INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Siamese Network with Wav2vec Feature for Spoofing Speech Detection

Yang Xie , Zhenchuan Zhang , Yingchun Yang[*]

Zhejiang University, China

wangxiaohu@zju.edu.cn, 11221052@zju.edu.cn, yyc@zju.edu.cn

Abstract

Automatic speaker verification is vulnerable to spoofing attacks with synthesized or converted speech. Although high-performance anti-spoofing countermeasures can achieve high accuracy when the training and testing spoofing attack examples are similarly distributed, their performance degrades significantly when confronted with out-of-distribution spoofing speech, which is created by increasingly advanced unseen speech synthesis and voice conversion methods. Since it is unrealistic to collect enough labeled data from each new spoofing attack method, we argue that addressing the problem of out-of-distribution generalization for spoofing speech detection is essential. In this work, we propose a two-phase representation learning system based on a Siamese network for spoofing speech detection tasks. During the representation learning phase, an embedding Siamese neural network is trained with the wav2vec features to distinguish whether the speech samples in a pair belong to the same category. The proposed system decreases the equal error rate from the state-of-the-art result of 4.07% to 1.15% on the ASVspoof 2019 evaluation set.

Index Terms: antispoofing, logical access, ASVspoof 2019 challenge, representation learning, Siamese network

1. Introduction

Automatic speaker verification (ASV) systems have attracted increasing interest as a form of biometric authentication. The main deterrent to the adoption of ASV is that it is vulnerable to malicious spoofing attacks. For ASV, two types of spoofing attacks have been identified: logical access (LA) spoofing attacks (i.e., using voice conversion (VC) and speech synthesis) and physical access (PA) spoofing attacks or replay attacks (i.e., replaying a recording of a target speaker). As audio spoofing techniques continue to be developed, such as the voice conversion model i-vector VC [1] and the speech synthesis model FastSpeech [2], it is necessary to augment an ASV system with a logical attack (LA) detection system in practical applications. In this work, we aim to provide countermeasures against spoofing attacks based on speech synthesis and voice conversion.

The performance of spoofing speech detection systems is highly dependent on the design of discriminative features. Specifically, constant Q cepstral coefficients (CQCCs) [3], which are used as baseline features, perform better than traditional Mel frequency cepstral coefficients (MFCCs). In recent years, linear frequency cepstral coefficients (LFCCs) [4] and group delay (GD) and modified group delay (MGD) [5] features based on phase, log power magnitude spectra (logspec) [6] and X-vector embedding [7] have achieved improved performance for spoofing speech detection tasks. Furthermore, powerful feature extractors based on deep learning have been proposed [8, 9]. Other works have explored back-end classifiers. For example, the classic Gaussian mixture model (GMM) with LFCC [4] has been used as the baseline system. Zhang [6] proposed a model—channel consistency DenseNeXt—that reduces the number of parameters and amount of computing power without loss of performance. Furthermore, Lai [10] proposed a model based on SENet and ResNet with statistical pooling, which achieved considerable improvement over previous models.

In recent years, pretrained models have rapidly advanced the state-of-the-art for many computer vision (CV), natural language processing (NLP), and speech tasks. In NLP, pretrained models such as ELMo[11] and BERT [12] are effective for a variety of NLP tasks, including document classification, dialog system, and abstractive summarization. Many CV methods [13, 14, 15] that pretrain a feature representation from a large-scale dataset such as ImageNet have shown effectiveness for downstream tasks. In speech processing, the unsupervised pretrained model wav2vec [16] outperforms the best character-based model on speech recognition. According to previous research [17, 15], the features extracted from the model pretrained on a large-scale dataset can be used for tasks that may not have sufficient labeled data or for which it is impossible to collect training data from every possible domain. Therefore, here we use wav2vec features instead of traditional acoustic features to improve the generalization of the model to unknown attacks.

