A Appendix A

A.1 Detailed explanation of continuous nature of similarity

In this section, we expand on our observation that similarity between training samples is not binary. Consider the images shown in Figure 6. Let the anchor image and the four images at the bottom be part of a batch of training data (possibly along with many other samples). Note that the similarity of the anchor image ranges from 'very similar' to 'highly dissimilar' and that it is not simply binary. However, Existing methods for contrastive training only use a binary notion for similarity, and categorize the samples in a batch into "positive" and "negative" sets. As a consequence, the models fail to correctly learn associations between different data samples.

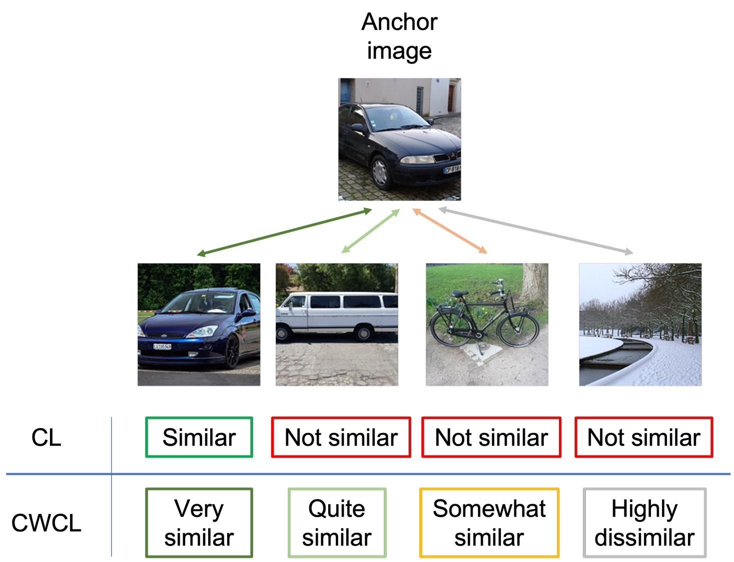

Figure 6: An illustration of the continuous nature of similarity in training data. In this example, the anchor image is similar (or dissimilar) to the other images to various degrees. Existing methods choose a subset of images and consider them to be 'positive examples', and consider the rest of the examples as 'negative' examples. Once these subsets are chosen, the embedding of the anchor image is *aligned* (to an equal degree) to those of the positive examples and *contrasted* with those of the negative examples. As a consequence, any similarity between the anchor image and the so-called 'negative' examples is completely ignored. Further, all 'positive' examples are considered to be *equally similar*, although this might not be the case.

A.2 Experimental details for aligning image and text modalities

A.2.1 Model training details

We build upon the code repository in [50]. We train our models for a total of 70 epochs, where each epoch uses a subset of 6 million images,. The batch size is set to 16000. Note the number of training steps in this case is equal to 26,250. We train on 4 A100 GPUs. Note that we experimented with different sizes for the subset used in each epoch (ranging from 6 million to the full dataset) and we obtained the best performance when the size was 6 million (for our method and the baseline methods that we train). We use a learning rate of 0.001, AdamW optimizer with $\beta_1 = 0.9, \beta_2 = 0.999$ and a weight decay of 0.0001 [51].

A.2.2 Simple templates to test model robustness

One of the advantages of cross-modal 0-shot transfer is the ability of the trained models to be used on downstream tasks without any further training. However, the downstream task still needs to be adapted to the task of modality alignment. We discuss this adaptation in the context of image classification and provide details about our experiments reported in Section 4.1.2.

Table 3: Simple template sentences that we use to generate classifier embeddings.

a photo of a { }
an image of a { }
a picture of a { }
this is a { }
a snap of { }
a shot of { }
an illustration of { }
an example of { }
a { } is pictured here
In this picture, we can see a { }

Table 4: CWCL improves upon the CL-based alignment method for image-text retrieval.

Method	I →T retrieval			T→I retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
CL	30.42	54.32	65.82	24.17	49.04	61.05
CWCL (Ours)	**35.10**	**61.52**	**73**	**25.69**	**50.04**	**61.59**

In [1, 2], the downstream task of image classification task is solved by first changing the class labels to sentences. The sentences are then converted to embeddings using the text encoder. Given a test image, the text embedding that it aligns the most with determines its class. In particular, both works use a set of 80 "template sentences" to convert each label to 80 sentences. The text embedding representing a given label is then computed as the average of the embeddings of these 80 sentences.

