Improving Multimodal Speech Recognition by Data Augmentation and Speech Representations

Anonymous ACL submission

Abstract

Multimodal speech recognition aims to improve the performance of automatic speech recognition (ASR) systems by leveraging additional visual information that is usually associated to the audio input. While previous approaches make crucial use of strong visual representations, *e.g.* by finetuning pretrained image recognition networks, significantly less attention has been paid to its counterpart: the speech component. In this work, we investigate ways of improving the base speech recognition system by following similar techniques to the ones used for the visual encoder, namely, transferring representations and data augmentation. First, we show that starting from a pretrained ASR significantly improves the state-of-the-art performance; interestingly, even when building upon a strong unimodal system, we still find gains by including the visual modality. Second, we employ speech data augmentation techniques to encourage the multimodal system to attend to the visual stimuli. This technique replaces previously used word masking and comes with the benefits of being conceptually simpler and yielding consistent improvements in the multimodal setting. We back up our conclusions by empirical results on three multimodal datasets, including the newly introduced Localized Narratives.

1 Introduction

With the advent of video sharing platforms (such as YouTube or Vimeo), multimodal data involving audio, visual and language are becoming ubiquitous. In many types of video, such as instructional videos, documentaries, movies, what is spoken is related (grounded) to the visual channel. In this paper we build upon this observation and address the task of automatic speech recognition in the context of visual information, also known as multimodal speech recognition. Concretely, we assume that we have two inputs (the acoustic signal and a related visual modality, such as a video or an image) and we want to output the transcription of the input utterance.

The recent work on multimodal speech recognition makes crucial use of deep end-to-end architectures (Palaskar et al., 2018; Caglayan et al., 2019; Srinivasan et al., 2020a,b,c; Paraskevopoulos et al., 2020; Ghorbani et al., 2021). We follow their suite and develop an end-to-end multimodal speech recognition system. Compared to previous work, we first experiment with two fusion mechanisms for combining the audio and visual modalities—either concatenation along the embedding dimension or concatenation along the temporal dimension. Second, and more importantly, we improve the pipeline by focusing on the speech component by (i) transferring pretrained speech representations and (ii) performing audio-level data augmentations.

Transferring representations. All recent papers on multimodal speech recognition transfer visual representations, obtained as activations or softmax predictions of a pretrained visual classification network. Depending on the training classes (objects, as in (Sun et al., 2016; Moriya and Jones, 2018; Palaskar et al., 2018; Srinivasan et al., 2020b); scenes, as in (Miao and Metze, 2016; Gupta et al., 2017; Srinivasan et al., 2020a); actions, as in (Miao and Metze, 2016; Caglayan et al., 2019; Paraskevopoulos et al., 2020); faces, as in (Miao and Metze, 2016; Moriya and Jones, 2019)), the visual encoder is more sensible to pick up certain types of visual information. However, none of these prior works make use of pretrained speech representations. In this paper we not only show the importance of starting from a good representation for both the audio and visual channels, but crucially we provide an answer to question of whether the visual information is helpful for a stronger baseline system.

Data augmentation. Augmenting the training set with perturbed samples is a common technique to enforce invariants for high-capacity deep learn-

ing models. For image classification, images are altered by horizontal flips and small affine transformations, while for speech recognition the speed of an utterance is modified by time warping. We use these ideas, in particular those related to speech augmentation, to improve the multimodal models. Our intuition is that perturbing the audio signal will make the model more reliant on the visual channel. The inspiration stems from the work of Srinivasan *et al.* which have shown that multimodal models improve over the baseline ASR even when audio-image pairs are mismatched (incongruent) (Srinivasan et al., 2019), but if the multimodal models were trained on masked audio signals, this behaviour is alleviated (Srinivasan et al., 2020a). Compared to the previous approaches (Srinivasan et al., 2019, 2020a,b), we do not limit ourselves to temporal masking of words, but randomly mask temporal and frequency segments, as in (Park et al., 2019); as a consequence our approach is more general and convenient to use. Another key distinction is that we do not carry the evaluation on the masked data, but consider the more realistic scenario of assuming clean speech at test time and performing alterations only at train time.

To summarize our contributions are: (i) improve the speech component of multimodal models by transfer learning and data augmentation; (ii) explore fusion techniques for the audio and visual modalities; (iii) report state-of-the-art performance on multiple multimodal speech recognition datasets such as Flickr8K and How2, and new results on the recently introduced Localized Narratives dataset.