In recent years, there has been great progress in supervised, semisupervised, and unsupervised representation learning. In CV, the representation learning framework SimCLR [18] has achieved top-1 accuracy on the ImageNet [19] dataset. In NLP, Sentence-BERT [20] using Siamese and triplet networks has significantly improved the performance of the model on sentence-pair regression and sentence classification tasks. Siamese networks have become a basic structure in these representation learning methods. Considering the empirical success of Siamese networks, we propose to use Siamese networks to learn a similarity metric for the task of spoofing speech detection.

In this paper, we propose systems with a Siamese network for representation learning and a back-end multilayer perceptron (MLP) [21] classifier for the task of spoofing speech detection. More generally, the embedding extractor based on a Siamese network extracts the representation vectors from the wav2vec features. During the phase of representation learning, the Siamese network [22] is trained to distinguish the positive sample, in which two types of speech belong to the same category, from the negative sample, in which two types of speech belong to different categories. For the Siamese network, we investigate five different network architectures and compare their representation learning abilities. During the classifier training phase, a two-layer fully connected neural network using representations obtained from the Siamese network is trained to minimize cross-entropy loss for spoofing speech detection.

The remainder of this paper is organized as follows. Section 2 provides the details of the wave2vec feature. Section 3 briefly describes the proposed Siamese network for spoofing detection.

[*] corresponding author

http://dx.doi.org/10.21437/Interspeech.2021-847

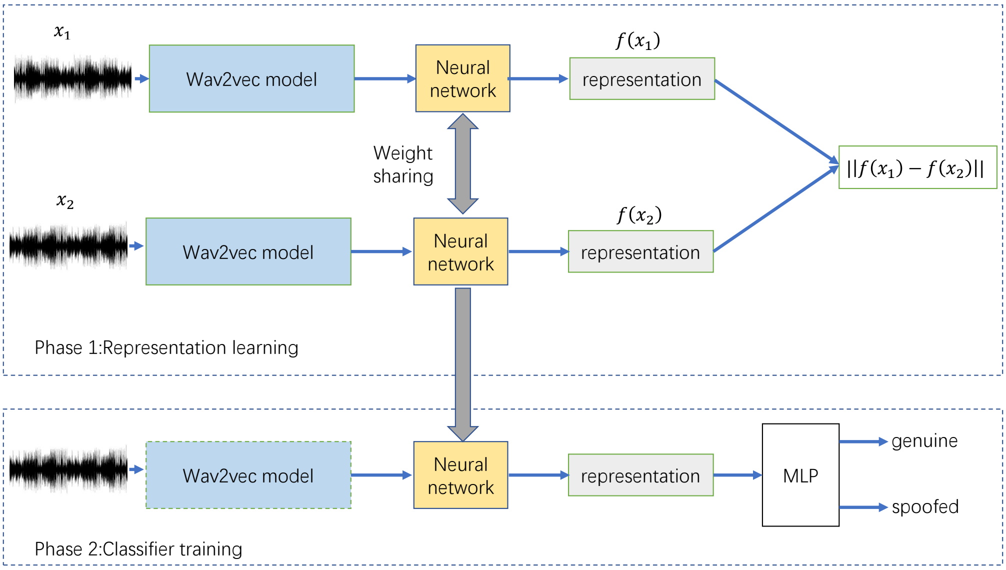

Figure 1: *Pipeline for spoofing speech detection using a Siamese network with wav2vec features. There are two phases: the representation learning phase and classification training phase. After representation learning, the embedding generated by the Siamese network is used as the input features of the classifier.*

In Section 4, the experimental setup is presented. Section 5 reports the overall results for different Siamese network architectures and the works of other researchers. Finally, the conclusion is presented in Section 6.

2. Wav2vec Feature

Hand-crafted features have traditionally worked well for speech recognition tasks. Commonly used features include MFCCs, LFCCs, and CQCCs. However, although these features work well for tasks where data are sufficient, such as automatic speech recognition, it is still difficult to use them for spoofing speech detection due to a lack of labeled data. Therefore, pretraining for representation learning has received increasing attention as an approach to obtain better discriminative features of audio, image, and text for tasks where labeled data are scarce.