We observe that the classification accuracy depends on the choice of these template sentences, as also seen in [5]. To illustrate this, we formulate $k = 1, 5, 10$ **simple** template sentences and use them to generate the classifier embeddings. We list these sentence in Table 3. Note that for $k = 1$, we use the first sentence only and for $k = 5$, we use the first 5 sentences. Our motivation in choosing simple sentences is to mimic the process of an end user who may not have the resources to carefully design the template sentences. Our goal is to test our model's robustenss under such a scenario. As shown in Figure 5, a model trained using standard contrastive tuning shows poor performance as the number of template sentences is reduced. This shows that to achieve high accuracy, an end user must design template sentences that are complex enough. However, a model trained using CWCL maintains its performance across varying number of template sentences, even when only simple templates are used. Our hypothesis is that owing to the continuous nature of the similarity used during training, the model has learnt better cross-modal associations.

A.2.3 Cross-modal retrieval

We also examine the 0-shot image-text retrieval capabilities of our proposed method. Note that our experiments are only towards comparing standard contrastive loss with CWCL. We leave the task of training with larger datasets [1, 2, 3] and using multi-objective training (which maybe used in conjuntion with contrastive tuning to obtain better retrieval performance) [30, 25, 18] for future exploration.

In our experiment, we simply compare the performance of models trained with contrastive loss (as done in [2]) to that of models trained using CWCL. We use the MS-COCO validation dataset [52] to study zero-shot retrieval performance of these models. We report our results in Table 4. Models trained with CWCL outperform those trained using the standard contrastive loss function.

A.3 Speech-Text Appendix

In this section, we provide additional details about training the speech-text alignemnt models.

A.3.1 Model training details

We train each model for a total of 20 epochs, where one epoch consumes the whole training data equal to 1,013,630 samples. We use a batch size of 20 with the 12,500 warmup steps and train on 1 A100 GPU. We use a learning rate of 0.00003, AdamW optimizer with $\beta_1 = 0.9, \beta_2 = 0.999$, a weight decay of 0.0001, and gradient clipping norm of 10.

A.3.2 Effects of using pre-trained model weights, locking location, and batch size

Each reported number in this section is Top-1 accuracy (%) on SLURP data for speech intent classification.

Start the speech model from scratch VS pre-trained weights: In Table 5, we compared starting multi-modal training from scratch and from pre-trained weights. The performance is significantly boosted by initializing the speech encoder using weights from the encoder part of the Whisper ASR model [42]. However, regardless of using random weights and pre-trained model weights, training with CWCL results in a much better downstream performance.

Table 5: Comparison between using randomly initialized weights and pre-trained weights for speech encoders during training: Top-1 accuracy (%) on SLURP data

Method	Random initialization	Pre-trained weights
CL	13.80	22.73
CWCL	**26.17**	**53.12**

Locking location: We have 4 ways to lock our model during multi-modal training since we have pre-trained speech and text models. We compared all the locking options and the result is shown in Table 6. In both baseline and CWCL losses, locking the text model works best. This can be seen as transferring the knowledge of semantic relationships in text models to speech models.

Table 6: Locking location vs. performance: Top-1 accuracy (%) on SLURP data

Locking location	none	speech	text	both
CL	**18.77**	7.89	24.03	9.82
CWCL	17.50	**27.39**	**53.12**	**16.70**

Batch size vs. performance: Since the large batch size was shown to improve performance with contrastive loss in computer vision, we also did a similar experiment to see how the batch size affects the performance as it gets larger. As the batch size increases, we also increased the learning rate proportionally, e.g., if bs=20 has lr=1, bs=40 has lr=2. The results are in Table 7.

Table 7: Effective batch size vs. performance: Top-1 accuracy (%) on SLURP data

batch size	20	40	80
CL	24.03	25.20	24.51
CWCL	**53.12**	**53.94**	**51.80**

A.3.3 Further evidence of modality alignment due to CWCL

In Figure 3, we showed the alignment (measured as inner product) between speech features and text features obtained from models trained using just CL and those trained using CWCL. We use the speech and text data from the SLURP test dataset. We illustrated that speech and text embeddings that belong to the same intent class were much more aligned compared to speech and text from mismatched classes. In this section, we provide more examples that support this observation.

In Figures 7, 8, we show the alignment between the speech and text embeddings where the speech and text samples belong to classes other than those used in Figure 3. We again see that the alignment between samples in the same class is much higher than that between samples in different classes. In general, we observe the same pattern to hold across all the classes in the dataset, thus confirming that our results are not due to sampling bias.