2 Related work

In this section we discuss the main categories of multimodal models and present our task in the context of related problems.

A taxonomy of multimodal models. Perhaps unsurprisingly, the techniques for multimodal speech recognition have been following the trends in speech processing and computer vision. Based on the choices of the two main components (namely, the audio and visual pipelines), we distinguish three types of systems.

The first approaches (Mukherjee and Roy, 2003; Fleischman and Roy, 2008) date back to the 2000s and rely on the Hidden Markov Models and Gaussian Mixture Models (the HMM-GMM paradigm) for speech recognition and hand-crafted features for the visual channel. These methods also assumed more constrained and simplified settings to account for the lack of data.

The second category of multimodal systems (Miao and Metze, 2016; Gupta et al., 2017; Sun et al., 2016; Moriya and Jones, 2018; Oneață and Cucu, 2021) uses Hidden Markov Models and Deep Neural Networks (the HMM-DNN paradigm) for speech recognition, while the visual component relies on pretrained networks. While many of these approaches fuse the two components at the last stage (language modeling) (Gupta et al., 2017; Sun et al., 2016; Moriya and Jones, 2018; Oneață and Cucu, 2021), a notable exception is the work of Miao and Metze (2016), which advocates for early fusion, at the audio level, based on the observation that the acoustic conditions can correlate with the visual context.

Finally, the latest type of models leverage recent developments in end-to-end architectures and training (Palaskar et al., 2018; Caglayan et al., 2019; Srinivasan et al., 2020a,b,c; Paraskevopoulos et al., 2020; Ghorbani et al., 2021). For the audio part, the most common model involves recurrent networks for encoder and decoders, coupled through an attention mechanism, but other variants include using a connectionist temporal classification (CTC) model (as done in (Palaskar et al., 2018)) or the Transformer architecture, which involves attention-only layers (as done in (Paraskevopoulos et al., 2020)). Various fusion levels have been explored: encoder, decoder, and also at acoustic level. Of course these can be combined as done by Caglayan et al. (2019).

Our approach falls into the latter category, of end-to-end architectures. We share similarities to the work of Paraskevopoulos et al. (2020), in that we employ Transformer architecture and sub-word modeling, however our base speech recognition system is much stronger and we focus our empirical evaluation on the importance of transfered representations.

Related tasks. We distinguish our work from two closely related tasks, which also make use of audio and visual input modalities. A first task is audio-visual speech recognition (Mroueh et al., 2015; Petridis et al., 2018; Afouras et al., 2018), which also attempts to improve speech recognition, but it uses lip movement information. A key difference to our methodology is that for the lip-based recognition there is a much tighter (although arguably more difficult to model) relation between the video and the transcriptions, while for multi-

modal speech recognition, the relationship is at a semantic level and might affect only a few words, which have visual grounding. A second task is learning audio-visual correspondences, but without depending on the textual annotations. This formulation has the advantage of relying on less supervision and finds many uses, such as representation learning (Harwath et al., 2016; Harwath and Glass, 2015), learning linguistic units (Harwath et al., 2019), semantic keyword spotting (Kamper et al., 2019), speech-based image retrieval (Synnaeve et al., 2014; Harwath et al., 2016, 2018) and speech-based object localization (Harwath et al., 2018).

3 Methodology

In speech recognition, an audio input $\mathbf{a}$ is mapped to a transcription $\mathbf{t}$, usually represented as a sequence of tokens. In the usual encoder-decoder instantiation the output is obtained by composing the two components: $\mathbf{t} = \mathrm{Dec}(\mathrm{Enc}(\mathbf{a}))$. In the case of multimodal speech recognition, we assume that we have access to an additional input—the visual channel $\mathbf{v}$. The visual information is processed by a separate encoder, $\mathrm{Enc_v}$, and integrated into the network by a fusion function, which we denote by "$\bowtie$":

$$\mathbf{t} = \mathrm{Dec}(\mathrm{Enc}(\mathbf{a}) \bowtie \mathrm{Enc_v}(\mathbf{v})).$$

Next, we discuss each of the components: speech encoder and decoder in §3.1, visual encoder in §3.2, fusion in §3.3.

