Transductive Ensemble Learning with Graph Neural Network for Speech Emotion Recognition

Eliton L. Scardin Perin[1], Edson Takashi Matsubara[2]

[1]University Federal of Mato Grosso do Sul (UFMS), Cidade Universitária, Av. Costa e Silva - Pioneiros, MS, 79070-900
[2]University Federal of Mato Grosso do Sul (UFMS), Cidade Universitária, Av. Costa e Silva - Pioneiros, MS, 79070-900

Abstract

In this paper, we jointly use a Transductive Ensemble Learning with a Graph Convolutional Network to perform a task of Speech Emotion Recognition (SER). Additionally, we propose solving this task using ensemble learning methods with simple base classifiers such as Multilayer Perceptrons and k-Nearest Neighborhood. We extracted features using Wav2Vec and prosodic. The performance reaches 0.5248 in macro F1-score in the SER PROPOR 2022 dataset.

Keywords

Transductive Learning, Ensemble Learning, Graph Convolution Network

1. Introduction

The speech can express more than words; it can express feelings. With this objective in mind, we would like to identify sentiment, intention, opinion, genre, and humor using audio and language features [1, 2].

Discover the feeling present in the audio content is known as Speech Emotion Recognition (SER). Among the most known dataset for this task, the Portuguese are not among them. For this reason, the Internation Conference on the Computer Processing of Portuguese (PROPOR 2022) proposed a workshop to bring new researchers and enthusiasts to processing speech in Portuguese. One of the challenges of this conference is the SER for Portuguese.

Traditional methods of machine learning and deep learning techniques employ several solutions for SER with strong performances [3]. Ensemble methods are applied to this task, too, and improve the results for the base classifiers [3]. We use a method with a transductive ensemble learning to predict new labels with several outputs from base classifiers to capture the best characteristic from each method. Our method shows competitive results when compared with the baselines.

2. Methodology

The following sections describe the techniques of Ensemble Learning and Graph Neural Network.

Proceedings of the First Workshop on Automatic Speech Recognition for Spontaneous and Prepared Speech & Speech Emotion Recognition in Portuguese (SE&R 2022), co-located with PROPOR 2022. March 21st, 2022 (Online)
elitonperin@gmail.com (E. L. S. Perin); edsont@ufms.br (E. T. Matsubara)

All of them were inserted in our method Transductive Ensemble Learning with Graph Neural Network. After these techniques section, we introduced our proposal.

2.1. Ensemble Learning

An ensemble combines multiple base classifiers to obtain final predictions. A simple ensemble can be obtained performing majority voting or weighting the voting from base classifiers. Another type of ensemble is the technique that includes label ranking in the process of voting rules [4]. More advanced techniques focus on training metamodels from the output of the classifiers together with the training set.

In general, the combination of the base classifiers is a crucial step and is a non-trivial process in the presence of several base classifiers. It is common in the usage of ensemble methods to perform neural machine translation (NMT) [5, 6].

2.2. Graph Neural Network

Neural Networks with support for graph representation are known as Graph Neural Network (GNN). Originally proposed by [7, 8] in a recurrent neural network style. The insertion of a graph is helpful in other types of Artificial Neural Networks to perform similar tasks. An example of GCN usage for text classification tasks is the Text GCN; this method gets the corpus and builds a graph with word co-occurrence and documents word relation to categorize them. Besides, no word embeddings or external knowledge outperforms the state-of-the-art [9].

Node classification is another way to work with graphs and neural networks. This idea introduced by [10] with DCNNs (Diffusion-convolutional neural networks) shows interesting results. The work of [11] uses GCN to realize node classification in the semi-supervised way of training and outperforms several similar methods. A simplifying GCN (SGCN) published by [12] explore the power of the GCN to scale and do not negatively impact the accuracy, and SGCN was better 100 times against the FastGCN [13].

2.3. Transductive Ensemble Learning with Graph Convolution Network

In this work, we propose Ensemble Learning with Transductive Learning for SER problem. Transductive learning [14] is similar to semi-supervised learning, but it considers a search space of hypothesis where all testing data is known beforehand. Our proposal of ensemble learning runs over a Graph Convolutional Network (GCN). We configure a graph with the training set that carries the relation between samples, labels, and models.

