ADVERSARIAL INPUT ABLATION FOR AUDIO-VISUAL LEARNING

David Xu, David Harwath

Department of Computer Science, The University of Texas at Austin
Austin, Texas, 78712, USA
{david.duanmu.xu,harwath}@utexas.edu

ABSTRACT

We present an adversarial data augmentation strategy for speech spectrograms, within the context of training a model to semantically ground spoken audio captions to the images they describe. Our approach uses a two-pass strategy during training: first, a forward pass through the model is performed in order to identify segments of the input utterance that have the highest similarity score to their corresponding image. These segments are then ablated from the speech signal, producing a new and more challenging training example. Our experiments on the SpokenCOCO dataset demonstrate that when using this strategy: 1) content-carrying words tend to be ablated, forcing the model to focus on other regions of the speech; 2) the resulting model achieves improved speech-to-image retrieval accuracy; 3) the number of words that can be accurately detected by the model increases.

Index Terms— visually grounded speech, self-supervised representation learning, adversarial training

1. INTRODUCTION AND RELATED WORK

Self-supervised representation learning for speech audio has attracted significant research attention in recent years. A primary motivation for this research direction is that self-supervised learning objectives enable models to learn useful representations from un-annotated data, which can then be repurposed for other downstream tasks. Various self-supervised training objectives have been proposed, such as predicting future speech frames conditioned on past frames [1, 2], in-painting regions of an input spectrogram which have been masked [3, 4], BERT-style [5] masked language modelling applied to a tokenized representation of the speech derived from clustering [6, 7], and learning to associate speech waveforms with context derived from a separate modality, such as vision [8]. These efforts have led to the creation of shared benchmarks to drive continued progress [9, 10].

One line of self-supervised learning for speech has shown that it is possible to train neural models to directly associate spoken image captions with the visual scenes that they describe, without the need for text transcripts or supervised automatic speech recognition (ASR) [11]. These models of Visually-Grounded Speech (VGS) have been shown to implicitly learn to predict the presence of individual words in the speech waveform, as well as detectors for visual objects and properties such as color, size, and material [12, 13, 14, 15]. The representations learned by VGS models have proven to be useful for improving the robustness of supervised ASR models [16, 17], various zero-resource speech processing tasks [18], and semantic image search systems [19].

In this paper, we introduce a novel form of data augmentation in the context of a VGS model trained to retrieve images given their spoken descriptions. Our approach is based on *adversarial input masking*, in which we use a current snapshot of the model weights to identify segments of an input spectrogram that would cause the greatest decrease in the similarity between an image and its description were they to be masked out. The ablated version of the spectrogram is then used as an additional training example, forcing the model to learn to pay attention to regions of the audio signal that were previously ignored. Our approach is motivated by the observation previously published VGS models tend to focus on a small number of audio regions which generally contain words that are highly predictive of the image content being described [20, 15]. Other regions of the audio tend to be ignored, despite the fact that they still may contain words that describe the image scene. We show that the accuracy of a speech-image retrieval system improves when training on examples that have been masked in our proposed fashion. Our analysis also shows that the adversarial masking strategy tends to ablate spectrogram regions containing content-carrying words and that models trained with this form of data augmentation learn to reliably detect a larger number of unique word types than models trained with randomly selected masks.

Data augmentation strategies that rely on spectrogram masking are commonly used in current state-of-the-art speech recognition systems, with SpecAugment [21] being a popular method. SpecAugment is a simple augmentation strategy that randomly selects temporal segments and frequency bands of an input spectrogram to mask out, forcing the model to more broadly utilize the information in the speech signal rather than overfitting to acoustic cues which may, for example, occur in a narrow frequency band. Adversarial data augmentation strategies have also been explored in the context of making

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

speech recognition systems robust to adversarial attacks, in which an adversary corrupts an input waveform with an imperceptible amount of noise that exploits brittleness in the model's decision boundaries [22]. While our setting does not assume the presence of an adversarial attacker, it does assume a "lazy" model that only focuses on a subset of the words contained in the input speech. We find that adversarial masking is effective under these circumstances, and outperforms a random masking strategy.

2. TECHNICAL APPROACH

2.1. Task and Baseline Model

We study the task of image-speech retrieval. In this setting, a model takes as input a speech waveform describing the content of a desired image. The model then searches over a library of candidate images and returns the closest matches to the query speech, according to a similarity function that is computed between the model's representations of the input speech and each image. Alternatively, we can use images as queries and spoken descriptions serve as the items to be retrieved. Performance is measured on a held-out test set of unseen images and descriptions using Recall at K (R@K), representing the probability that the ground-truth target item matching a given query was in the top K retrieved results. For model training, a set of images paired with their spoken descriptions are assumed to be available.

