Project 2: Transformers

Mikołaj Rzepiński, Krzysztof Wolny

Presentation plan

- Models
- Methodology
- Learning Rate
- Batch Size
- Data Augmentation
- Best models for 4 label dataset
- Results for all label dataset
- Summary

Models

- **Wav2Vec2.0**: A self-supervised model that learns speech representations from raw audio. It pretrains on unlabeled data and is fine-tuned for tasks like speech recognition, reducing the need for large labeled datasets.
- **LSTM-CNN Network(LSTM)**: A custom-built model combining CNNs for extracting features from audio and LSTMs for capturing temporal patterns, trained from scratch.
- **Audio Spectrogram Transformer (AST)**: A transformer-based model that operates on spectrograms, leveraging transformer architecture strengths for audio understanding.

Methodology

- Models were first trained on a small dataset (4 labels: yes, no, stop, go), with 1000 samples per label.
- Data split: 70% training, 15% validation, 15% testing.
- Multiple hyperparameter and augmentation combinations were tested.
- Best-performing setups were then trained on the full dataset with all labels.
- AST and Wav2Vec2.0: fine-tuned from pretrained weights (feature extractors frozen).
- LSTM-CNN: trained from scratch.
- Training done in TensorFlow on Google Colab for 10 epochs per model.

Methodology

- Learning Rate:
  - 0.0001
  - 0.001
  - 0.01
- Batch Size:
  - 32
  - 64
- Data Augmentations:
  - Background Noise Injection (BN)
  - Time Masking (TM)
  - No Augmentation

Learning rate

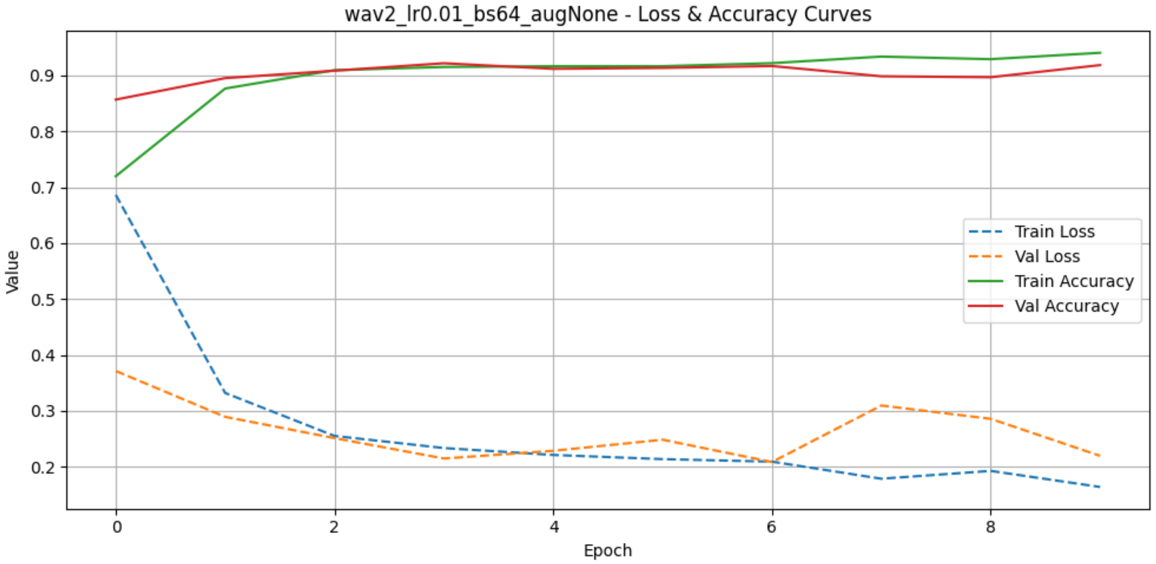
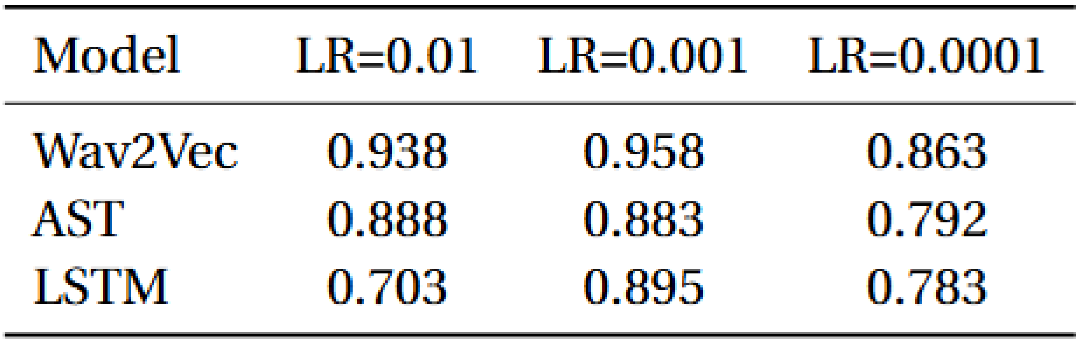
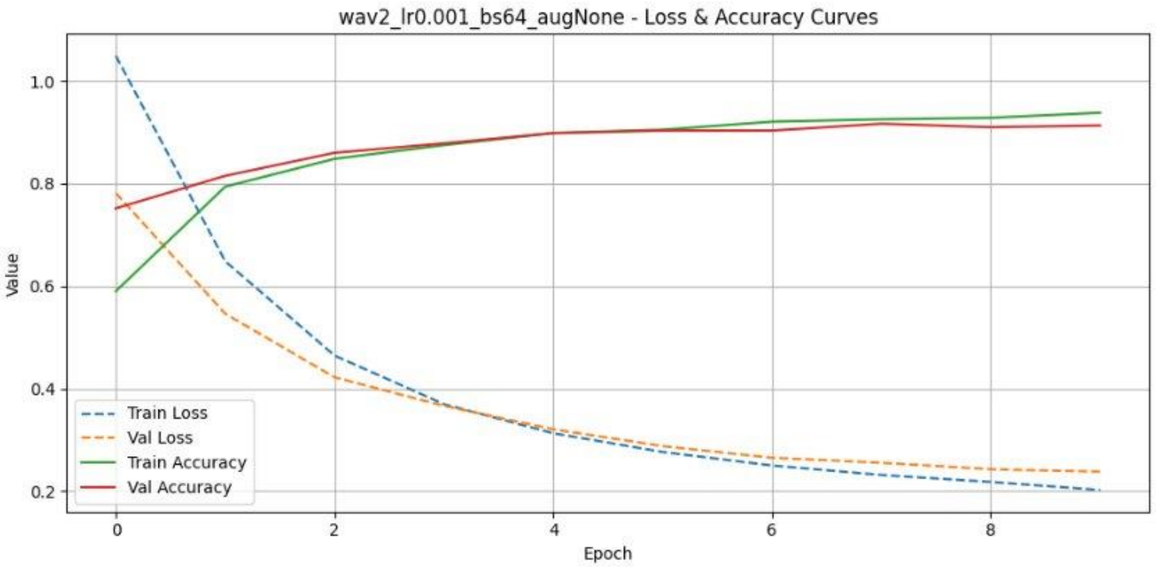