In a GCN, we build the graph that connects all data samples with all the labels for the respective models of classifiers. Figure 1 shows the representation of this bipartite graph. One set of bipartite nodes represents the samples. The *tr_0* is a node representing the first example in the training dataset, and nodes below until *tr_i* are the $i-th$ last one in the training set. These nodes receive a label from training. The *te_0* is a node for the first sample in the testing set. The following nodes till *te_j* belong to the $j-th$ example in the test dataset. We set nodes from the testing set with an invalid label not used in the training set.

The model nodes are another set of bipartite nodes. The nodes $m1c0$, $m1c1$ e $m1c2$, represent the model 1 for each valid label, and this setting repeat for the others model nodes $m2c0$ until

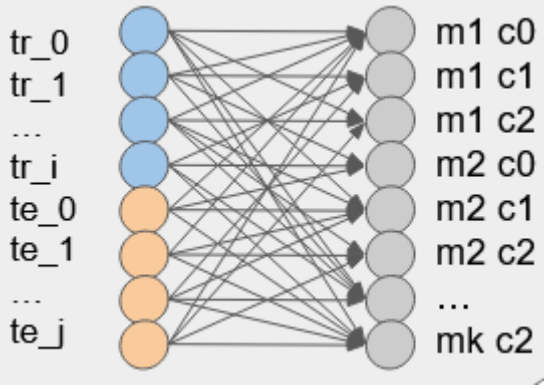

Figure 1: The input graph for GCN.

the *mkc*2, the $k - th$ model. The edges between model and samples nodes are the probability of the output from models to respective samples.

The representation of the Graph Convolution Network is an adjacency matrix. Rows and columns represent nodes of the graph. If a node has a connection, the value in the matrix cell represents the probability that the node of instance belongs to the node class. The matrix of adjacency is normalized and becomes a piece of the function of propagation of data [11]:

$$H^{[i+1]} = \sigma(W^{[i]}H^{[i]}A^{*}) \tag{1}$$

where $H^{[i]}$ is the vector of input features to layer i, W is the weights of layer i of the GCN, σ is the activation function, and A^{*} is the normalized adjacency matrix. From the adjacency matrix is extracted the identity matrix that is used as network input.

The Figure 2 shows a diagram of a GCN with 4 layers, with respective sizes 512, 256, 128 and $c + 1$, where c is the number of classes. Note that the adjacency matrix is shared with all layers. The network has $c + 1$ outputs, c are for the class of the problem, and one more for the nodes that do not have a class. The proposal avoids examples without defined classes using masks to not participate in the CNN backpropagation algorithm.

The following steps of the algorithm are the same as the CNN. The problem is a classification with the CNN, where it propagates labels from training set to unlabeled set. The training uses the function and applies it to the backpropagation to adjust the weights with the respective derivatives.

3. Results

We used the dataset available by the competition for the SER task for the experiments. They are audios from the dataset called CORAA (Corpus of Annotated Audios). The distribution of examples and classes is shown in Table 1. We used two types of audio features as input to the base models. Both types were made available by the competition. The first type is the Wav2Vec, an unsupervised pre-training model for speech recognition [15]. The second is prosodic, which