Wav2vec is a self-supervised pretrained model and a multilayer convolutional neural network that takes audio signals as input and encodes them into a feature representation that can be input to downstream tasks. The wav2vec model consists of an encoder network and a context network. The encoder network $f : X \mapsto Z$ converts the speech signal into a low-frequency feature representation, and the output representation vector Z encodes approximately 30 ms of 16-kHz audio signal every 10 ms. Then, the context network $g : Z \mapsto C$ converts the output of the encoder network $Z_i ... Z_{i-v}$ into a context-level vector $C_i = g(Z_i ... Z_{i-v})$ for a receptive field size v. In practice, the context network has a 210-ms receptive field. The outputs of both the encoder network and the context network are used to compute the contrastive loss. Specifically, the model is trained to distinguish true future samples from distractor samples $\tilde{z}$ drawn from a distribution p_n. For each step $k = 1, ..., K$,

$$sim_k(Z_i, c_j) = log(\sigma(Z_i^T H_k c_j))) \tag{1}$$

$$L_K = -\Sigma(\log \sigma(sim_k(Z_{i+k}, c_i)) + \lambda \underset{\tilde{z} \sim p_n}{\mathbb{E}} (sim_k(-\tilde{z}, c_i)) \tag{2}$$

where $\sigma(x)$ denotes the logistic function and p_n denotes a proposal distribution. In practice, the model can choose $p_n(z) = \frac{1}{T}$, where T is the audio sequence length. H_k is an affine transformation for each step. The total loss of training is $L = \Sigma_{k=1}^{K} L_k$, summed over every step.

The wav2vec pretrained model variant "wav2vec large", which we use as a pretrained feature extractor using additional linear transformations and a larger context network, is trained on 1,000 hours of unlabeled English speech with noise. The model's downsampling factor is 160. Thus, there is a 512-dimensional vector for every 10 ms of speech [23]. After training, the representations produced by the context network are input to the spoofing speech detection model instead of the traditional acoustic features.

3. Proposed Method

3.1. Representation Learning

For the representation learning step, the proposed model takes two randomly selected wav2vec features of speech as input. These two embeddings are processed by a Siamese network consisting of a backbone network, such as a light convolutional neural network (LCNN) or ResNet. To learn a powerful and meaningful representation, we optimize the model parameters by minimizing a contrastive loss function, as shown in Figure 1. The contrastive loss is the loss function whose value approaches 0 when the distance between a similar pair of inputs is close to 0 and the distance between a dissimilar pair of inputs exceeds a certain threshold. In this task, similar means both recordings are either from genuine speech or spoofed speech while dissimilar means one recording from genuine speech and another from spoofed speech. Let $f(x_i)$ be an embedding neural network that extracts representation vectors from wav2vec features. Then, we define the distance between two representation vectors as $D_{ij} = \parallel f(x_i) - f(x_j) \parallel$, where $\parallel \cdot \parallel$ denotes the Euclidean distance. For a minibatch of N samples, we define the contrastive prediction task on the pairs of examples derived from the N samples. Let Y be a binary label for each pair. $Y = 0$ if x_i and x_j belong to the same class, and $Y = 1$ if x_i and x_j belong to different classes. Then, the loss function is

$$L_{contrast}(Y, x_i, x_j) = (1-Y)\frac{1}{2}(D_{ij})+(Y)\frac{1}{2}max(0, m - D_{ij}) \tag{3}$$

where margin m is a positive number that makes the dissimilar pairs contribute to the contrastive loss only if the distance between two samples is smaller than m. After representation learning, the Siamese network yields embeddings that keep similar speech datapoints close while pushing dissimilar speech datapoints apart.

3.2. Siamese Network Architecture

In this section, the neural network architectures that we use for representation learning are described. State-of-the-art performance for spoofing detection tasks has been achieved by using LCNN, ResNet, and SENet in recent works, so we use these three network architectures as our backbone networks for Siamese networks.