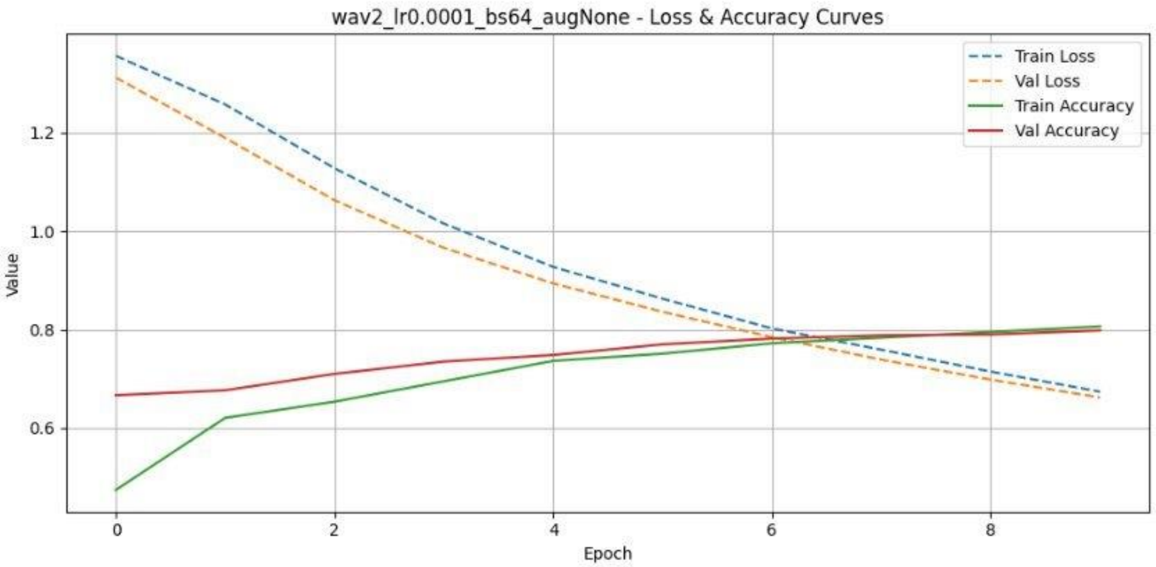
Table 4.2. Best test accuracy for different learning rates



Model	LR=0.01	LR=0.001	LR=0.0001
Wav2Vec	0.938	0.958	0.863
AST	0.888	0.883	0.792
LSTM	0.703	0.895	0.783