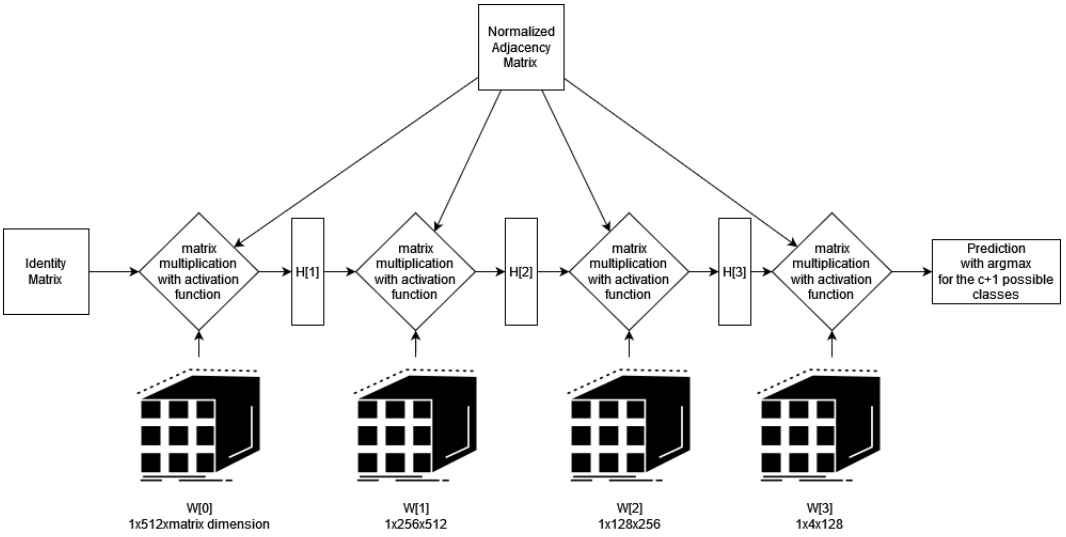

Figure 2: Representation of a GCN.

extracts features from the intonations and energy curves of audios [16]. The base models chosen for this task are:

- **Multilayer Perceptron (MLP):** is an artificial neural network with few layers. It is a supervised learner, able to build a universal function to approximate [17, 18].
- **k-Nearest Neighborhood (kNN):** a classifier that performs the predictions based on voting across the metric of distance between train and test samples [19].

Classes	Training set	Testing set
neutral	491	-
non-neutral-female	89	-
non-neutral-male	45	-
Total	625	308

Table 1
Distribution of dataset

Table 2 shows the name, parameters, features, and performance of models adopted by ensemble learning. To choose the parameters, we run cross-validation to get the best performance for each model. The best model for the training set was the MLP with Wav2Vec features; this model was employed as the baseline Wav2Vec in the competition.

Table 3 presents the performance final of the Transductive Ensemble Learning with GCN. To compare the results, we added the baseline Wav2Vec and prosodic given by the competition. The test set is closed, and we can not submit more than three sets of predictions. So, the experiments were split with a small validation set to analyze the performance of models. The best parameter for these experiments with a small validation set was a learning rate with 0.001, for epochs

Name	parameters	features	F1-score macro with cross-validation with train set
MLP	activation=logistic, iterations=3000	Wav2Vec	0.5544
MLP	activation=tanh, iterations=3000	Wav2Vec	0.5553
MLP	activation=relu, iterations=3000	Wav2Vec	0.5418
kNN	k=1	Wav2Vec	0.5258
kNN	k=3	Wav2Vec	0.4474
MLP	activation=logistic, iterations=3000	prosodic	0.5371
MLP	activation=tanh, iterations=3000	prosodic	0.5088
MLP	activation=relu, iterations=3000	prosodic	0.4873
kNN	k=1	prosodic	0.4523
kNN	k=3	prosodic	0.4635

Table 2
Metrics for base classifiers.

equal to 1000 steps, and layers with size 512, 256, 128 e 4 number of neurons from beginning to the end of GCN. A graphical representation is shown in Figure 2.

Method	Size of training set	Size of validation set	Accuracy on validation set	F1-score macro on validation set	F1-score macro on real test set
Transductive Ensemble Learning with GCN	615	10	0.8000	0.2962	0.5248
Transductive Ensemble Learning with GCN	500	125	0.7952	0.4532	-
Transductive Ensemble Learning with GCN	625	0	-	-	-
Baseline Wav2Vec	625	0	-	-	0.5356
Baseline prosodic	625	0	-	-	0.5335

Table 3
Show the metrics for the method of Transductive Ensemble Learning with GCN.

4. Conclusion

The Transductive Ensemble Learning with GCN is a competitive method for the task of SER. The results do not exceed baseline metrics of Wav2Vec or prosodic, but we do not use any new feature for the network. We plan to use features from other models with fine-tuning for this dataset for future works. Apply new features into nodes as a piece of the learning data to improve the results of the method. We can combine other models of neural networks with attention mechanism.

Acknowledgments

This work has been supported by the following Brazilian research agencies: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) and Fundação de Apoio ao Desenvolvimento de Ensino, Ciência e Tecnologia do MS (FUNDECT).