3.2.1. LCNN

The greatest contribution of the LCNN [24] architecture is the proposed max feature map (MFM) activation function, which is an extension of maxout activation. In this work, we use a 9-layer Light CNN that contains 5 convolution layers, 4 max-pooling layers with kernels of size 2 × 2 and strides of 2, 4 Network in Network (NIN) layers, and Max-Feature-Map layers. In contrast to the classic LCNN structure, we use only the front-end convolutional neural network for feature extraction. In addition, we use a fully connected layer to obtain a 512-dimensional embedding from the high-dimensional original feature. For the next SENet and ResNet models, we take the same approach.

3.2.2. ResNet

Based on the application of ResNet [25] in the field of CV, many studies have also used ResNet in audio spoofing detection [6, 26]. The skip connections in ResNet help preserve the gradient during back-propagation. In this work, we use ResNet18 as an embedding network.

3.2.3. SENet

The main concept behind the SENet architecture is squeeze-and-excitation (SE) blocks [27], which improve the quality of representations by using interdependencies between the channels to automatically emphasize useful features and suppress insignificant features. In this work, we use SENet18, SENet34, and SENet50 as embedding networks.

3.3. Classifier

The embeddings learned by the Siamese network will be used as the input features of classifiers such as support vector machines and MLP for spoofing detection tasks. In this study, we employ a two-layer fully connected (fc) neural network using the representations obtained from the Siamese network. The classifier MLP has batch normalization (BN) [28]applied to its hidden fc layer, and its output fc layer does not have BN. The dimension of the classifier's input is 512, the dimension of the output is 2, and the dimension of the hidden layer is 256.

4. Experimental Setup

The proposed model was evaluated on the LA partition of the ASVspoof 2019 challenge, which was divided into three parts: 54,000 audio samples in the training set, 29,700 audio samples in the development set, and 124,730 audio samples in the evaluation set. The LA dataset contains spoofing audio generated by voice conversion and text-to-speech, and the evaluation set contains unknown attacks generated by different speech synthesis and voice conversion technologies. During the representation learning phase, the training data are randomly divided into minibatches of 64. To maintain a balance of negative and positive samples, we ensured that a minibatch contained half spoofing speech and half genuine speech. Then, we randomly selected 50 pairs of samples from the minibatch for representation learning.

We applied t-SNE [29] to visualize the high-dimensional representation, as shown in Figures 2 and 3. To fairly compare the representation learning capabilities of different network structures, we selected the representation model with the best performance. Representation learning performed best at 50 epochs, and the classifier performed best at 10 epochs.

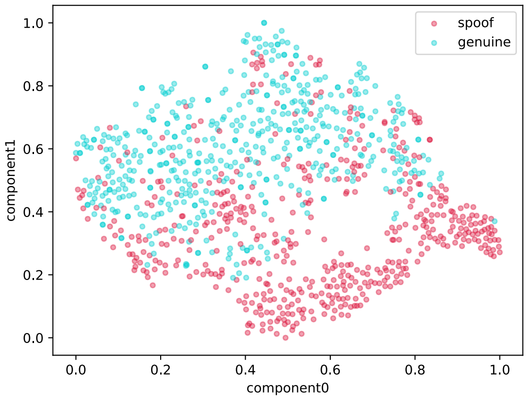

Figure 2: *The t-SNE visualization of the representations on the ASVspoof 2019 LA evaluation dataset after 1 epoch*

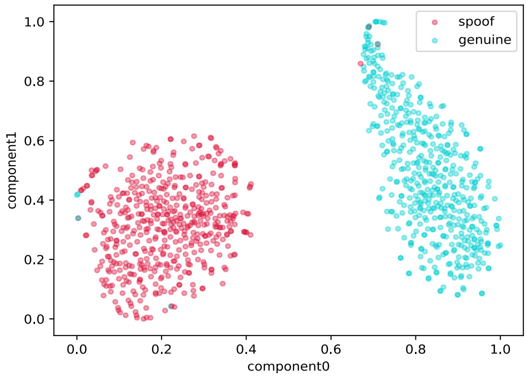

Figure 3: *The t-SNE visualization of the representations on the ASVspoof 2019 LA evaluation dataset after 50 epochs*

