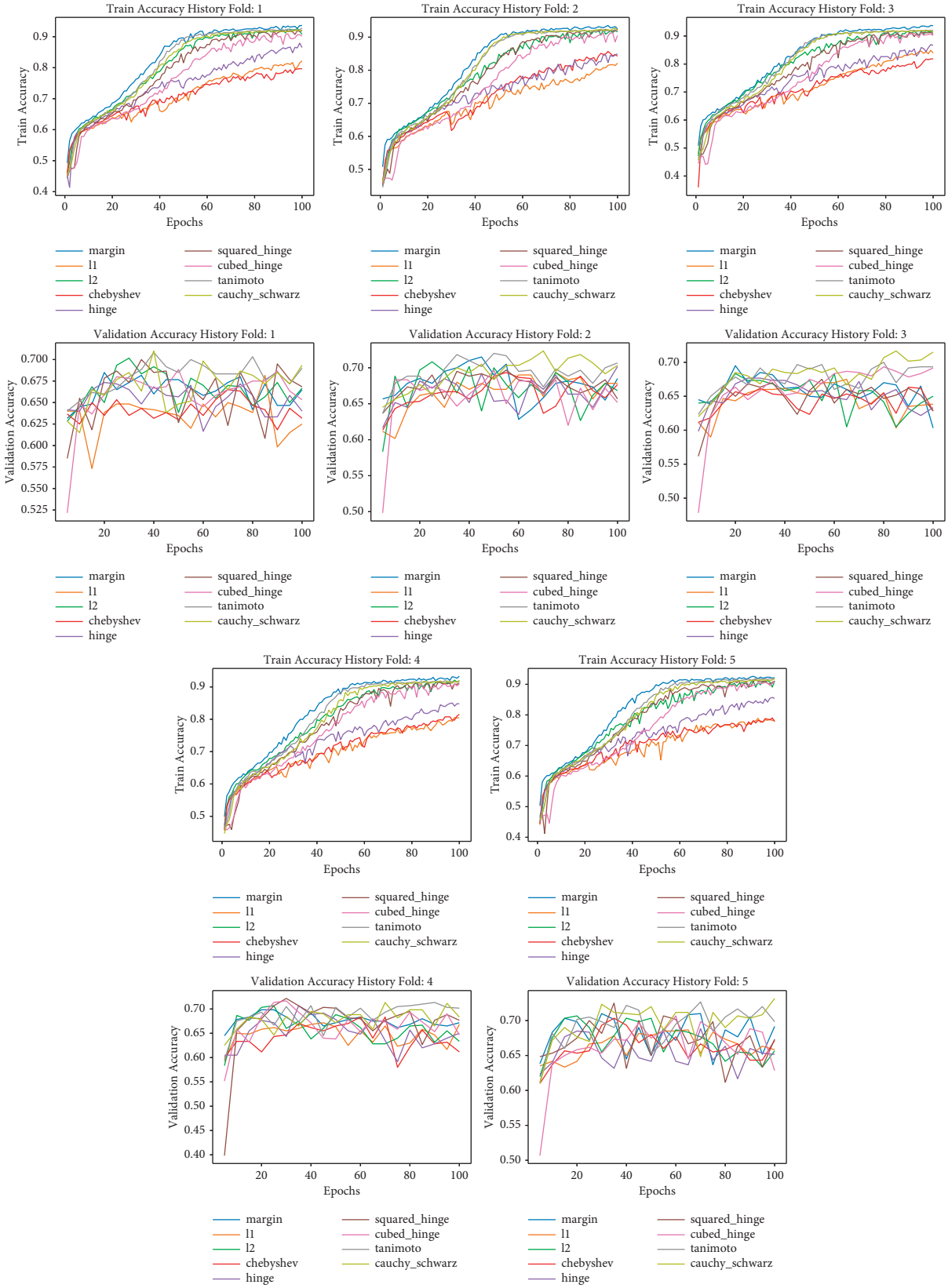

Figure 5: SHEMO train and validation accuracy history for capsule model and different loss functions.

Table 5: SHEMO class F1 scores and overall accuracies.