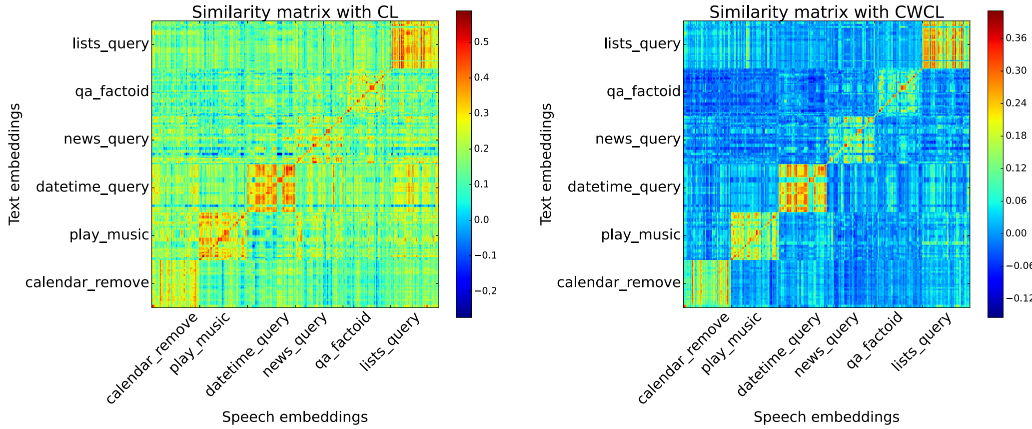

Figure 7: Cosine-similarity between speech and text embeddings obtained by sampling 6 classes randomly from the SLURP test dataset. The block diagonal structure of the matrix on the right shows that using CWCL results in a strong alignment between speech (and text) samples that share a similar intent. In this case, the sampled classes are different from those used in Figures 3 and 8.

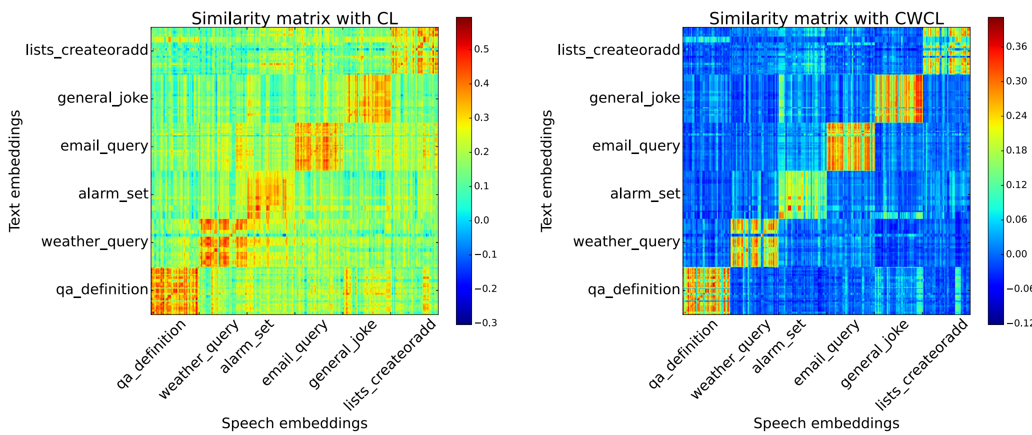

Figure 8: Another example of alignment between speech and text embeddings. The sampled classes are different from those used in Figures 3 and 7.

A.3.4 Additional tables for reference

Table 8 additionally shows the Top-5 accuracy over speech-text experiments. Since most of the previous works did not report this metric, we only include our own experimental results. Table 9 shows the existing supervised model performances where the models are either trained or fine-tuned on the labeled Google Speech Command Dataset V2 for performing the keyword spotting (KWS) task, while our methods did not require any labeled KWS data for performing the task.

A.3.5 General template as a python list

To test the speech-text alignment models, we use a “general" set of templates in addition to the one obtained by using the text from the training data itself. This general set of templates aims to mimic a scenario where the no example texts maybe available. We list the set of template sentences used in the general set here.

Table 8: Top-5 accuracy for zero-shot speech-to-intent classification (SLURP and STOP) and KWS on Google Speech Command Dataset V2. Superscript # is used to indicate use of general templates.

Method	Text model	SLURP	SLURP#	STOP	STOP#	KWS	KWS#
CL	RoBERTa+S	69.57	49.86	98.19	94.03	82.22	82.53
CL	BART+Y	52.97	24.87	95.27	81.63	84.02	78.14
CWCL (Ours)	RoBERTa+S	84.53	68.58	99.38	96.52	91.20	92.42
CWCL (Ours)	BART+Y	79.48	57.34	99.48	97.71	93.79	94.30
Text-intent	RoBERTa+S	95.66	83.36	98.93	95.20	100	98.20
(upper bound)	BART+Y	99.58	73.82	99.45	98.40	100	100

Table 9: Keyword spotting Top-1 accuracies on Google Speech Command Dataset V2 from existing supervised models.