We use the ResDAVEnet/Resnet-50 dual-encoder model proposed by [23], which uses a pair of fully-convolutional neural network encoders that map speech waveforms and visual images into a shared, multimodal embedding space. We represent each 16kHz speech waveform as a 80-dimensional log-mel spectrogram computed using 25ms Hamming windows and a 10ms shift, truncated or zero-padded as needed to 2048 frames. All utterances in the dataset are less than 2048 frames long [24], so our truncation/padding strategy guarantees that no information is lost from any utterance. The spectrogram is fed as input to the audio network to produce an output feature map $A \in \mathbb{R}^{T \times D}$. D represents the dimensionality of the shared embedding space, and T represents the number of timesteps in the output, which corresponds to a downsampling of the input spectrogram's framerate by a factor of 16. We follow the same image resizing and cropping strategy as [23], and represent each input image as a 224-by-224 pixel RGB array. Using the image encoder network, an input image is mapped to a feature map $I \in \mathbb{R}^{W \times H \times D}$, whose width W and height H will both be 7 for an input image size of 224.

To compute the similarity score between a image and a speech waveform, we first perform global average pooling, applied over the time dimension for the speech output A and over the spatial dimensions for the image output I. This results in a pair of D-dimensional vectors $\bar{I}$ and $\bar{A}$, and we take their inner product to compute the similarity score $\bar{I}^T \bar{A}$. At retrieval time, a query can be scored against all candidate items in the library, the resulting scores sorted, and the top scoring items retrieved. At training time, we optimize the Masked Margin Softmax [25] loss. Assume we have a training batch of B image-speech pairs, where the model's output for the i^{th} pair is denoted by $\bar{I}_i$ and $\bar{A}_i$. We then compute the similarity score for every image in the batch with every caption in the batch, $Z_{ij} == \bar{I}_i^T \bar{A}_j$. The overall loss training loss is $L = L_{IA} + L_{AI}$, where

$$L_{IA} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{e^{\mathbf{Z}_{ii}-\delta}}{e^{\mathbf{Z}_{ii}-\delta} + \sum_{j=1}^{B} \mathbf{M}_{ij} e^{\mathbf{Z}_{ij}}} \qquad (1)$$

$$L_{AI} = -\frac{1}{B}\sum_{j=1}^{B} \log \frac{e^{\mathbf{Z}_{jj}-\delta}}{e^{\mathbf{Z}_{jj}-\delta} + \sum_{i=1}^{B} \mathbf{M}_{ij} e^{\mathbf{Z}_{ij}}} \qquad (2)$$

$$\mathbf{M}_{ij} = \begin{cases} 0, & \text{if the } i^{th} \text{ image is described by the } j^{th} \text{ caption} \\ 1, & \text{otherwise} \end{cases}$$

Here, δ is a margin hyperparameter, which we fix at 1. Because more than one ground-truth caption may exist for an image within a training batch, $\mathbf{M}$ is used to prevent these captions from interfering with one another as negative examples.

2.2. Proposed Spectrogram Ablation Strategies

When no ablation is applied, the inputs to the ResDAVEnet audio branch are will contain the entire contents of their spoken descriptions, with the exception of any zero-padding frames. In this section, we propose several ablation-based data augmentation strategies that erase temporal segments within an input spectrogram (depicted in Figure 2.2) [1].

Adversarial ablation based on oracle word boundaries. Given that our motivation is to remove entire words from the input spectrogram that the model becomes over-reliant on, we first explore an oracle approach in which we assume that we have knowledge of the ground-truth word boundaries. To derive these, we force-align the SpokenCOCO caption transcripts with the speech waveforms using the Kaldi [26] toolkit. During training, we use the following step-by-step ablation strategy: **1)** Perform a forward pass through the model with the image and its non-ablated spectrogram to obtain $\bar{I}$ (with global average pooling applied) and A (*without* global average pooling applied). **2)** Let $\mathbb{A} = \{A^1, A^2, \dots A^{N_w}\}$, where each A^i is temporally aligned with the i^{th} word in the ground-truth transcription of the speech. Note that A has a temporal downsampling factor of 16 with respect to the duration of the input spectrogram, which we accomodate by simply dividing the spectrogram-frame word alignments by 16. Next, apply temporal average pooling to each tensor segment, obtaining a sequence of D dimensional vectors $\bar{A}^1, \bar{A}^2, \dots \bar{A}^{N_w}$. **3)** Compute per-word similarity scores with the image $S^1, S^2, \dots, S^{N_w}$ where