Learning rate


wav2_lr0.001_bs64_augNone - Loss & Accuracy Curves


wav2_lr0.0001_bs64_augNone - Loss & Accuracy Curves

Batch size

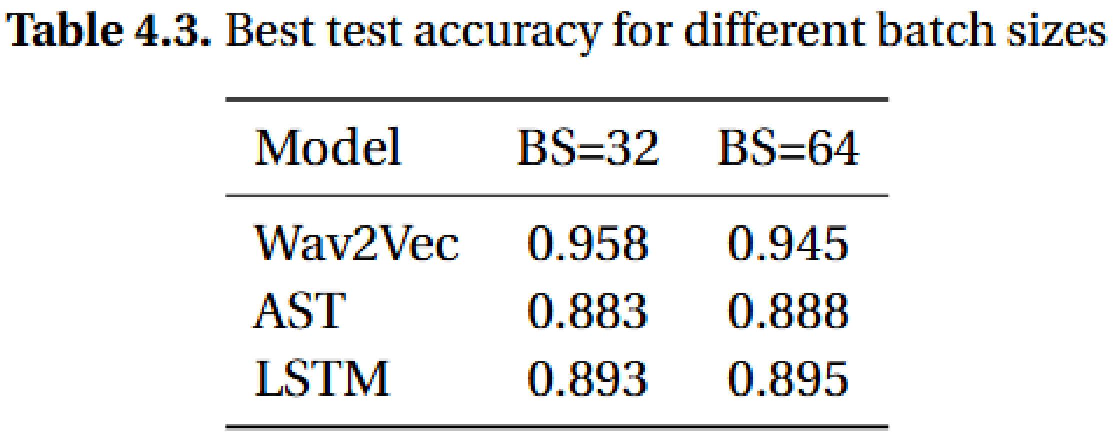
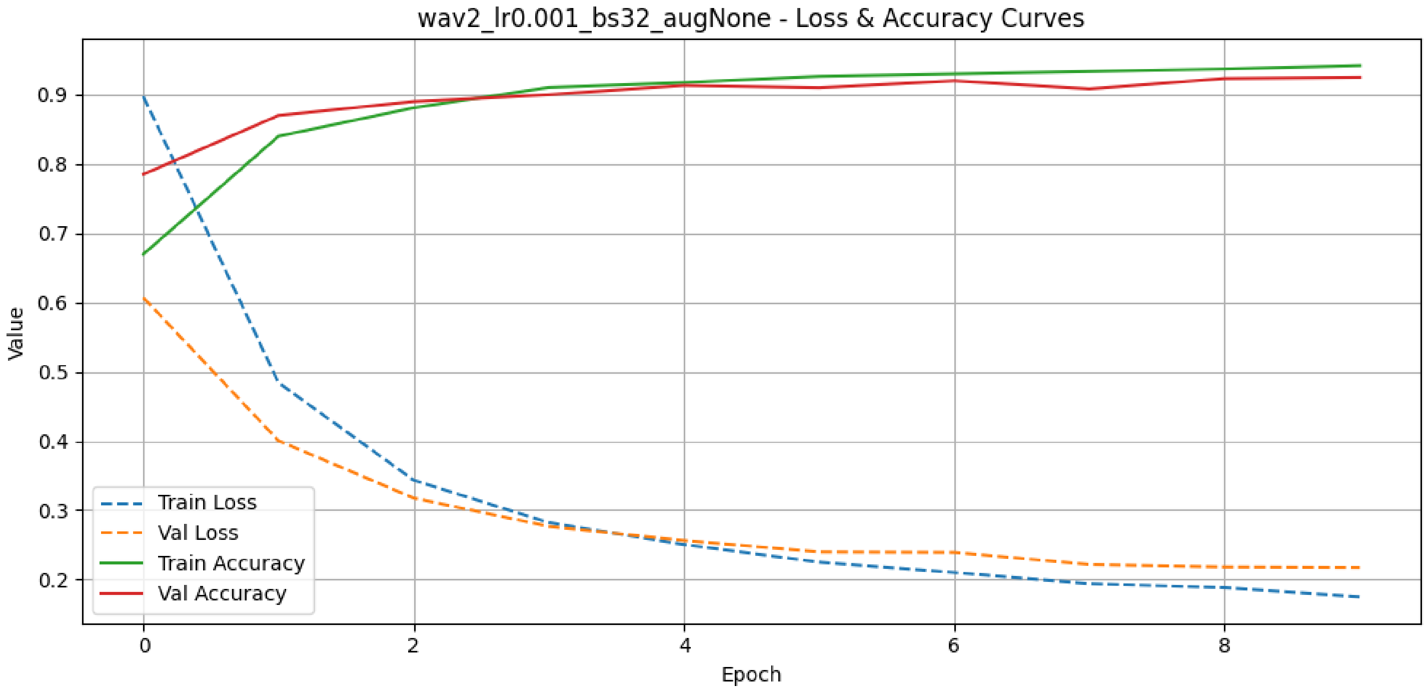
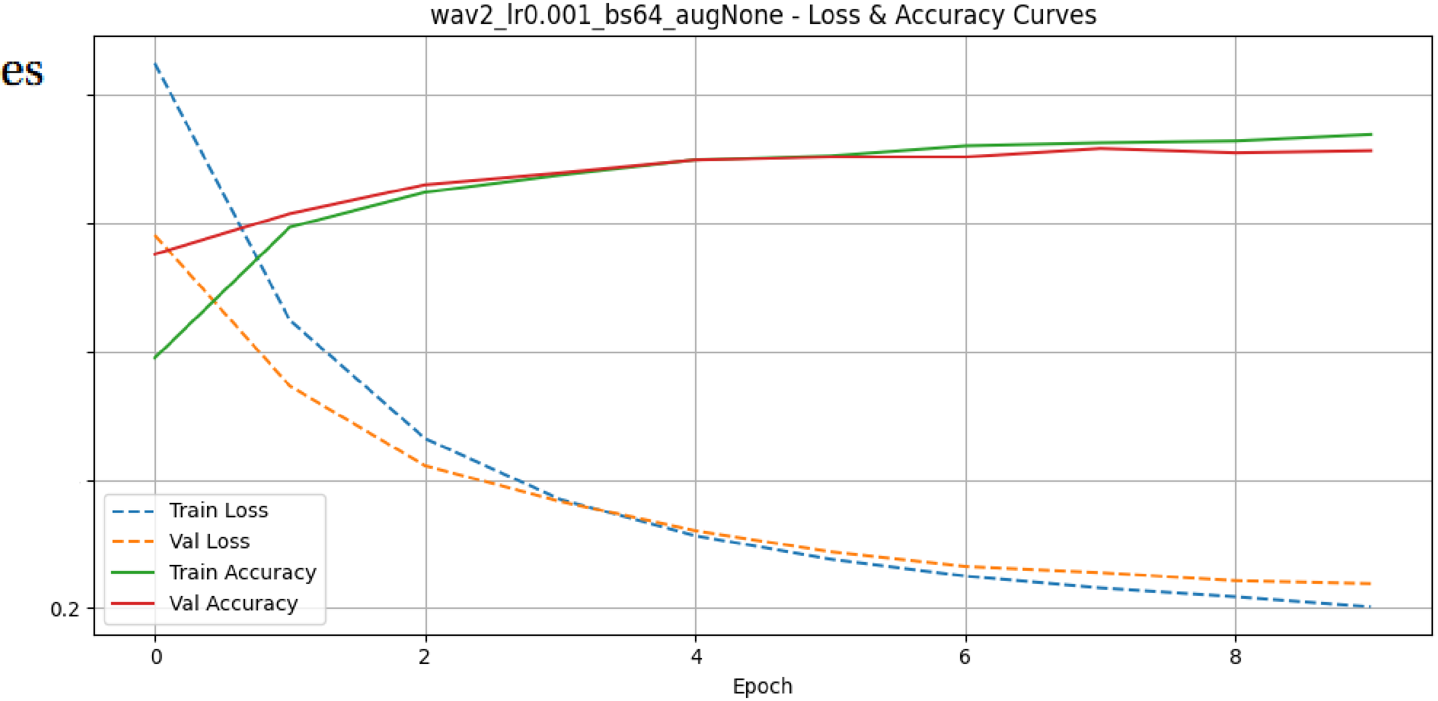


Table 4.3. Best test accuracy for different batch sizes

Model	BS=32	BS=64
Wav2Vec	0.958	0.945
AST	0.883	0.888
LSTM	0.893	0.895



wav2_lr0.001_bs32_augNone - Loss & Accuracy Curves



wav2_lr0.001_bs64_augNone - Loss & Accuracy Curves

Data Augmentation

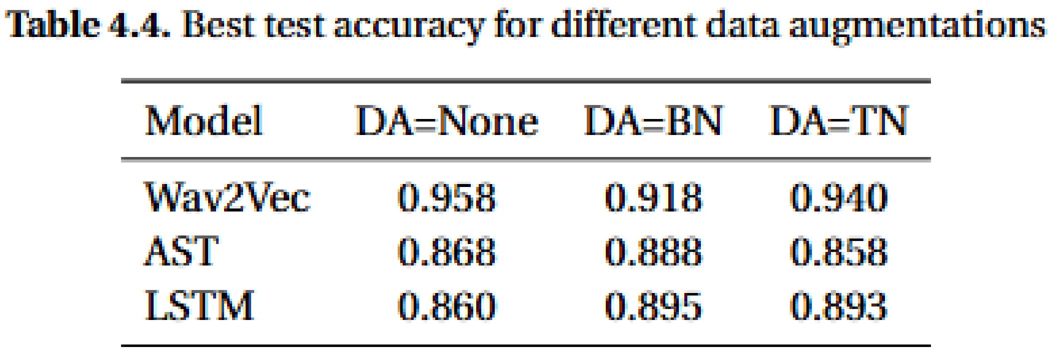
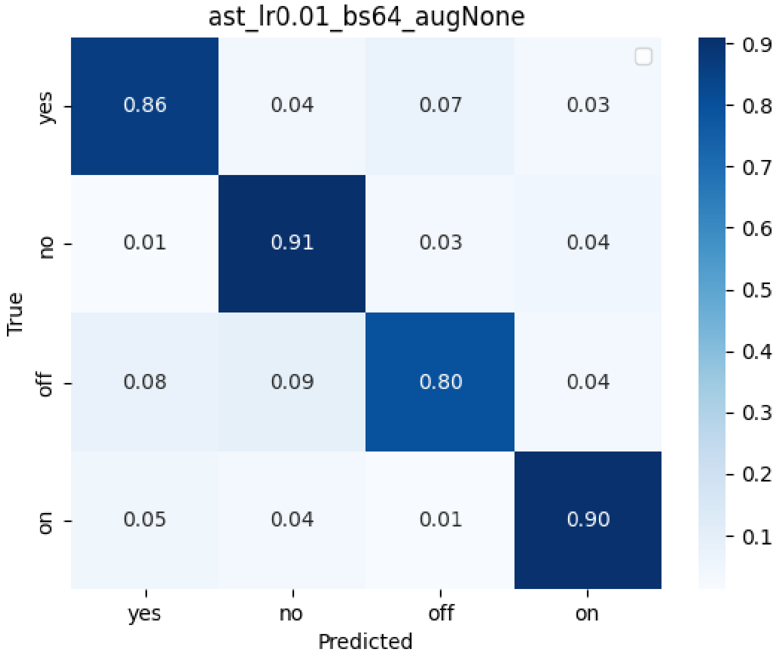


