Modular Training of Deep Neural Networks for Text Classification in Guarani

José Luis Vázquez Noguera
Centro de Investigación
Universidad Americana
Asunción, Paraguay
jose.vazquez@ua.edu.py

Carlos U. Valdez
Facultad de Ciencias y Tecnología
Universidad Autónoma de Asunción
Asunción, Paraguay
cavaldez@uaa.edu.py

Julio César Mello-Román
Facultad Politécnica
Universidad Nacional de Asunción
San Lorenzo, Paraguay
juliomello@pol.una.py

Marvin M. Aguero
Facultad de Ciencias y Tecnología
Universidad Autónoma de Asunción
Asunción, Paraguay
marvin-aguero@outlook.com

José D. Colbes
Facultad Politécnica
Universidad Nacional de Asunción
San Lorenzo, Paraguay
jcolbes@pol.una.py

Sebastián A. Grillo
Facultad de Ciencias y Tecnología
Universidad Autónoma de Asunción
Asunción, Paraguay
sgrillo@uaa.edu.py

Abstract

We present a modular training approach for deep text classification in Guarani, where networks are split into sectors trained independently and later combined. This sector-wise backpropagation improves stability, reduces training time, and adapts to standard architectures like CNNs, LSTMs, and Transformers. Evaluated on three Guarani datasets—emotion, humor, and offensive language—our method outperforms traditional Bayesian-optimized training in both accuracy and efficiency.

1 Introduction

Natural language processing (NLP) for low-resource languages has gained attention due to the need for more inclusive technologies (Joshi et al., 2020). Guarani, an indigenous language spoken by over eight million people in Paraguay and neighboring countries, remains underrepresented in digital resources.It lacks open corpora, standard models, and suffers from frequent code-switching with Spanish (Estigarribia, 2016), which complicates data collection. These particularities of the Guarani language, coupled with the scarcity of labeled data and pretrained modules, make it challenging to train deep neural networks that generalize well to downstream tasks such as sentiment analysis, which are standard benchmark tasks (Mao et al., 2023) for high-resource languages like English.

Some efforts in low-resource NLP for Guarani have focused on corpus creation and benchmarking. Chiruzzo et al. (2020) expanded initial Guarani-Spanish sentence pairs into larger parallel collections, later unified and quality controlled as the `Jojajovai` corpus (Chiruzzo et al., 2022). Preliminary Guarani BERT (Devlin et al., 2019) variants (including continuous-pretrained and trained from scratch) have been trained on Wikipedia-derived texts containing only ∼800K tokens (Agüero-Torales et al., 2023), and Guarani was added to large multilingual initiatives such as 'No Language Left Behind' (NLLB Team et al., 2022) and `Google Translate` (Bapna et al., 2022). With regard to the text classification task, there are some works with diverse results, mainly for affective computing such as (i) (Agüero-Torales et al., 2023) explores various deep neural text classification techniques for multidimensional affective analysis; and (ii) sentiment analysis (Ríos et al., 2014), covering approaches that range from lexicon-based or traditional machine learning models (bag-of-words) to more sophisticated methods such as fine-tuning multilingual transformer models (Vaswani et al., 2017).

On the other hand, traditional text classification approaches in high-resource settings rely on end-to-end backpropagation over large corpora and big pretrained embeddings. When applied to Guarani, these methods tend to overfit quickly or fail to converge, since the number of tunable parameters far exceeds the available supervision. Recent work on low-resource NLP has mitigated these issues through transfer learning and cross-lingual embeddings (e.g. Schuster et al. (2019)), or adapting models trained in related languages or synthetic data (Lucas et al., 2024). However, these strategies remain monolithic: they update most network parameters at once, risking catastrophic forgetting of pretrained knowledge or uneven adaptation across layers (Kirkpatrick et al., 2017; Roy, 2024).

In parallel, modular and layer-wise training has been proposed in other domains (e.g. vision) to control the capacity of deep architectures (Tabrizi

Proceedings of the 1st Workshop on Advancing NLP for Low-Resource Languages associated with RANLP 2025, pages 76–81, Varna, Bulgaria, Sep 13, 2025.
https://doi.org/10.26615/978-954-452-100-4-008

et al., 2024). By isolating each layer (or 'sector') and optimizing its weights separately, these methods reduce the dimensionality of each learning step, reducing overfitting, and accelerating convergence (Belilovsky et al., 2020). However, to our knowledge, no prior work has applied a fully sector-wise backpropagation scheme to text classification in a truly low-resource language.