Method	KWS
Attention RNN [47]	93.9
KWT-2 [41]	97.74
Wav2Vec2 [48]	96.6
M2D [49]	95.4
M2D - Fine tuned [49]	98.5

General template sentences: `[it is about { }, it was about { }, it will be about { }, this is about { }, this was about { }, this will be about { }, it is related to { }, it was related to { }, it will be related to { }, this is related to { }, this was related to { }, this will be related to { }, it is talking about { }, it was talking about { }, it will be talking about { }, this is talking about { } this was talking about { }, this will be talking about { }, I am talking about { }, I was talking about { }, I will be talking about { }, You are talking about { }, You were talking about { }, You will be talking about { }, They are talking about { }, They were talking about { }, They will be talking about { }, We are talking about { }, We were talking about { }, We will be talking about { }, it talks about { }, it talked about { }, it will talk about { }, this talks about { }, this talked about { }, this will talk about { }, I talk about { }, I talked about { }, I will talk about { }, You talk about { }, You talked about { }, You will talk about { }, They talk about { }, They talked about { }, They will talk about { }, We talk about { }, We talked about { }, We will talk about { }]`

References

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[2] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer, "Lit: Zero-shot transfer with locked-image text tuning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18123–18133.

[3] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 4904–4916.

[4] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu, "Coca: Contrastive captioners are image-text foundation models," *arXiv preprint arXiv:2205.01917*, 2022.

[5] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao, "Test-time prompt tuning for zero-shot generalization in vision-language models," in *Advances in Neural Information Processing Systems*, 2022.

[6] Basil Mustafa, Carlos Riquelme Ruiz, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby, "Multimodal contrastive learning with limoe: the language-image mixture of experts," in *Advances in Neural Information Processing Systems*, 2022.

[7] Emanuele Bastianelli, Andrea Vanzo, Pawel Swietojanski, and Verena Rieser, "Slurp: A spoken language understanding resource package," *arXiv preprint arXiv:2011.13205*, 2020.

[8] Pete Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," *arXiv preprint arXiv:1804.03209*, 2018.

[9] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz, "Contrastive learning of medical visual representations from paired images and text," in *Machine Learning for Healthcare Conference*. PMLR, 2022, pp. 2–25.

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Everest Hinton, "A simple framework for contrastive learning of visual representations," 2020.

[11] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He, "Improved baselines with momentum contrastive learning," *arXiv preprint arXiv:2003.04297*, 2020.

[12] Yash Patel, Yusheng Xie, Yi Zhu, Srikar Appalaraju, and R Manmatha, "Simcon loss with multiple views for text supervised semantic segmentation," *arXiv preprint arXiv:2302.03432*, 2023.

[13] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang, "Clap learning audio concepts from natural language supervision," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[14] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan, "Supervised contrastive learning," *Advances in neural information processing systems*, vol. 33, pp. 18661–18673, 2020.

[15] Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello, "Wav2clip: Learning robust audio representations from clip," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 4563–4567.

[16] Zachary Novack, Saurabh Garg, Julian McAuley, and Zachary C Lipton, "Chils: Zero-shot image classification with hierarchical label sets," *arXiv preprint arXiv:2302.02551*, 2023.

[17] Jonatas Wehrmann, Camila Kolling, and Rodrigo C Barros, "Adaptive cross-modal embeddings for image-text alignment," in *Proceedings of the AAAI conference on artificial intelligence*, 2020, vol. 34, pp. 12313–12320.

[18] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," *Advances in neural information processing systems*, vol. 34, pp. 9694–9705, 2021.

[19] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Ce Liu, Lu Yuan, and Jianfeng Gao, "Unified contrastive learning in image-text-label space," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19163–19173.

[20] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel, "Audioclip: Extending clip to image, text and audio," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 976–980.

[21] Ilaria Manco, Emmanouil Benetos, Elio Quinton, and George Fazekas, "Contrastive audio-language learning for music," in *Ismir 2022 Hybrid Conference*, 2022.

[22] Rong Ye, Mingxuan Wang, and Lei Li, "Cross-modal contrastive learning for speech translation," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Seattle, United States, July 2022, pp. 5099–5113, Association for Computational Linguistics.