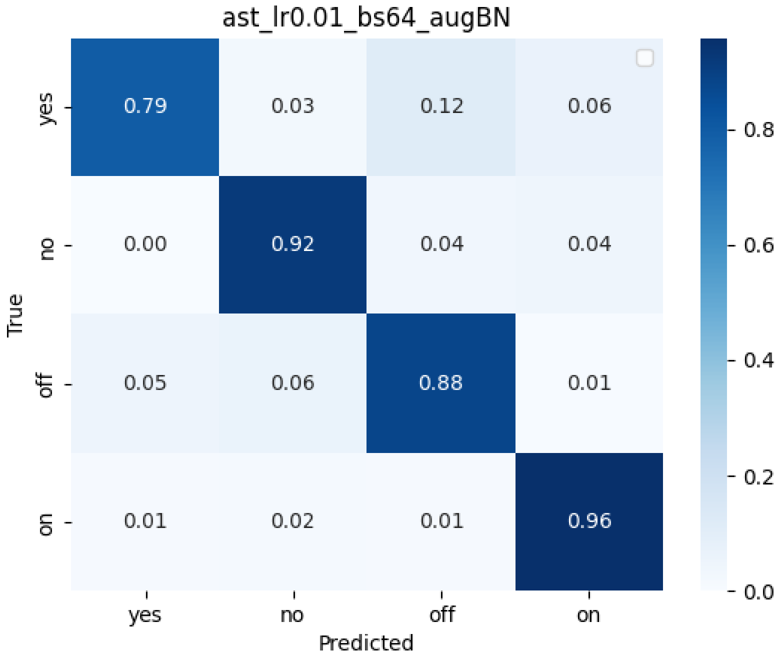
Table 4.4. Best test accuracy for different data augmentations

Model	DA=None	DA=BN	DA=TN
Wav2Vec	0.958	0.918	0.940
AST	0.868	0.888	0.858
LSTM	0.860	0.895	0.893



ast_lr0.01_bs64_augNone

True \ Predicted	yes	no	off	on
yes	0.86	0.04	0.07	0.03
no	0.01	0.91	0.03	0.04
off	0.08	0.09	0.80	0.04
on	0.05	0.04	0.01	0.90



ast_lr0.01_bs64_augBN

True \ Predicted	yes	no	off	on
yes	0.79	0.03	0.12	0.06
no	0.00	0.92	0.04	0.04
off	0.05	0.06	0.88	0.01
on	0.01	0.02	0.01	0.96

Best models for 4 label dataset

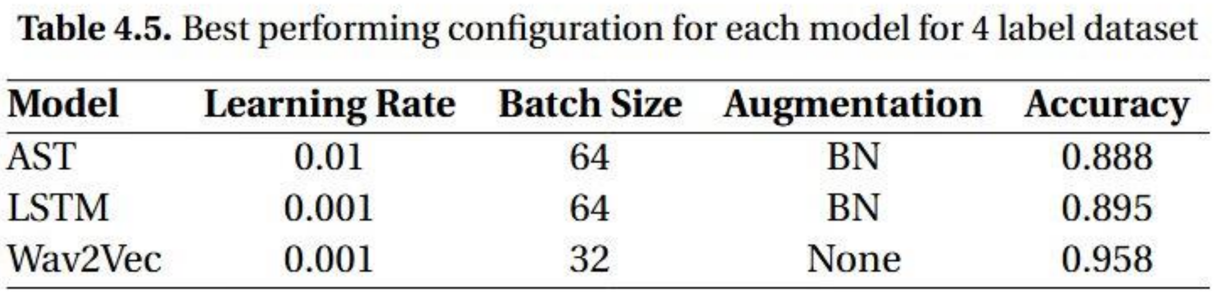
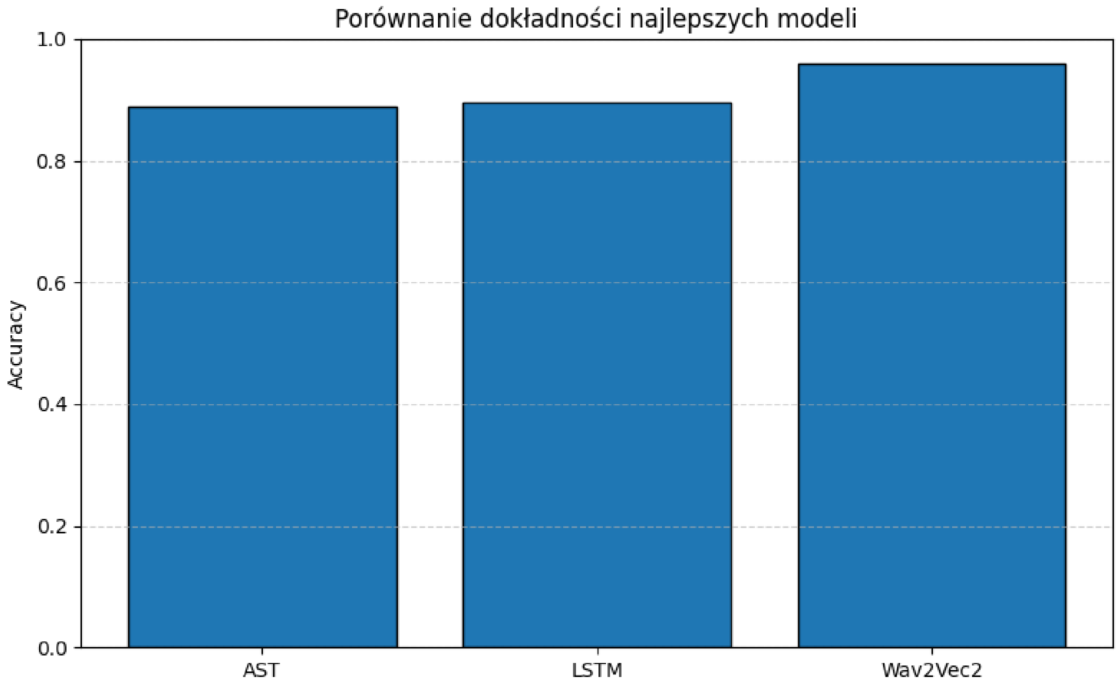
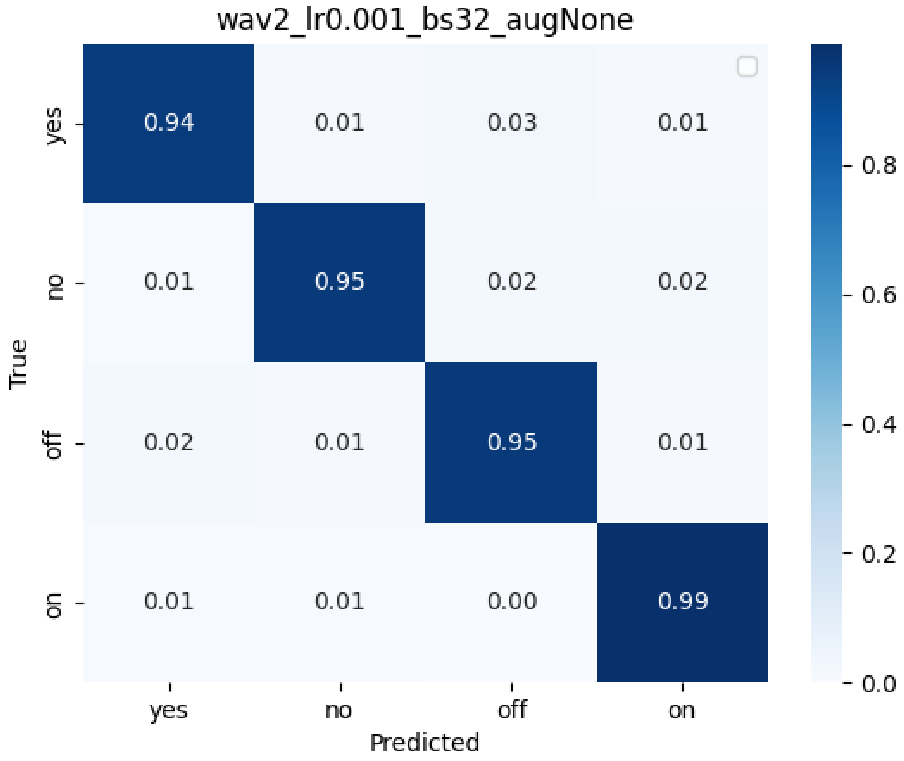