This work is based on the layer-wise loss assignments approach for layer-wise training (Belilovsky et al., 2020, 2019), which trains each layer using an auxiliary coupled model that can have several layers. Our approach decomposes a deep network into successive parameterized sectors, each trained as a shallow subnetwork on intermediate representations. We then recombine the trained sectors into a full model, preserving both pretrained knowledge and local adjustments. This sector-wise backpropagation delivers the following benefits:

- It constrains the number of parameters updated at each step, resulting in more stable training curves on small Guarani datasets.
- It preserves cross-sector knowledge transfer by propagating learned representations forward between stages.
- It consistently integrates with any architecture built from standard layers (e.g., convolutional (LeCun et al., 1989), Long Short-Term Memory (Hochreiter and Schmidhuber, 1997, LSTM) or transformers (Vaswani et al., 2017)), allowing the adaptation of existing models.

We validate our proposal on three Guarani corpora for affective computing (Agüero-Torales et al., 2023), namely: i) *gn-humor-detection*, ii) *gn-offensive-language-identification*, and iii) *gn-emotion-recognition*. In experiments, our sector-wise method outperforms conventional end-to-end training and standard baselines by significant margins. The remainder of this paper is structured as follows. Section 2 details our sector-wise optimization algorithm. Section 3 presents the experimental results, and Section 4 concludes our work.

2 Sector-wise Backpropagation

The modular optimization applied in this work is based on the concept of *sector*. A sector consists of a parameterized layer and all subsequent non-parameterized layers until the next parameterized layer. For example, in a network with architecture C_1–P_1–P_2–C_2–P_3–C_3 (where C_i are fully connected layers and P_i are pooling layers), the sectors would be:

$S_1 = C_1$–P_1–P_2, $S_2 = C_2$–P_3, $S_3 = C_3$.

Given a network D and a training set X, for each epoch, the method proceeds in three main steps:

1. For each sector S_i (excluding the last), construct a shallow network N_i composed of S_i and an output layer identical to that of D.
2. Train each N_i using transformed instances $f_{i-1}(x)$, where $f_0(x) = x$, and we define $f_i(x)$ as the output of N_i with its output layer removed.
3. Rebuild D by stacking the trained sectors and removing the auxiliary output layers, except for the final one.

In the earlier example, the auxiliary networks created would be:

$N_1 = C_1$–P_1–P_2–C_3', $N_2 = C_2$–P_3–C_3,

where C_3' replicates C_3. N_1 is trained on $x \in X$, and N_2 is trained on transformed outputs $f_1(x)$. Algorithm 1 formalises the proposal.

Algorithm 1 Sector-wise Local Backpropagation and Network Reconstruction

Initialize: Architecture D, sectors $S_1, \ldots, S_n$, null network R_0
while stop condition not met **do**
  Sector Backpropagation
  for $i = 1$ to $n - 1$ **do**
    Create N_i by adding a layer similar to the last layer of D on top of sector S_i
    Train N_i for one epoch using instances $f_{i-1}(x)$ for $x \in X$, with the same label as x
    Compute $f_i(x)$ for each $x \in X$ by evaluating the penultimate layer output of N_i
  end for
  Network Reconstruction
  $R_0 \leftarrow \varnothing$
  for $i = 1$ to $n - 2$ **do**
    Extract S_i from trained N_i preserving learned parameters
    Connect S_i to R_{i-1} according to D, forming R_i
  end for
  Connect R_{n-2} to N_{n-1} according to D, forming R_{i-1}
end while
return R_{i-1}

3 Results

Experiments were conducted on three datasets (over their train-dev-test splits): *gn-humor-detection* (fun and no-fun classes), *gn-offensive-language-identification* (offensive and no-offensive

classes), and *gn-emotion-recognition* (happy, angry, sad and other classes) (Agüero-Torales et al., 2023); using 10-fold cross-validation. Three model architectures were tested on each dataset: a **1D convolutional network** (Omernick and Chollet, 2019; Waibel et al., 1989), a **transformer-based model** (Nandan, 2020; Vaswani et al., 2017), and a **bidirectional LSTM** (Chollet, 2020; Schuster and Paliwal, 1997).

Each model was trained under three configurations: i) Standard backpropagation with fixed hyperparameters, ii) Backpropagation with Bayesian hyperparameter optimization and iii) Sector-based backpropagation (the proposed method).

For configurations 1 and 3, training was performed using a *batch size of 32*, *learning rate of 0.001*, the *Adam optimizer*, and *sparse categorical cross-entropy* loss. For configuration 2, Bayesian optimization was applied with the following domains: i) optimizer $\in$ {`adam`, `rmsprop`, `sgd`}, ii) learning rate $\in$ (1e-5, 1e-1) with a log-uniform distribution and iii) Batch size $\in$ [16, 128].