Loss type	Anger (F1)	Fear (F1)	Happiness (F1)	Neutral (F1)	Sadness (F1)	Surprise (F1)	Overall (F1)	Overall (Acc.)
Margin loss	79.60	0.00	14.65	76.69	50.99	52.03	45.66	69.70
L1 loss	78.07	0.00	4.34	74.67	53.02	32.11	40.37	68.10
L2 loss	80.37	0.00	17.69	77.45	55.39	50.89	46.97	70.43
Chebyshev loss	77.43	0.00	0.95	74.71	50.28	45.09	41.41	67.63
Hinge loss	78.58	0.00	3.33	76.25	52.61	48.13	43.15	69.27
Square hinge loss	80.58	4.00	**26.82**	76.96	49.04	51.93	48.22	70.17
Cubed hinge loss	78.90	0.00	19.21	78.06	52.18	47.23	45.93	69.87
Tanimoto loss	**81.15**	0.00	26.73	**78.47**	**56.52**	51.71	49.10	**71.43**
Cauchy–Schwarz loss	80.55	**4.44**	25.42	77.80	54.52	**57.26**	**50.00**	71.06

Table 6: Comparison with previous works (unweighted accuracies).

Model	Data set			
	RAVDESS (%)	EMODB (%)	CAFE (%)	SHEMO (%)
Capsule	69.03	**73.46**	47.01	**71.43**
CNN-BiGRU [7]	**70.07**	66.92	44.13	67.47
Head fusion [28]	57.85	68.04	41.45	70.60
LSTM [29]	68.19	71.59	**48.18**	69.77

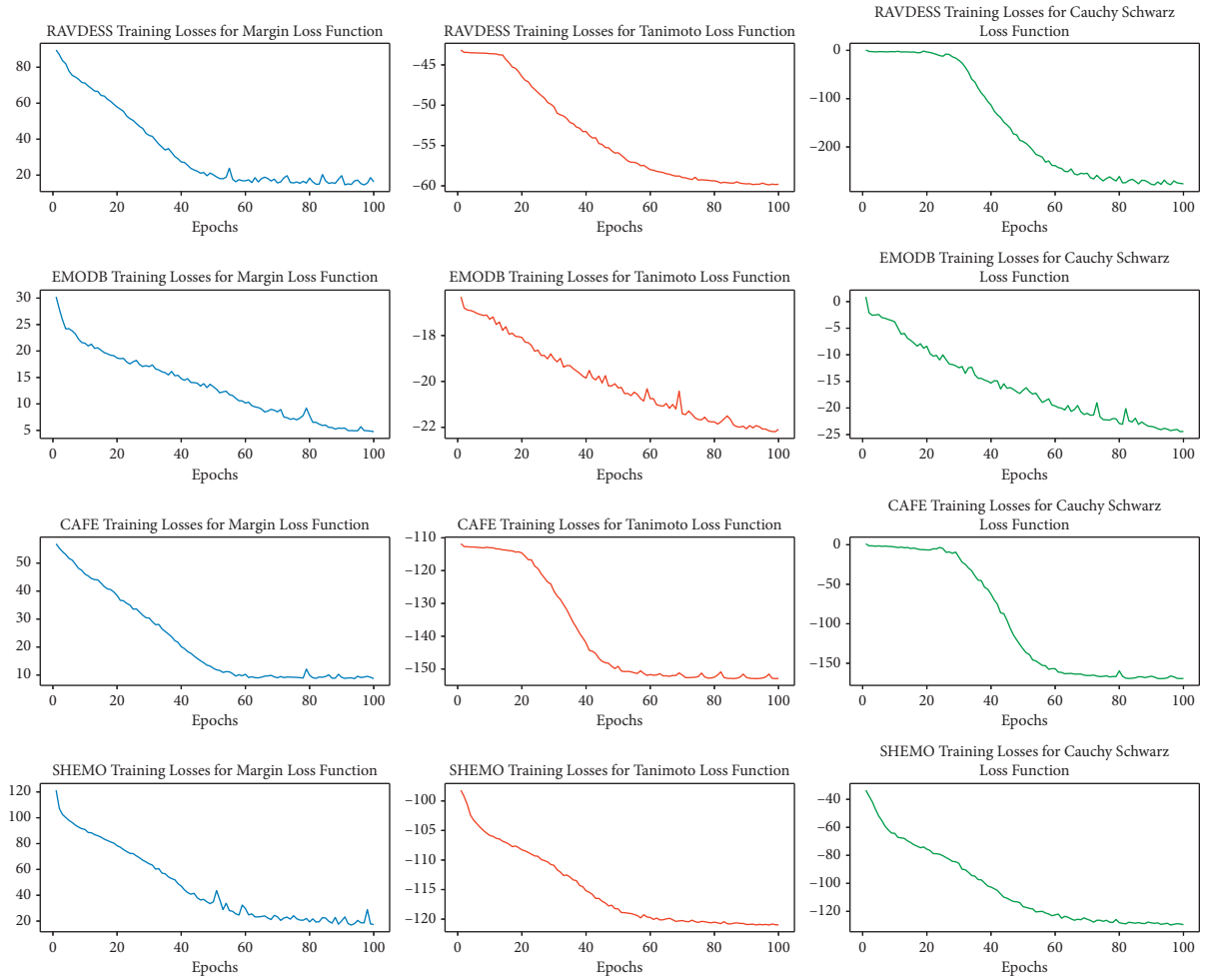

Figure 6: Training losses in a single fold for margin, Tanimoto, and Cauchy–Schwarz loss functions.