Table 4.5. Best performing configuration for each model for 4 label dataset

Model	Learning Rate	Batch Size	Augmentation	Accuracy
AST	0.01	64	BN	0.888
LSTM	0.001	64	BN	0.895
Wav2Vec	0.001	32	None	0.958



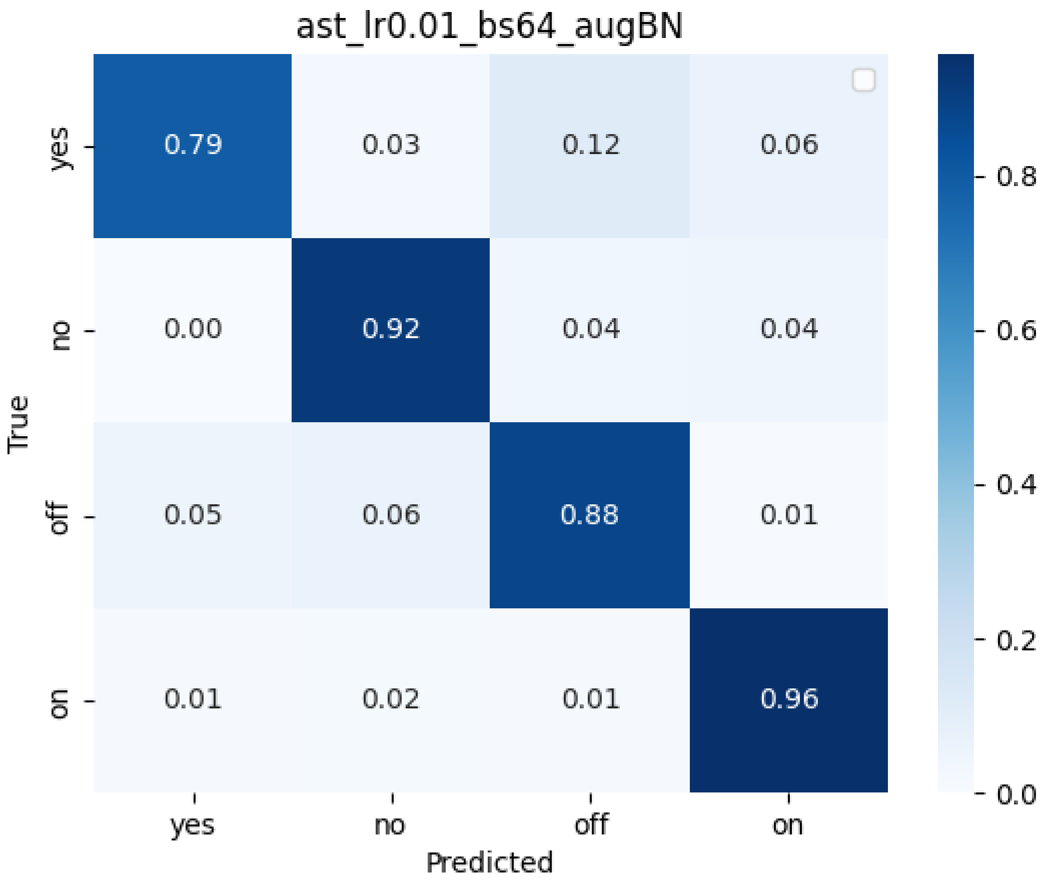
Porównanie dokładności najlepszych modeli