Table 1: Average accuracy on the gn-humour-detection dataset as the number of training epochs increases. Model 1 is a 1D ConvNet, model 2 is a Transformer, and model 3 is a Bidirectional LSTM.

Mod.	Epoch	Simp.	Bayes.	Prop.
1	2	**71.27**	**71.27**	70.27
1	4	69.92	70.38	**71.46**
1	6	70.19	69.95	**71.27**
1	8	65.58	68.99	**71.22**
1	10	66.12	68.78	**71.76**
2	2	71.27	71.27	**73.28**
2	4	71.27	71.25	**73.52**
2	6	71.82	71.27	**74.09**
2	8	62.33	71.27	**73.55**
2	10	59.62	71.27	**73.98**
3	2	64.54	64.66	**68.92**
3	4	64.85	66.02	**69.16**
3	6	58.27	65.39	**70.46**
3	8	63.04	65.18	**69.40**
3	10	64.23	65.15	**70.54**

Table 1 presents the corresponding results for the *gn-humor-detection* dataset. They are grouped according to the models (first column), considering different epochs (second column), followed by the average accuracy for each configuration. Considering each model, the transformer-based one achieved the highest accuracy among the others. More interestingly, our proposal obtained a better performance in nearly all cases (except for model 1 with 2 epochs). In terms of accuracy, the best configuration recorded (74.09%) is the transformer-based architecture when trained with the proposal for 6 epochs. Moreover, for the first two configurations, the average accuracy generally decreases slightly as the number of epochs increases. This behaviour does not appear in our proposal.

Table 2: Average accuracy on the gn-offensive-language-identification dataset as the number of training epochs increases. Model 1 is a 1D ConvNet, model 2 is a Transformer, and model 3 is a Bidirectional LSTM.

Mod.	Epoch	Simp.	Bayes.	Prop.
1	2	83.87	84.22	**85.12**
1	4	80.41	78.96	**85.02**
1	6	82.72	82.35	**86.31**
1	8	81.11	81.66	**87.00**
1	10	70.28	81.27	**86.94**
2	2	83.87	84.15	**89.84**
2	4	84.10	82.42	**89.59**
2	6	83.40	83.96	**89.77**
2	8	82.40	85.97	**90.09**
2	10	82.32	86.89	**89.95**
3	2	86.87	**88.32**	87.72
3	4	70.74	87.81	**88.41**
3	6	85.71	88.04	**88.20**
3	8	86.25	88.00	**88.02**
3	10	87.48	88.44	**89.51**

For the *gn-offensive-language-identification* dataset, the results are presented in Table 2. In general, the average accuracies are higher than in the first dataset (>80% in almost all combinations). As before, our proposal achieved better performance in nearly all cases (except for model 3 with 2 epochs), and by a significantly larger margin for the 1D ConvNet and transformer-based models. In this dataset, the best configuration recorded (90.09%) is the transformer-based architecture when trained with the proposal for eight epochs.

Table 3: Average accuracy on the gn-emotion-recognition dataset as the number of training epochs increases. Model 1 is a 1D ConvNet, model 2 is a Transformer, and model 3 is a Bidirectional LSTM.

Mod.	Epoch	Simp.	Bayes.	Prop.
1	2	37.78	48.92	**55.43**
1	4	41.27	49.30	**55.05**
1	6	49.84	50.98	**55.43**
1	8	50.16	50.06	**55.97**
1	10	45.71	49.21	**56.10**
2	2	37.78	36.00	**48.29**
2	4	45.08	39.87	**51.71**
2	6	47.30	47.62	**55.27**
2	8	47.62	47.84	**55.87**
2	10	51.75	53.08	**56.03**
3	2	45.71	52.06	**55.17**
3	4	50.19	52.48	**56.92**
3	6	51.83	52.86	**58.35**
3	8	52.70	51.30	**57.75**
3	10	52.06	53.33	**57.84**

Table 3 shows the results for the last dataset,

gn-emotion-recognition. In this case, the accuracy values are substantially lower than those presented in Tables 1 and 2; therefore, it is the most challenging dataset. Another interesting point is that, for all epoch values, the results for the bidirectional LSTM-based models are superior to those of the other models. As with the previous datasets, our proposal consistently outperforms the other configurations. The best accuracy (58.35%) corresponds to the bidirectional LSTM model with the proposal, trained for six epochs.