[1] Implementation and model weights can be found at https://github.com/DavidXu9000/AdversarialAblation

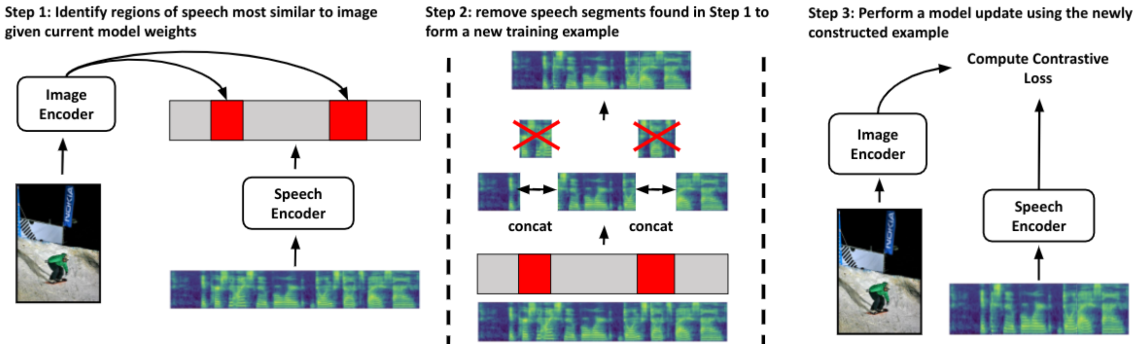
Step 1: Identify regions of speech most similar to image given current model weights
Image Encoder
Speech Encoder
Step 2: remove speech segments found in Step 1 to form a new training example
concat
concat
Step 3: Perform a model update using the newly constructed example
Compute Contrastive Loss
Image Encoder
Speech Encoder

Fig. 1. Overview of training using our proposed adversarial input ablation strategy.

$S^i = \bar{I}^T \bar{A}^i$. **4)** Choose the top k tensor segments with the highest similarity scores S^i. For each of these, decide whether or not to ablate them using a Bernoulli random variable with parameter p. Let the set of tensor segments that will be ablated be $\mathbb{A}'$. **5)** Map each segment in $\mathbb{A}'$ back to spectrogram-frames by multiplying the start and end indices by 16. Completely remove the spectrogram frames between these intervals, stitching together the remaining spectrogram frames along the temporal axis. Using this ablated spectrogram re-padded to 2048 frames, perform a forward pass through the model to obtain $\hat{A}$, and use it in place of A when computing the loss according to Equations 1 and 2.

Unsupervised adversarial ablation based on frame selection. In the nominal case, oracle word boundary information is not available. Thus, rather than following the previously outlined ablation strategy that uses ground-truth word alignments, we select the k indices of the vector $A\bar{I}$ that have the largest values, which correspond to the indices of the temporal frames of A with the highest similarity to $\bar{I}$. We map these selected frames back to the same resolution as the input spectrogram by multiplying them by 16, resulting in a set of frame indices $\{c_1, \ldots, c_k\}$. We then use use these indices as the centers of a set of k segments $\{C_1, \ldots, C_k\}$, where the segment C_i spans from the spectral frame at index $c_1 - w_i$ to the spectral frame at index $c_1 + w_i$. w_i is independently chosen for each segment using a uniform distribution over the interval $[12, 25]$. We independently decide whether or not to keep each segment C_i using a Bernoulli random variable with parameter p. The segments chosen for ablation are spliced out of the spectrogram, resulting in a new training example. This ablated sample is re-padded to 2048 frames and then used in a forward pass through the model to compute $\hat{A}$, which is then used to compute the losses in Equations 1 and 2.

Random ablation based on frame selection. This strategy is the same as unsupervised adversarial ablation, except we simply choose the center frames of the spectrogram intervals to be ablated from a uniform random distribution over the spectrogram frames (not including zero-padding frames).

3. EXPERIMENTS

3.1. Dataset and Implementation Details

We perform our experiments on the SpokenCOCO dataset [24], which contains spoken versions of the MSCOCO [27] text captions. The spoken captions were collected by presenting Amazon Mechanical Turkers with the text captions and asking them to record themselves reading the captions out loud. The dataset contains approximately 123k images, each with 5 spoken captions. In our experiments, we use the training and validation splits defined by the 2017 COCO evaluation.