wav2_lr0.001_bs32_augNone

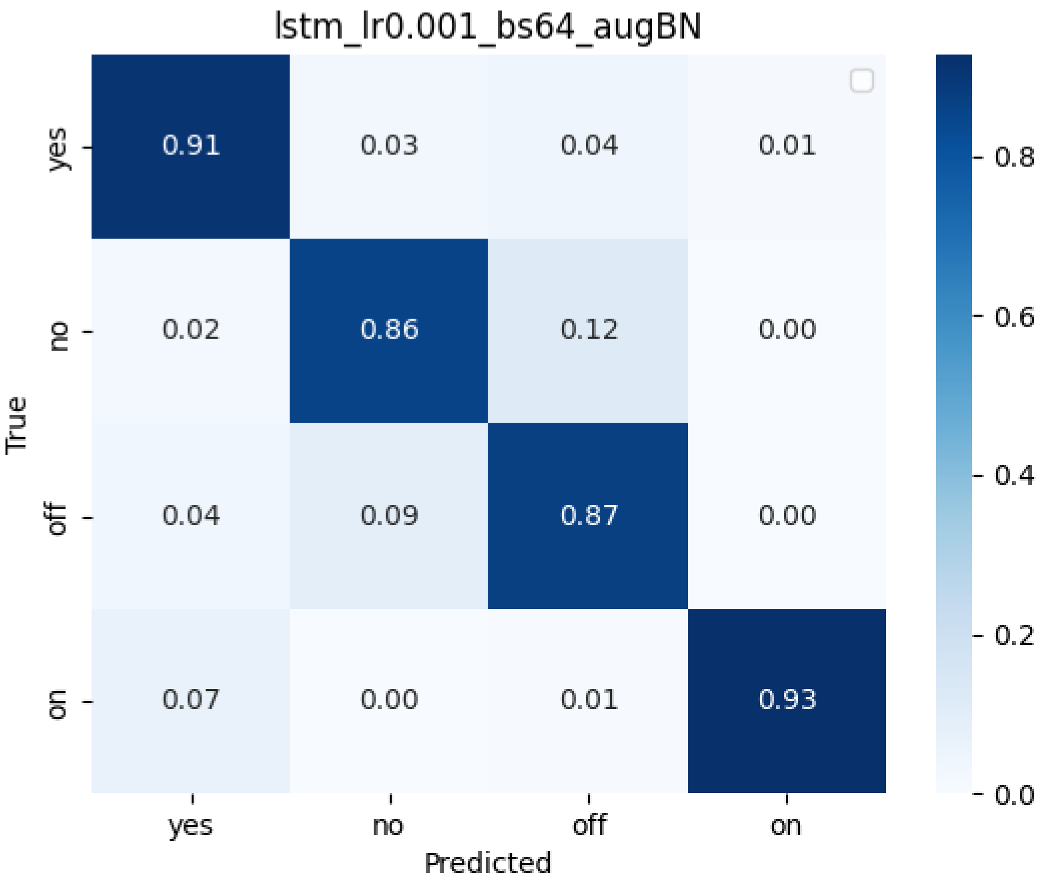
True \ Predicted	yes	no	off	on
yes	0.94	0.01	0.03	0.01
no	0.01	0.95	0.02	0.02
off	0.02	0.01	0.95	0.01
on	0.01	0.01	0.00	0.99

Best models for 4 label dataset



ast_lr0.01_bs64_augBN

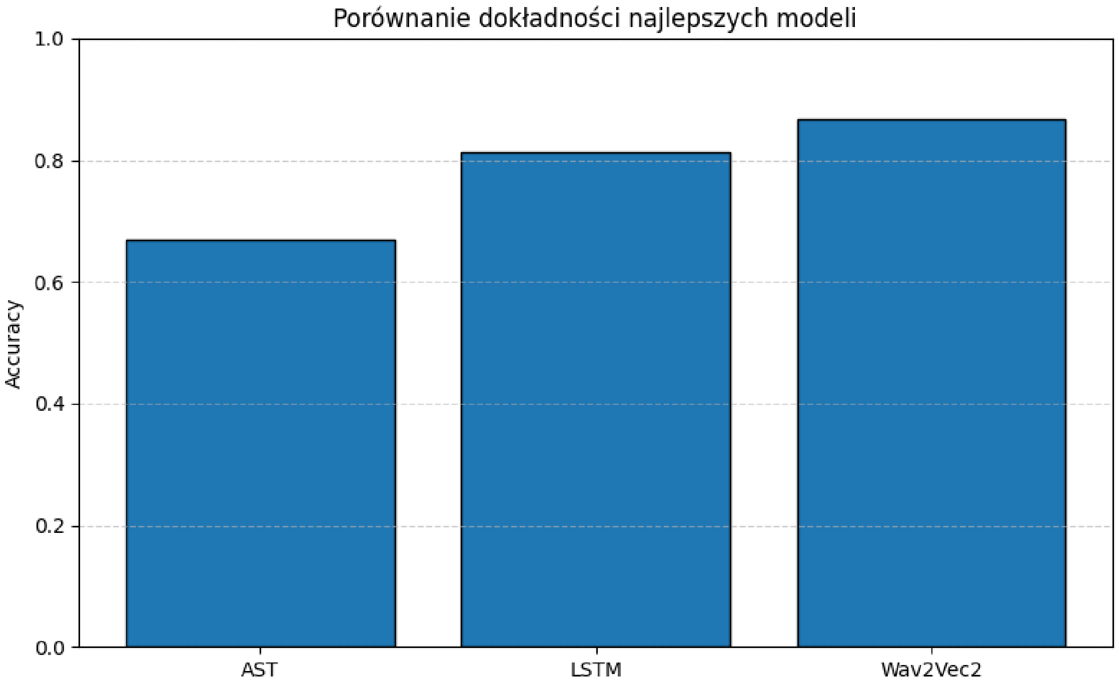
True \ Predicted	yes	no	off	on
yes	0.79	0.03	0.12	0.06
no	0.00	0.92	0.04	0.04
off	0.05	0.06	0.88	0.01
on	0.01	0.02	0.01	0.96