Figures 1, 2, and 3 illustrate the average execution times observed across models. The results suggest that execution time is more strongly influenced by the network architecture than by the dataset itself. For the 1D ConvNet and bidirectional LSTM architectures, sector-based training achieved execution speeds approximately two and three times faster, respectively, compared to standard backpropagation. The proposal yielded speedups of up to 32× in relation to traditional backpropagation with Bayesian optimization. In the case of the Transformer architecture, sector-based training incurred an execution time up to 10% longer than traditional training; however, with Bayesian optimization, it demonstrated a 12× improvement in efficiency.

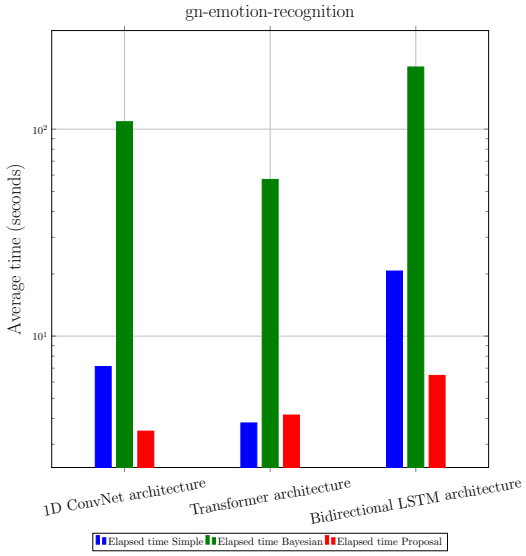

Figure 1: Average execution time for the gn-emotion-recognition dataset.

4 Conclusion

The experiments as a whole showed three notable advantages of sector training over traditional methods for text classification in Guarani using deep architectures. Firstly, for each dataset and algorithm, the highest average accuracy was always achieved by sector training during some epoch. This advantage ranged from less than 1% to almost 6% compared to the best value achieved by traditional methods. Second, the average accuracy is more stable for sector training, which does not show significant declines in later epochs, as can happen with traditional methods. Finally, the greatest advantage identified is the efficiency in execution time of sector training, which was not always lower than traditional simple backpropagation, but was nevertheless 12 to 32 times less costly than traditional backpropagation with Bayesian optimisation and with superior accuracy. This is noteworthy because traditional backpropagation with Bayesian optimisation represents the best traditional configuration in terms of average accuracy.

As future work, we plan to evaluate the approach on multi-class classification tasks with alternative loss functions, extend experiments to more tasks and languages, and analyze its scalability with different sector sizes and smaller datasets.

Limitations

The evaluation was restricted to three small Guarani affective computing datasets, which may limit generalization to other tasks or languages. Moreover, the scalability of sector-wise backpropagation to larger architectures and broader benchmarks remains to be explored.

Disclaimer

During the preparation of this work the authors used generative tools in order to fix misspellings and improve writing. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Code Availability

The code for reproducing the experiments presented in this paper is publicly accessible at `https://gitlab.com/pinv01-401/dloptimizer`.

Acknowledgments

This work was supported by the CONACYT, Paraguay, under Grant PINV01-401.

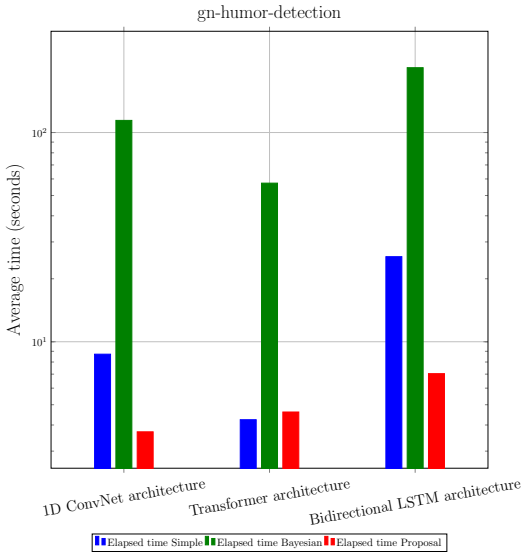

Figure 2: Average execution time for the gn-humor-detection dataset.