Figure 5 and Table 5 show the results for the SHEMO data set. The first thing that is relatively clear from Table 5 is low scores under the fear class with only 38 samples. Despite that the model was able to achieve an accuracy of 71% with the Tanimoto loss. Both Tanimoto and Cauchy-Schwarz divergence losses once again performed the best. Significant improvements were observed in the minority classes, such as fear, happiness, sadness, and surprise. If measures were to be taken to address the imbalance problem, the accuracy might increase, but the effect of these two losses might be less significant instead.

Finally, three more baseline models are implemented from other works for comparison. The first model is a combination of a CNN and a bidirectional gated recurrent unit network (BiGRU) model with focal loss function proposed by Zhu et al. [7]. In this model, the spectrogram features are passed through the CNN, after their temporal properties are analysed by the BiGRU. Next, the second model is a CNN model with a custom attention mechanism called head fusion [28], which is based on multihead attention. Finally, the third model is an LSTM model with a regular attention mechanism as described by Xie et al. [29]. All the baseline models use the same set of features as the capsule network. As shown in Table 6, the best capsule network accuracy is taken and compared with the previous works. Across the data sets, the capsule network performs as well as an LSTM especially on the EMODB data set. The ability of the capsule to encode spatial information would most likely complement well with an LSTM's affinity for encoding temporal information. The combination of both can be a good new research direction to consider. Another mechanism to consider is an attention mechanism, but its addition can be highly redundant to the dynamic routing.

4.2. Convergence Analysis for Tanimoto, Cauchy–Schwarz, and Margin Loss. To provide a better understanding for the reason of the Tanimoto and Cauchy–Schwarz loss functions' better performance, the training losses (in a single fold) for each type of loss are plotted as shown in Figure 6. It is clear in the RAVDESS data set that Tanimoto and Cauchy–Schwarz perform better because they converge a bit later than margin loss. On other data sets, the performances of Tanimoto and Cauchy–Schwarz in comparison with Margin loss are relatively similar; hence, they have similar curves and converge at roughly similar times. One thing to also note is that Tanimoto and Cauchy–Schwarz on both RAVDESS and CAFE data sets do not immediately have lowering losses within the first 20 epochs. This may mean that these loss functions are taking their time in learning in the initial portion of training.

5. Conclusion

This paper analyses the use of a capsule network and several different loss functions on SER data sets. Results showed that Tanimoto and Cauchy–Schwarz losses can highly improve capsule network performance by improving on the minority classes. Comparisons of the capsule network with previous deep learning models in the field also show that the capsule network performs marginally better. Future research directions will experiment on the use of capsule networks combined with LSTMs to use both their capabilities in learning spatial and temporal information, respectively.

Data Availability

The data are available at https://zenodo.org/record/1188976# (RAVDESS), https://zenodo.org/record/1478765# (CAFE), and https://github.com/mansourehk/ShEMO (ShEMO).

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (grant no. 61772023) and the National Key Research and Development Program of China (grant no. 2019QY1803).

References

[1] N. Fragopanagos and J. G. Taylor, "Emotion recognition in human-computer interaction," *Neural Networks*, vol. 18, no. 4, pp. 389–405, 2005.

[2] K. Vicsi and D. Sztahó, "Emotional state recognition in customer service dialogues through telephone line," in *Proceedings of the 2011 2nd International Conference on Cognitive Infocommunications (CogInfoCom)*, Budapest, Hungary, 2011.

[3] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[4] D. Issa, M. Fatih Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks," *Biomedical Signal Processing and Control*, vol. 59, Article ID 101894, 2020.

[5] D. Li, Y. Zhou, Z. Wang, and D. Gao, "Exploiting the potentialities of features for speech emotion recognition," *Information Sciences*, vol. 548, pp. 328–343, 2021.