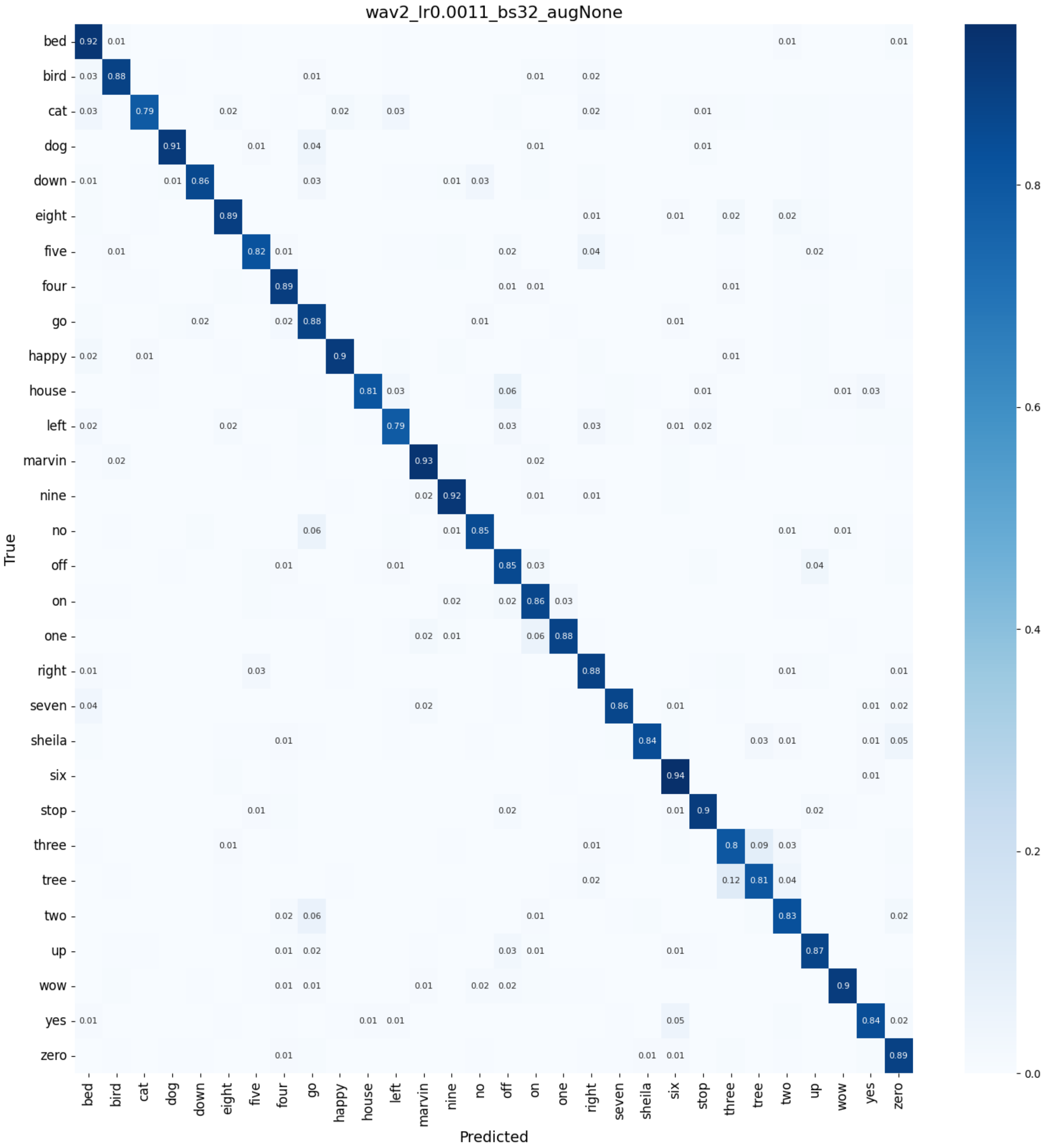


lstm_lr0.001_bs64_augBN

True \ Predicted	yes	no	off	on
yes	0.91	0.03	0.04	0.01
no	0.02	0.86	0.12	0.00
off	0.04	0.09	0.87	0.00
on	0.07	0.00	0.01	0.93