Our models were optimized by the Adam optimizer [30] with a two-staged learning rate scheduler for 50 epochs. The first stage was similar to the gradual warmup strategy proposed in [31] and lasted for approximately one and a half epochs. During this stage, the learning rate gradually increased with each iteration. During the second stage, the learning rate slowly decayed with each iteration. Consequently, the learning rate can be described by the following:

$$lr = \alpha \min(\frac{it\_steps + 1}{\gamma}, \frac{1}{\sqrt{it\_steps + 1}}) \quad (4)$$

$$\gamma = N\sqrt{2N} \quad (5)$$

where N is the batch count within one epoch and γ controls the length of the first stage.

5. Results and Analysis

In this section, we compare the results of the state-of-the-art systems with those of the wav2vec features for the LCNN, ResNet, and SENet models in Table 1. To confirm the effect of the Siamese network for representation learning, we also compare the performance of the original network with that of the proposed model based on the Siamese network in Table 2. Finally, the results obtained using different margins are presented in Table 3.

5.1. Comparison of Different Models

Table 1 shows the results of the baseline systems that use the GMM with LFCC and CQCC features provided by the ASVspoof 2019 organizer. We also report other models that achieve state-of-the-art results. We observe that the results of LC-GRNN+LDA, FG-LCNN, and the proposed models are close for the development set that contains only known attacks; however, the performance of the other models decreases significantly when they encounter unseen attacks in the evaluation set. This confirms that the features extracted by the pretrained model can help neural networks generalize to out-of-distribution spoofing attacks. The proposed system wav2vec-SENet34 has the best performance compared with the other models, with a relative improvement of 67% compared with the state-of-the-art system FG-LCNN. LCNN and SeResNet both obtain performance improvements using the wav2vec feature compared with the LFCC and logspec features.

Table 1: *EER(%) for different backbone networks with the wav2vec feature and their comparison with some known single systems on the ASVspoof 2019 logical access evaluation and development set.*

System	Dev	Eval
LFCC-GMM [4]	11.9	13.54
CQCC-GMM [3]	9.87	11.04
LPS-FFT-LCNN [9]	-	4.53
SENet34+logspec [6]	-	9.94
DenseNet+logspec [6]	-	6.78
LC-GRNN+LDA [32]	0.00	6.28
FG-LCNN [33]	0.002	4.07
wav2vec-LCNN	0.009	3.02
wav2vec-ResNet18	0.004	1.62
wav2vec-SENet18	0.003	1.31
wav2vec-SENet34	0.001	**1.21**
wav2vec-SENet50	0.002	1.59

5.2. Comparison of Our Models with the Original Model

Table 2 reports the comparison of the performance of the proposed Siamese network with that of the original network. We compare the LCNN, ResNet, and SENet architectures with cross-entropy loss with the Siamese network approach with contrastive loss. Compared with the original model, the proposed approach obtains an average relative improvement in performance of 13%. Moreover, because Siamese networks share parameters, the proposed method does not add too many new parameters.

Table 2: *Performance of the proposed model and its comparison with the original model on the ASVspoof 2019 logical access evaluation and development set.*

Model	Original		Siamese	
	Dev.	Eval.	Dev.	Eval.
LCNN	0.009	3.02	0.006	2.54
ResNet18	0.004	1.62	0.003	1.33
SENet18	0.003	1.31	0.002	1.25
SENet34	0.001	1.21	0.002	1.16
SENet50	0.002	1.59	0.004	**1.15**

5.3. Comparison of Different Margins

Table 3 presents the results obtained by using different margins. Considering Equation 3, the margin is a very important parameter that defines a radius, and only the pair of samples whose distance within the margin can contribute to the loss function. Based on the results of the experiment, the margin of 2 is better than 1 and 0.5, so margin m is set to 2 for the entire experiment.