For the visual encoder, we use the ImageNet pre-trained weights for the ResNet-50 model. The audio encoder is not pre-trained. For frame-based and random ablation, we conducted a grid search on the dropout rate p and k, the number of segments to drop. Our search for p was over the intervals $[0.15, .60]$ with a .15 increment, and $[1, 3]$ for k. Models were trained with an initial learning rate set to 0.0002, a batch size of 128, and a learning rate decay of .985 every 3 epochs. The best hyperparameter settings for both frame-based ablation and random ablation was $p = 0.45$ and $k = 2$.

3.2. Image-Speech Retrieval

In Table 1, we summarize the performance of the baseline model and models trained with our ablation strategies. The SpokenCOCO validation dataset consists of 5000 images with 5 corresponding captions per image. We split the validation set into 25 subsets of 1000 image-audio pairs, such that no two captions in the same subset correspond to the same image. All models were evaluated on the same 25 subsets, and we report the averaged recall scores across these subsets.

The results in Table 1 show that models trained with ablation of the spectrogram outperform the baseline model on R@1, R@5, and R@10 metrics. The frame-based ablation strategy shows the largest improvement with a 3% absolute improvement in the R@10 metric. This demonstrates that ablating the spectrogram in an adversarial manner improves the model's generalization ability. We now turn our attention to analyzing the types of words that tend to be selected for ablation, as well as the impact of our training strategy on the model's ability to detect the presence of particular words.

3.3. Analysis of Ablated Intervals

To confirm that our adversarial ablation strategies are selecting frames contained in content-carrying words, we compute statistics over the words containing the frames that would have been selected for ablation during a single epoch, using model fully trained with the adversarial frame-based selection strategy. We display a histogram over the 20 most commonly ablated words in Table 2, all of which are nouns corresponding to commonly occurring objects in the MSCOCO images.

Next, using the baseline model (with no adversarial ablation applied), for all training examples we compute the average similarity score between the image and each word in its paired caption. We average the similarity scores across all instances of each word type. We then perform a single epoch of training using frame-based adversarial ablation, and recompute the per-word average similarity scores. The top 20 words that experienced the largest absolute change in similarity scores are displayed in Table 3. We note that these words are also predominantly nouns, and tend to be less generic (e.g. "panda", "chocolate") than the nouns appearing in Table 2.

Table 2. Histogram of words that most frequently contained the frame chosen for ablation using the unsupervised frame-based ablation strategy.

word	count	word	count
cat	12022	baseball	9153
tennis	11675	bus	8765
woman	11623	kitchen	8724
man	11194	plate	8530
train	11003	bathroom	8480
dog	10676	water	8224
pizza	10666	bed	7956
street	10085	beach	7530
people	9565	toilet	7363
table	9534	sign	7313

Table 3. Words with the highest increase in image similarity scores between frame-based and baseline models.

word	Δ sim. score	word	Δ sim. score
panda	47.861	photo	29.248
pictures	47.464	cat	27.995
window	47.439	chocolate	27.364
pickup	43.535	smokes	25.848
smoke	40.681	palm	25.174
ripe	37.38	prop	25.012
horsed	35.839	policeman	24.723
papers	34.74	cow	23.267
we	32.566	outdoor	23.042
row	32.309	suitcase	21.95

3.4. Word Presence Detection with Linear Probes

Finally, we perform a quantitative analysis to demonstrate that our models trained with adversarial ablation are able to more reliably detect the presence of different words within an utterance than the baseline models. We train a set of linear classification probes for every residual block of the ResDAVEnet model to detect the presence or absence of a particular target word in an utterance, using the ground-truth word transcriptions. To select target words, we consider all words with a minimum frequency of 100 in the training set and select the 1000 words with the largest Inverse Document Frequency (IDF) score, treating each spoken caption as a document. Each probe consists of a gradient blocker followed by a single linear neuron with a sigmoid activation, trained using binary cross entropy. For each residual block, we train 1000 probes, one for each target word. The input features of the probes are the average-pooled outputs of its corresponding block. Since the audio inputs are zero-padded to a fixed-length, we remove the frames corresponding to the padding before computing the probe features. In Table 4, we report the mean average precision (mAP) across all probes for each layer block, for both the frame-based and baseline (no ablation) models. We see that the mAP of the model trained with frame-based adversarial ablation achieves a consistently higher mAP than the baseline model, indicating its improved ability to detect content-carrying words in the speech.

Table 1. Validation R@10 for the various ablation strategies.