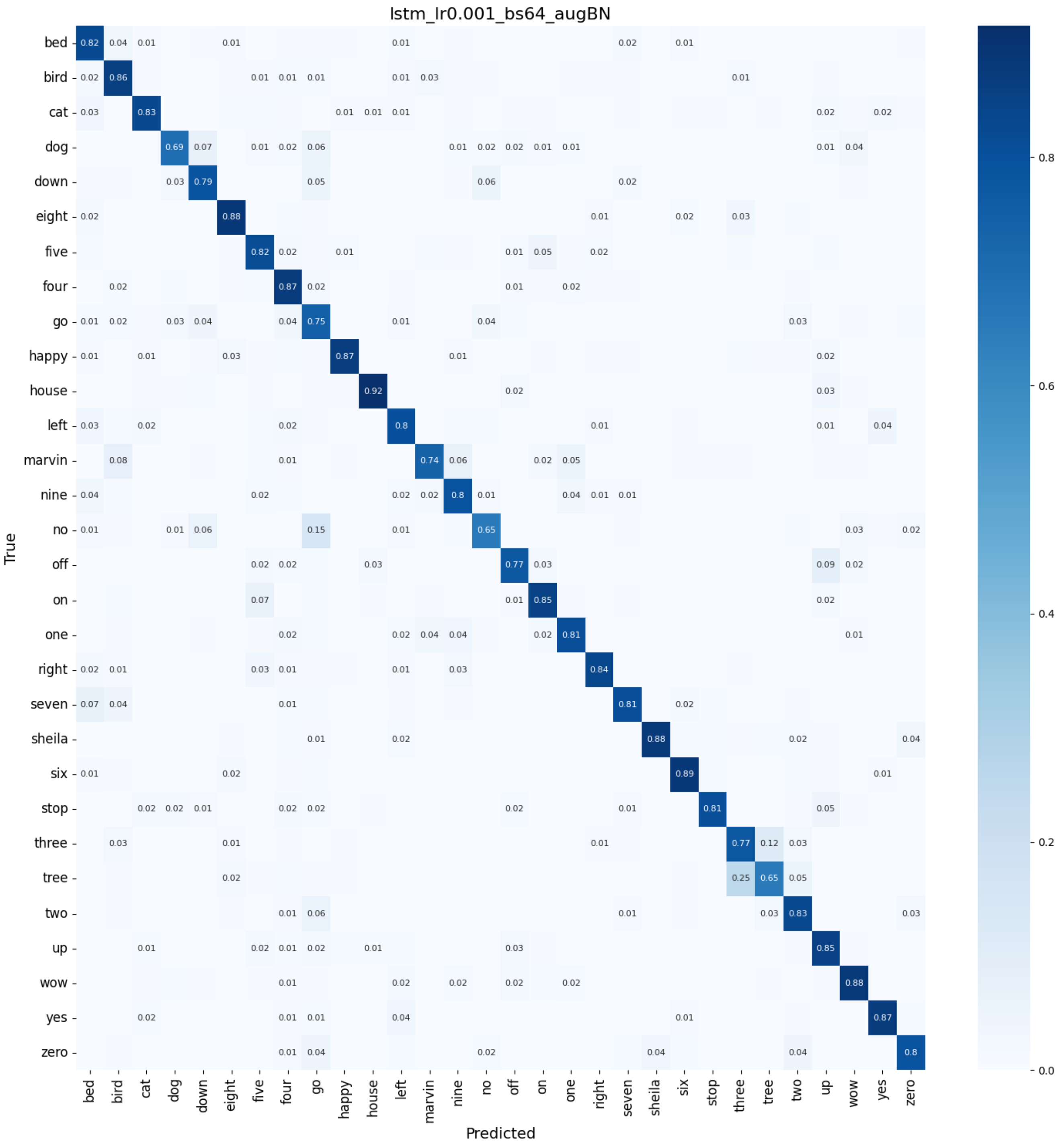
Results for all label dataset - Wav2Vec



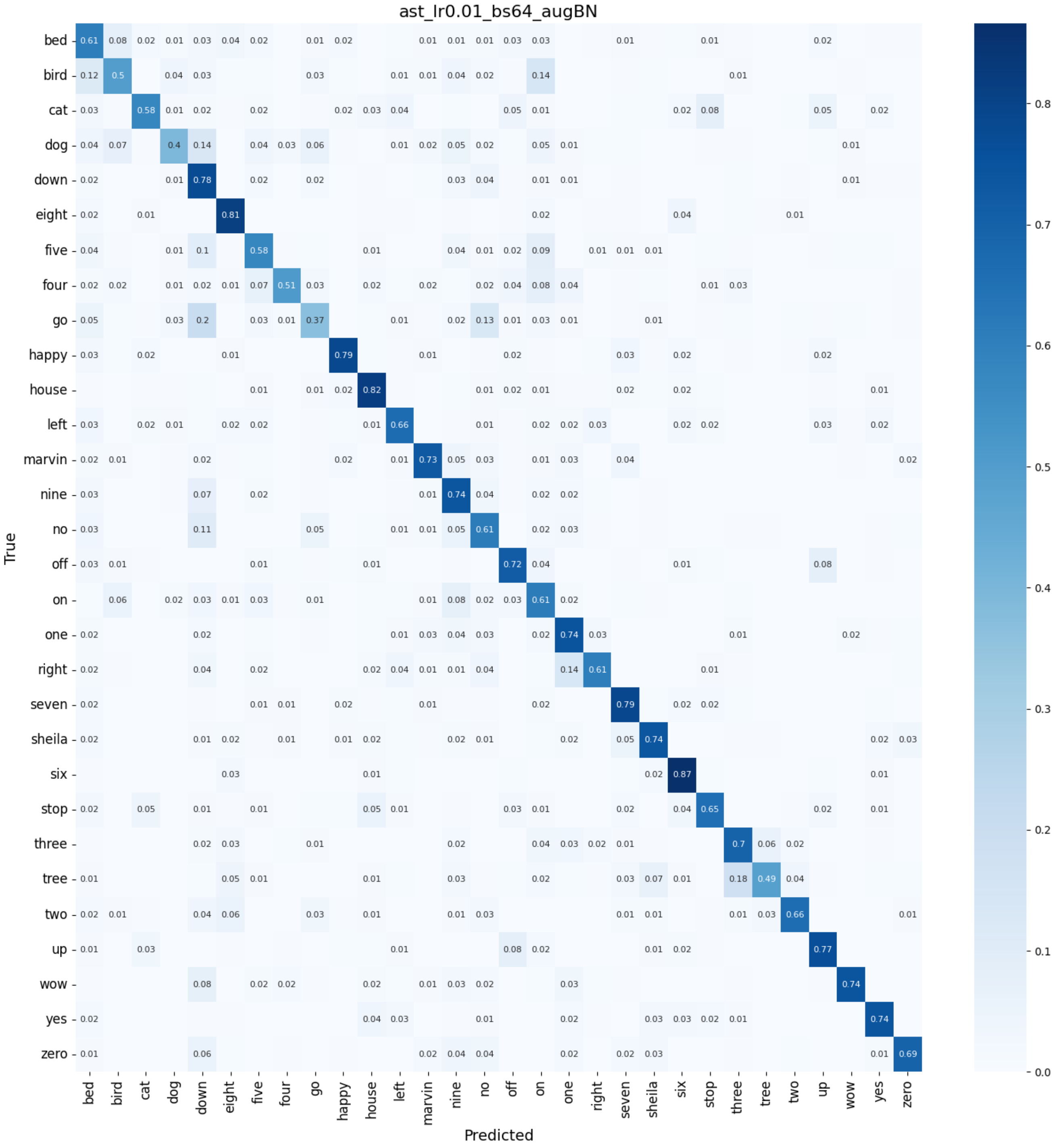


	bed	bird	cat	dog	down	eight	five	four	go	happy	house	left	marvin	nine	no	off	on	one	right	seven	sheila	six	stop	three	tree	two	up	wow	yes	zero
bed	0.92	0.01																								0.01				0.01
bird	0.03	0.88							0.01								0.01		0.02											
cat	0.03		0.79			0.02				0.02		0.03							0.02				0.01							
dog				0.91			0.01		0.04								0.01						0.01							
down	0.01			0.01	0.86				0.03					0.01	0.03															
eight						0.89													0.01			0.01		0.02		0.02				
five		0.01					0.82	0.01								0.02			0.04								0.02			
four								0.89								0.01	0.01							0.01						
go					0.02			0.02	0.88						0.01							0.01								
happy	0.02		0.01							0.9														0.01						
house											0.81	0.03				0.06							0.01					0.01	0.03	
left	0.02					0.02						0.79				0.03			0.03			0.01	0.02							
marvin		0.02											0.93				0.02													
nine													0.02	0.92			0.01		0.01											
no									0.06					0.01	0.85											0.01		0.01		
off								0.01				0.01				0.85	0.03										0.04			
on														0.02		0.02	0.86	0.03												
one													0.02	0.01			0.06	0.88												
right	0.01						0.03												0.88							0.01				0.01
seven	0.04												0.02							0.86		0.01							0.01	0.02
sheila								0.01													0.84				0.03	0.01			0.01	0.05
six																						0.94							0.01	
stop							0.01									0.02						0.01	0.9				0.02			
three						0.01													0.01					0.8	0.09	0.03				
tree																			0.02					0.12	0.81	0.04				
two								0.02	0.06								0.01									0.83				0.02
up								0.01	0.02							0.03	0.01					0.01					0.87			
wow								0.01	0.01				0.01		0.02	0.02												0.9		
yes	0.01										0.01	0.01										0.05							0.84	0.02
zero								0.01													0.01	0.01								0.89

Results for all label dataset - LSTM



Results for all label dataset - AST



Summary

- A learning rate of 0.001 generally got the best performance
- Wav2Vec achieved its highest accuracy with a smaller batch size (32), whereas AST and LSTM slightly benefited from a larger batch size (64).
- Data augmentation using background noise (BN) improved performance for AST and LSTM, while Wav2Vec performed best without any augmentation.
- On the 4-label dataset, Wav2Vec achieved the highest accuracy (95.8%), followed by LSTM (89.5%) and AST (88.8%) with their respective best configurations.
- On the dataset with all labels, Wav2Vec also outperformed other models (86.7%), with LSTM reaching 80.2% and AST achieving 65.7%.

Bibliography

- Y. Gong, Y. Chung, and J. R. Glass, “AST: audio spectrogram transformer”, CoRR, vol. abs/2104.01778, 2021. arXiv: 2104.01778. [Online]. Available https://arxiv.org/abs/2104.01778.
- [2] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, “Wav2vec 2.0: A framework for self-supervised learning of speech representations”, CoRR, vol. abs/2006.11477, 2020.arXiv: 2006.11477. [Online]. Available: https://arxiv.org/abs/2006.11477.
- [3] J. Donahue, L. A. Hendricks, S. Guadarrama, et al., “Long-term recurrent convolutional networks for visual recognition and description”, CoRR, vol. abs/1411.4389, 2014. arXiv: 1411.4389. [Online]. Available: http://arxiv.org/abs/1411.4389.
- https://machinelearningmastery.com/cnn-long-short-term-memory-networks/
- https://huggingface.co/facebook/wav2vec2-base
- https://huggingface.co/docs/transformers/en/model_doc/audio-spectrogram-transformer