Table 3: *EER(%) for representation learning using SENet34 as the backbone network over three different margins*

Different_Margins	Dev.	Eval.
Siamese-SENet34_margin_2	0.002	1.16
Siamese-SENet34_margin_1	0.003	1.24
Siamese-SENet34_margin_0.5	0.015	1.42

6. Conclusions

In this study, we used supervised representation learning with the proposed two-phase Siamese network to protect antispoofing models against LA attacks and found that the wav2vec features extracted from the pretrained model and the Siamese network architectures significantly improve the performance of spoofing speech detection. We evaluated five different backbone networks in the Siamese network with the wav2vec features and showed that the proposed models are able to improve performance without increasing the number of parameters. As a result, through the use of the proposed representation learning system, EER was reduced from 4.07 to 1.15. In future work, we will improve the Siamese network architecture to generate more discriminative representations. We also encourage future work to investigate additional speech pretraining models, such as transformers based on masked predictive coding [34], for spoofing speech detection tasks.

7. Acknowledge

The work was supported by Chinese Technology Innovation2030 Major Project(2020AAA0140004). Yingchun Yang is the corresponding author.

8. References

[1] T. Kinnunen, L. Juvela, P. Alku, and J. Yamagishi, “Non-parallel voice conversion using i-vector PLDA: towards unifying speaker verification and transformation,” in *ICASSP*. IEEE, 2017, pp. 5535–5539.

[2] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Liu, “Fastspeech: Fast, robust and controllable text to speech,” in *NeurIPS*, 2019, pp. 3165–3174.

[3] M. Todisco, H. Delgado, and N. W. D. Evans, “Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification,” *Comput. Speech Lang.*, vol. 45, pp. 516–535, 2017.

[4] M. Sahidullah, T. Kinnunen, and C. Hanilçi, “A comparison of features for synthetic speech detection,” in *INTERSPEECH*. ISCA, 2015, pp. 2087–2091.

[5] X. Xiao, X. Tian, S. Du, H. Xu, E. Chng, and H. Li, “Spoofing speech detection using high dimensional magnitude and phase features: the NTU approach for asvspoof 2015 challenge,” in *INTERSPEECH*. ISCA, 2015, pp. 2052–2056.

[6] C. Zhang, J. Cheng, Y. Gu, H. Wang, J. Ma, S. Wang, and J. Xiao, “Improving replay detection system with channel consistency densenext for the asvspoof 2019 challenge,” in *INTERSPEECH*. ISCA, 2020, pp. 4596–4600.

[7] J. Williams and J. Rownicka, “Speech replay detection with x-vector attack embeddings and spectral features,” in *INTERSPEECH*. ISCA, 2019, pp. 1053–1057.

[8] A. G. Alanís, A. M. Peinado, J. A. González, and A. M. Gomez, “A gated recurrent convolutional neural network for robust spoofing detection,” *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 12, pp. 1985–1999, 2019.

[9] G. Lavrentyeva, S. Novoselov, A. Tseren, M. Volkova, A. Gorlanov, and A. Kozlov, “STC antispoofing systems for the asvspoof2019 challenge,” in *INTERSPEECH*. ISCA, 2019, pp. 1033–1037.

[10] C. Lai, N. Chen, J. Villalba, and N. Dehak, “ASSERT: anti-spoofing with squeeze-excitation and residual networks,” in *INTERSPEECH*. ISCA, 2019, pp. 1013–1017.

[11] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, “Deep contextualized word representations,” in *NAACL-HLT*. Association for Computational Linguistics, 2018, pp. 2227–2237.

[12] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: pre-training of deep bidirectional transformers for language understanding,” in *NAACL-HLT (1)*. Association for Computational Linguistics, 2019, pp. 4171–4186.