[6] B. T. Atmaja and M. Akagi, "Speech emotion recognition based on speech segment using LSTM with attention model," in *Proceedings of the 2019 IEEE International Conference on Signals and Systems (ICSigSys)*.

[7] Z. Zhu, W. Dai, Y. Hu, and J. Li, "Speech emotion recognition model based on Bi-GRU and focal loss," *Pattern Recognition Letters*, vol. 140, pp. 358–365, 2020.

[8] S. Sabour, N. Frosst, and G. Hinton, "Dynamic routing between capsules," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, CA, USA, December 2017.

[9] H. Yao, Y. Tan, C. Xu, J. Yu, and X. Bai, "Deep capsule network for recognition and separation of fully overlapping handwritten digits," *Computers & Electrical Engineering*, vol. 91, Article ID 107028, 2021.

[10] J. Poncelet, V. Renkens, and H. Van hamme, "Low resource end-to-end spoken language understanding with capsule networks," *Computer Speech & Language*, vol. 66, Article ID 101142, 2020.

[11] P. Afshar et al., "COVID-CAPS: a capsule network-based framework for identification of COVID-19 cases from x-ray images," 2020, https://arxiv.org/abs/2004.02696.
[12] K. Lei, Q. Fu, and Y. Liang, "Multi-task learning with capsule networks," in *Proceedings of the 2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8.
[13] H. Yao, P. Zhang, C. Jiang, and Z. Han, "Capsule network assisted IoT traffic classification mechanism for smart cities," *IEEE Internet of Things Journal*, vol. 6, pp. 7515–7525, 2019.
[14] B. T. Atmaja and M. Akagi, "Evaluation of error- and correlation-based loss functions for multitask learning dimensional speech emotion recognition," *Journal of Physics: Conference Series*, vol. 1896, no. 1, Article ID 012004, 2021.
[15] K. Janocha and W. Czarnecki, "On loss functions for deep neural networks in classification," *Schedae Informaticae*, vol. 25, 2017.
[16] S. Langari, H. Marvi, and M. Zahedi, "Efficient speech emotion recognition using modified feature extraction," *Informatics in Medicine Unlocked*, vol. 20, Article ID 100424, 2020.
[17] L. Kerkeni, Serrestou, Raoof, M. Mbarki, A. A. Mahjoub, and C. Cleder, "Automatic speech emotion recognition using an optimal combination of features based on EMD-TKEO," *Speech Communication*, vol. 114, pp. 22–35, 2019.
[18] T. Özseven, "A novel feature selection method for speech emotion recognition," *Applied Acoustics*, vol. 146, pp. 320–326, 2019.
[19] Z. Yao, Y. Liu, and J. Pan, "Speech emotion recognition using fusion of three multi-task learning-based classifiers: HSF-DNN, MS-CNN and LLD-RNN," *Speech Communication*, vol. 120, pp. 11–19, 2020.
[20] J. Zhao, M. Mao, and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," *Biomedical Signal Processing and Control*, vol. 47, pp. 312–323, 2019.
[21] K. Lee, H. Joe, H. Lim et al., "Sequential routing framework: fully capsule network-based speech recognition," *Computer Speech & Language*, vol. 70, Article ID 101228, 2021.
[22] X. Wu, S. Liu, Y. Cao et al., "Speech emotion recognition using capsule networks," in *Proceedings of the ICASSP 2019—*, p. 5.
[23] R. Jain, "Improving performance and inference on audio classification tasks using capsule networks," 2019, https://arxiv.org/abs/1902.05069.
[24] S. R. Livingstone and F. A. Russo, "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*, vol. 13, no. 5, Article ID e0196391, 2018.
[25] F. Burkhardt, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proceedings of the INTERSPEECH 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology*, p. 4, liston, portugal.
[26] P. Gournay, O. Lahaie, and R. Lefebvre, "A canadian French emotional speech dataset," in *the 9th ACM Multimedia Systems Conference*, pp. 399–402, Amsterdam, Netherland, June 2018.
[27] O. Mohamad Nezami, P. Lou, and M. Karami, "ShEMO—a large-scale validated database for Persian speech emotion detection," *Language Resources and Evaluation*, vol. 53, 2019.
[28] M. Xu, "Head fusion: improving the accuracy and robustness of speech emotion recognition on the IEMOCAP and RAVDESS dataset," *IEEE Access*, vol. 9, pp. 74539–74549, 2021.
[29] Y. Xie, R. Liang, Z. Liang, C. Huang, C. Zou, and B. Schuller, "Speech emotion classification using attention-based LSTM," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 11, pp. 1675–1685, 2019.