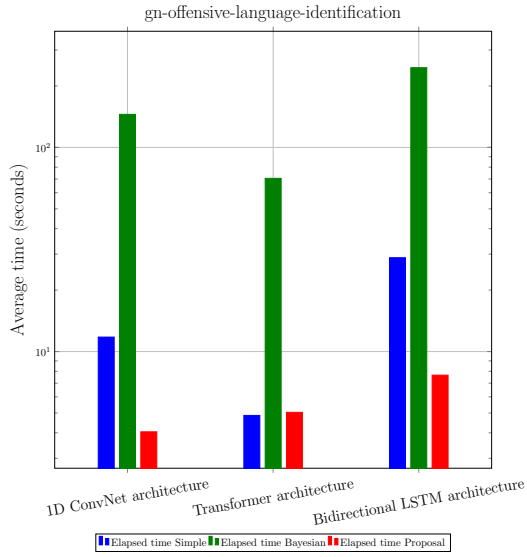

Figure 3: Average execution time for the gn-offensive-language-identification dataset.

References

Marvin M Agüero-Torales, Antonio G López-Herrera, and David Vilares. 2023. Multidimensional affective analysis for low-resource languages: A use case with guarani-spanish code-switching language. *Cognitive Computation*, 15(4):1391–1406.

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Baljekar, Xavier Garcia, Wolfgang Macherey, Theresa Breiner, Vera Axelrod, Jason Riesa, Yuan Cao, Mia Xu Chen, Klaus Macherey, Maxim Krikun, Pidong Wang, Alexander Gutkin, Apurva Shah, Yanping Huang, Zhifeng Chen, Yonghui Wu, and Macduff Hughes. 2022. Building machine translation systems for the next thousand languages.

Eugene Belilovsky, Michael Eickenberg, and Edouard Oyallon. 2019. Greedy layerwise learning can scale to imagenet. In *International conference on machine learning*, pages 583–593. PMLR.

Eugene Belilovsky, Michael Eickenberg, and Edouard Oyallon. 2020. Decoupled greedy learning of cnns. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 748–758. PMLR.

Luis Chiruzzo, Pedro Amarilla, Adolfo Ríos, and Gustavo Giménez Lugo. 2020. Development of a Guarani - Spanish parallel corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2629–2633, Marseille, France. European Language Resources Association.

Luis Chiruzzo, Santiago Góngora, Aldo Alvarez, Gustavo Giménez-Lugo, Marvin Agüero-Torales, and Yliana Rodríguez. 2022. Jojajovai: A parallel Guarani-Spanish corpus for MT benchmarking. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2098–2107, Marseille, France. European Language Resources Association.

François Chollet. 2020. Bidirectional lstm on imdb. `https://keras.io/examples/nlp/bidirectional_lstm_imdb/`.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Bruno Estigarribia. 2016. Guaraní aquí, jopara allá. reflexiones sobre la (socio)lingüística paraguaya, written by penner, hedy. *Journal of Language Contact*, 9(2):397 – 403.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell.

2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. 1989. Backpropagation applied to handwritten zip code recognition. *Neural Comput.*, 1(4):541–551.

Agustín Lucas, Alexis Baladón, Victoria Pardiñas, Marvin Agüero-Torales, Santiago Góngora, and Luis Chiruzzo. 2024. Grammar-based data augmentation for low-resource languages: The case of Guarani-Spanish neural machine translation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6385–6397, Mexico City, Mexico. Association for Computational Linguistics.

Rui Mao, Qian Liu, Kai He, Wei Li, and Erik Cambria. 2023. The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection. *IEEE Trans. Affect. Comput.*, 14(3):1743–1753.

Apoorv Nandan. 2020. Text classification with transformer. `https://keras.io/examples/nlp/text_classification_with_transformer/`.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Mark Omernick and François Chollet. 2019. Text classification from scratch. `https://keras.io/examples/nlp/text_classification_from_scratch/`.

Kaushik Roy. 2024. *Lifelong Learning with Neural Network*. Ph.D. thesis, MONASH University.

Adolfo A. Ríos, Pedro J. Amarilla, and Gustavo A. Giménez Lugo. 2014. Sentiment categorization on a creole language with lexicon-based and machine learning techniques. In *2014 Brazilian Conference on Intelligent Systems*, pages 37–43.

Mike Schuster and Kuldip K. Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. Cross-lingual transfer learning for multilingual task oriented dialog. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3795–3805, Minneapolis, Minnesota. Association for Computational Linguistics.

Melika Sadeghi Tabrizi, Ali Karimi, Ahmad Kalhor, Babak N Araabi, and Mona Ahmadian. 2024. Layer-wise learning of cnns by self-tuning learning rate and early stopping at each layer. In *35th British Machine Vision Conference 2024, BMVC 2024, Glasgow, UK, November 25-28, 2024*. BMVA.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Alex Waibel, Manfred Hanazawa, Gregory Hinton, Kevin Shikano, and Kevin J. Lang. 1989. Phoneme recognition using time-delay neural networks. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(3):328–339.