References

[1] D. Bertero, P. Fung, Deep learning of audio and language features for humor prediction, in: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), European Language Resources Association (ELRA), Portorož, Slovenia, 2016, pp. 496–501. URL: https://aclanthology.org/L16-1079.

[2] D. Chathuranga, L. Jayaratne, Automatic music genre classification of audio signals with machine learning approaches, GSTF Journal on Computing (JoC) 3 (2013) 1–12.

[3] M. B. Akçay, K. Oğuz, Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers, Speech Communication 116 (2020) 56–76.

[4] H. Werbin-Ofir, L. Dery, E. Shmueli, Beyond majority: Label ranking ensembles based on voting rules, Expert Systems with Applications 136 (2019) 50–61.

[5] E. Garmash, C. Monz, Ensemble learning for multi-source neural machine translation, in: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, The COLING 2016 Organizing Committee, Osaka, Japan, 2016, pp. 1409–1418. URL: https://aclanthology.org/C16-1133.

[6] Y. Wang, L. Wu, Y. Xia, T. Qin, C. Zhai, T.-Y. Liu, Transductive ensemble learning for neural machine translation, Proceedings of the AAAI Conference on Artificial Intelligence 34 (2020) 6291–6298. URL: https://ojs.aaai.org/index.php/AAAI/article/view/6097. doi:10.1609/aaai.v34i04.6097.

[7] M. Gori, G. Monfardini, F. Scarselli, A new model for learning in graph domains, in: Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005., volume 2, 2005, pp. 729–734 vol. 2. doi:10.1109/IJCNN.2005.1555942.

[8] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, G. Monfardini, The graph neural network model, IEEE Transactions on Neural Networks 20 (2009) 61–80. doi:10.1109/TNN.2008.2005605.

[9] L. Yao, C. Mao, Y. Luo, Graph convolutional networks for text classification, Proceedings of the AAAI Conference on Artificial Intelligence 33 (2019) 7370–7377. URL: https://ojs.aaai.org/index.php/AAAI/article/view/4725. doi:10.1609/aaai.v33i01.33017370.

[10] J. Atwood, D. Towsley, Diffusion-convolutional neural networks, in: D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 29, Curran Associates, Inc., 2016. URL: https://proceedings.neurips.cc/paper/2016/file/390e982518a50e280d8e2b535462ec1f-Paper.pdf.

[11] T. N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, in: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings, OpenReview.net, 2017. URL: https://openreview.net/forum?id=SJU4ayYgl.

[12] F. Wu, A. Souza, T. Zhang, C. Fifty, T. Yu, K. Weinberger, Simplifying graph convolutional networks, in: K. Chaudhuri, R. Salakhutdinov (Eds.), Proceedings of the 36th International Conference on Machine Learning, volume 97 of *Proceedings of Machine Learning Research*, PMLR, 2019, pp. 6861–6871. URL: https://proceedings.mlr.press/v97/wu19e.html.

[13] J. Chen, T. Ma, C. Xiao, FastGCN: Fast learning with graph convolutional networks via importance sampling, in: International Conference on Learning Representations, 2018.

URL: https://openreview.net/forum?id=rytstxWAW.

[14] A. Gammerman, V. Vovk, V. Vapnik, Learning by transduction, in: Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, UAI'98, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1998, p. 148–155.

[15] S. Schneider, A. Baevski, R. Collobert, M. Auli, wav2vec: Unsupervised pre-training for speech recognition, arXiv preprint arXiv:1904.05862 (2019).

[16] I. Luengo, E. Navas, I. Hernáez, J. Sánchez, Automatic emotion recognition using prosodic parameters, in: Ninth European conference on speech communication and technology, Citeseer, 2005.

[17] G. E. Hinton, Connectionist learning procedures, in: Machine learning, Elsevier, 1990, pp. 555–610.

[18] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, in: Proceedings of the thirteenth international conference on artificial intelligence and statistics, JMLR Workshop and Conference Proceedings, 2010, pp. 249–256.

[19] E. Fix, J. L. Hodges, Discriminatory analysis. nonparametric discrimination: Consistency properties, International Statistical Review/Revue Internationale de Statistique 57 (1989) 238–247.