[13] K. He, H. Fan, Y. Wu, S. Xie, and R. B. Girshick, “Momentum contrast for unsupervised visual representation learning,” in *CVPR*. IEEE, 2020, pp. 9726–9735.

[14] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in *CVPR*. IEEE Computer Society, 2014, pp. 580–587.

[15] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, “Decaf: A deep convolutional activation feature for generic visual recognition,” in *ICML*, ser. JMLR Workshop and Conference Proceedings, vol. 32. JMLR.org, 2014, pp. 647–655.

[16] S. Schneider, A. Baevski, R. Collobert, and M. Auli, “wav2vec: Unsupervised pre-training for speech recognition,” in *INTERSPEECH*. ISCA, 2019, pp. 3465–3469.

[17] D. Hendrycks, X. Liu, E. Wallace, A. Dziedzic, R. Krishnan, and D. Song, “Pretrained transformers improve out-of-distribution robustness,” in *ACL*. Association for Computational Linguistics, 2020, pp. 2744–2751.

[18] C. Ting, K. Simon, N. Mohammad, and H. Geoffrey, “A simple framework for contrastive learning of visual representations,” *ICML*, pp. 1597–1607, 2020.

[19] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li, “Imagenet: A large-scale hierarchical image database,” in *CVPR*. IEEE Computer Society, 2009, pp. 248–255.

[20] N. Reimers and I. Gurevych, “Sentence-bert: Sentence embeddings using siamese bert-networks,” *EMNLP/IJCNLP (1)*, pp. 3980–3990, 2019.

[21] A. Kratzsch, W. Kästner, and R. Hampel, “Modelling the differential pressure at sieves with artificial neural networks (multilayer perceptron) - a contribution to reactor safety research,” in *EUSFLAT Conf. (1)*. Universitas Ostraviensis, 2007, pp. 469–472.

[22] R. Hadsell, S. Chopra, and Y. LeCun, “Dimensionality reduction by learning an invariant mapping,” in *CVPR (2)*. IEEE Computer Society, 2006, pp. 1735–1742.

[23] A. van den Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” *CoRR*, vol. abs/1807.03748, 2018.

[24] X. Wu, R. He, Z. Sun, and T. Tan, “A light CNN for deep face representation with noisy labels,” *IEEE Trans. Inf. Forensics Secur.*, vol. 13, no. 11, pp. 2884–2896, 2018.

[25] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *CVPR*. IEEE Computer Society, 2016, pp. 770–778.

[26] M. Alzantot, Z. Wang, and M. B. Srivastava, “Deep residual neural networks for audio spoofing detection,” in *INTERSPEECH*. ISCA, 2019, pp. 1078–1082.

[27] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” *CoRR*, vol. abs/1709.01507, 2017.

[28] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *ICML*, ser. JMLR Workshop and Conference Proceedings, vol. 37. JMLR.org, 2015, pp. 448–456.

[29] v. d. L. Maaten and G. Hinton, “Visualizing data using t-sne,” *JOURNAL OF MACHINE LEARNING RESEARCH*, pp. 2579–2605, 2008.

[30] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[31] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He, “Accurate, large minibatch sgd: Training imagenet in 1 hour,” *arXiv preprint arXiv:1706.02677*, 2017.

[32] A. G. Alanís, A. M. Peinado, J. A. González, and A. M. Gomez, “A light convolutional GRU-RNN deep feature extractor for ASV spoofing detection,” in *INTERSPEECH*. ISCA, 2019, pp. 1068–1072.

[33] Z. Wu, R. K. Das, J. Yang, and H. Li, “Light convolutional neural network with feature genuinization for detection of synthetic speech attacks,” in *INTERSPEECH*. ISCA, 2020, pp. 1101–1105.

[34] D. Jiang, X. Lei, W. Li, N. Luo, Y. Hu, W. Zou, and X. Li, “Improving transformer-based speech recognition using unsupervised pre-training,” *CoRR*, vol. abs/1910.09932, 2019.