Model	R@1	R@5	R@10
Baseline	0.364	0.708	0.827
Random Ablation	0.363	0.715	0.839
Frame-based	0.385	0.741	0.859
Oracle-based	0.390	0.739	0.856

Table 4. mAP of probes by layers

Layer #	0	1	2	3
baseline	0.116	0.3419	0.529	0.475
frame-based	0.128	0.398	0.585	0.538

4. CONCLUSION

In this paper, we presented a method for constructing adversarial training examples for a model that learns speech representations by grounding spoken captions to visual images. We showed that our method automatically selects spectrogram regions likely to contain words that are descriptive of the image being captioned, and by masking them out during training we force the model to leverage information in the speech beyond these regions. We verified that our strategy results in improved retrieval accuracy, and an improved ability of the model to detect the presence of content-carrying words in an utterance. In our future work, we plan to extend our method to other tasks such as automatic speech recognition.

5. REFERENCES

[1] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *ArXiv*, 2018.

[2] Y.-A. Chung and J. Glass, "Generative pre-training for speech with autoregressive predictive coding," in *ICASSP*, 2020.

[3] W. Wang, Q. Tang, and K. Livescu, "Unsupervised pre-training of bidirectional speech encoders via masked reconstruction," in *ICASSP*, 2020.

[4] A. T. Liu, S. Yang, P.-H. Chi, P.-C. Hsu, and H. yi Lee, "Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders," *ICASSP*, 2020.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *NAACL*, 2019.

[6] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *NeurIPS*, 2020.

[7] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," 2021.

[8] D. F. Harwath, A. Torralba, and J. R. Glass, "Unsupervised learning of spoken language with visual context," in *NeurIPS*, 2016.

[9] T. A. Nguyen, M. de Seyssel, P. Rozé, M. Rivière, E. Kharitonov, A. Baevski, E. Dunbar, and E. Dupoux, "The zero resource speech benchmark 2021: Metrics and baselines for unsupervised spoken language modeling," *SAS @ NeurIPS*, 2020.

[10] S. Y. et al, "SUPERB: speech processing universal performance benchmark," *CoRR*, vol. abs/2105.01051, 2021.

[11] G. Chrupała, "Visually grounded models of spoken language: A survey of datasets, architectures and evaluation techniques," *arXiv*, 2021.

[12] D. F. Harwath, A. Recasens, D. Surís, G. Chuang, A. Torralba, and J. R. Glass, "Jointly discovering visual objects and spoken words from raw sensory input," in *ECCV*, 2018.

[13] D. Suris, A. Recasens, D. Bau, D. Harwath, J. Glass, and A. Torralba, "Learning words by drawing images," in *CVPR*, 2019.

[14] H. Kamper, S. Settle, G. Shakhnarovich, and K. Livescu, "Visually grounded learning of keyword prediction from untranscribed speech," in *INTERSPEECH*, 2017.

[15] W. Havard, J. Chevrot, and L. Besacier, "Models of visually grounded speech signal pay attention to nouns: a bilingual experiment on English and Japanese," in *ICASSP*, 2019.

[16] W.-N. Hsu, D. F. Harwath, and J. Glass, "Transfer learning from audio-visual grounding to speech recognition," in *INTERSPEECH*, 2019.

[17] T. Srinivasan, R. Sanabria, F. Metze, and D. Elliott, "Multimodal speech recognition with unstructured audio masking," in *ACL International Workshop on Natural Language Processing Beyond Text*, 2020.

[18] A. Alishahi et al., "Zr-2021vg: Zero-resource speech challenge, visually-grounded language modelling track, 2021 edition," 2021.

[19] R. Sanabria, A. Waters, and J. Baldridge, "Talk, don't write: A study of direct speech-based image retrieval," in *INTERSPEECH*, 2021.

[20] D. F. Harwath, W.-N. Hsu, and J. R. Glass, "Learning hierarchical discrete linguistic units from visually-grounded speech," in *ICLR*, 2020.

[21] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple augmentation method for automatic speech recognition," in *INTERSPEECH*, 2019.

[22] S. Sun, C. feng Yeh, M. Ostendorf, M.-Y. Hwang, and L. Xie, "Training augmentation with adversarial examples for robust speech recognition," in *INTERSPEECH*, 2018.

[23] D. F. Harwath, A. Recasens, D. Surís, G. Chuang, A. Torralba, and J. Glass, "Jointly discovering visual objects and spoken words from raw sensory input," *IJCV*, 2019.

[24] W.-N. Hsu, D. F. Harwath, T. Miller, C. Song, and J. R. Glass, "Text-free image-to-speech synthesis using learned segmental units," in *ACL*, 2021.

[25] G. Ilharco, Y. Zhang, and J. Baldridge, "Large-scale representation learning from visually grounded untranscribed speech," in *CoNLL*, 2019.

[26] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *ASRU*, 2011.

[27] T.-Y. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014.