[23] Lu Wu, Chenyu Wu, Han Guo, and Zhihao Zhao, "A cross-modal alignment for zero-shot image classification," *IEEE Access*, vol. 11, pp. 9067–9073, 2023.

[24] Zhiqiu Lin, Samuel Yu, Zhiyi Kuang, Deepak Pathak, and Deva Ramana, "Multimodality helps unimodality: Cross-modal few-shot learning with multimodal models," *arXiv preprint arXiv:2301.06267*, 2023.

[25] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al., "Flamingo: a visual language model for few-shot learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 23716–23736, 2022.

[26] Mohammadreza Zolfaghari, Yi Zhu, Peter Gehler, and Thomas Brox, "Crossclr: Cross-modal contrastive learning for multi-modal video representations," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1450–1459.

[27] Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi, "A theoretical analysis of contrastive unsupervised representation learning," *arXiv preprint arXiv:1902.09229*, 2019.

[28] Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka, "Debiased contrastive learning," *Advances in neural information processing systems*, vol. 33, pp. 8765–8775, 2020.

[29] Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer, "How to train your vit? data, augmentation, and regularization in vision transformers," *arXiv preprint arXiv:2106.10270*, 2021.

[30] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia, July 2018, pp. 2556–2565, Association for Computational Linguistics.

[31] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut, "Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3558–3568.

[32] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li, "Yfcc100m: The new data in multimedia research," *Communications of the ACM*, vol. 59, no. 2, pp. 64–73, 2016.

[33] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[34] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar, "Do imagenet classifiers generalize to imagenet?," in *International conference on machine learning*. PMLR, 2019, pp. 5389–5400.

[35] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing, "Learning robust global representations by penalizing local predictive power," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[36] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al., “The many faces of robustness: A critical analysis of out-of-distribution generalization,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8340–8349.

[37] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song, “Natural adversarial examples,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15262–15271.

[38] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz, “Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models,” *Advances in neural information processing systems*, vol. 32, 2019.

[39] Libo Qin, Tianbao Xie, Wanxiang Che, and Ting Liu, “A survey on spoken language understanding: Recent advances and new frontiers,” *arXiv preprint arXiv:2103.03095*, 2021.

[40] Jerome R Bellegarda, “Spoken language understanding for natural interaction: The siri experience,” *Natural Interaction with Robots, Knowbots and Smartphones: Putting Spoken Dialog Systems into Practice*, pp. 3–14, 2013.

[41] Axel Berg, Mark O’Connor, and Miguel Tairum Cruz, “Keyword transformer: A self-attention model for keyword spotting,” in *Interspeech*, 2021.

[42] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, “Robust speech recognition via large-scale weak supervision,” *arXiv preprint arXiv:2212.04356*, 2022.

[43] Jack G. M. FitzGerald, Christopher Leo Hench, Charith S. Peris, Scott Mackie, Kay Rottmann, A. Patricia Domínguez Sánchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, L. Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and P. Natarajan, “Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages,” *ArXiv*, vol. abs/2204.08582, 2022.

[44] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber, “Common voice: A massively-multilingual speech corpus,” *arXiv preprint arXiv:1912.06670*, 2019.

[45] Paden Tomasello, Akshat Shrivastava, Daniel Lazar, Po-Chun Hsu, Duc Le, Adithya Sagar, Ali Elkahky, Jade Copet, Wei-Ning Hsu, Yossi Adi, et al., “Stop: A dataset for spoken task oriented semantic parsing,” in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 991–998.

[46] Siddhant Arora, Siddharth Dalmia, Pavel Denisov, Xuankai Chang, Yushi Ueda, Yifan Peng, Yuekai Zhang, Sujay Kumar, Karthik Ganesan, Brian Yan, et al., “Espnet-slu: Advancing spoken language understanding through espnet,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7167–7171.

[47] Douglas Coimbra De Andrade, Sabato Leo, Martin Loesener Da Silva Viana, and Christoph Bernkopf, “A neural attention model for speech command recognition,” *arXiv preprint arXiv:1808.08929*, 2018.

[48] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *Advances in neural information processing systems*, vol. 33, pp. 12449–12460, 2020.

[49] Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Noboru Harada, and Kunio Kashino, “Masked modeling duo: Learning representations by encouraging both networks to model the input,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[50] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt, “Openclip,” July 2021, If you use this software, please cite it as below.

[51] Ilya Loshchilov and Frank Hutter, “Decoupled weight decay regularization,” in *International Conference on Learning Representations*, 2019.

[52] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick, "Microsoft coco captions: Data collection and evaluation server," *arXiv preprint arXiv:1504.00325*, 2015.