3.1 Speech recognition system

The backbone of the multimodal system the is an end-to-end automatic speech recognition system. We use a Transformer network, which is based on self-attention modules for the encoder and attention modules in the decoder to pool information from the audio stream. The network predicts tokens in an autoregressive fashion, by modelling the probability of the next token given the audio and previously predicted tokens, $p(t_k|\mathbf{t}_{<k}, \mathbf{a})$.

Transfer learning. Instead of starting the training of the multimodal speech recognition system from scratch, we explore initializing the speech components (encoder and decoder) from a pretrained speech recognition model. The base system is a pretrained ASR system on a large unimodal dataset, the LibriSpeech corpus (Panayotov et al., 2015), whose weigths we transfer and then adapt on the target multimodal dataset via finetuning.

Data augmentations. We extend the set of speech samples with perturbed versions of the signal in order to make the system more robust and to encourage the decoder to attend to the visual component. The augmentations are based on SpecAugment (Park et al., 2019) and include time warping, frequency masking and time masking. The same transformations were used for training the base unimodal system on LibriSpeech and we also apply them when training the multimodal system. Previous approaches (Srinivasan et al., 2019, 2020a,c) used temporal masking, but in their case the removed segments corresponded to words (such as nouns and places). As such, these methods rely on additional components such as audio-text alignment and part-of-speech tagging, while our approach is unstructured and, consequently, free of these dependencies. Moreover, previous methods investigated masking with a different goal in mind (not as a data augmentation technique): to quantify how well the visual component is able to retrieve the masked words at test time.

3.2 Visual encoder

The visual encoder summarizes the information present at the input of the visual channel. We assume an image at input and build upon the popular ResNet architecture (He et al., 2016), which was also used in previous works on multimodal speech recognition, *e.g.*, (Caglayan et al., 2019; Srinivasan et al., 2019, 2020a). The visual encoder is initialized with the weights of a pretrained model on the ImageNet dataset (Russakovsky et al., 2015) and uses intermediate network activations as its encoding. Depending at which layer we take the activations, we obtain either (i) a single feature vector or (ii) a sequence of feature vectors. Concretely, the activations before the softmax layer (and after the global average pooling layer) yield a single fixed-sized vector, which encodes global information from the entire image. If we take the activations from one layer before (that is, before the global average pooling layer), we obtain a 7 $\times$ 7 grid of embeddings, which we sequence as a list of $K = 49$ embeddings. This second approach encodes more local information, which we hope will allow the model to use more fine-grained characteristics of the image. On top of the sequence of embeddings we optionally learn layers of gated

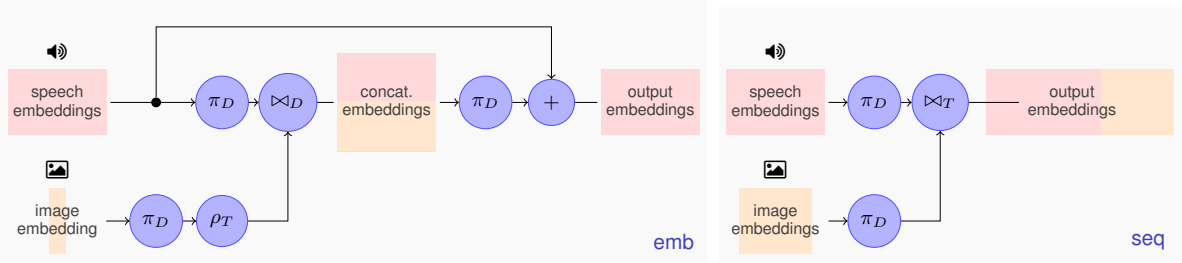

Figure 1: The two proposed fusion mechanisms of the audio and visual modalities: emb, fuses along the embedding dimension (left); seq, fuses along the sequence dimension (right). Additional operations (dense projection, denoted by π; repeat operation, denoted by ρ) ensure matching dimensions and better adapted embeddings. The subscript (D or T) indicates the axis along which each transformation is applied (embedding dimension or sequence dimension). The symbol "$\bowtie$" denotes concatenation.

multilayer perceptrons (gMLP) (Liu et al., 2021), a recently introduced substitute for self-attention layers, which alternates channel-wise with sequence-wise dense layers. Compared to the attention layer, the gMLP architecture requires less computation and memory, while still maintaining the performance.

Relation to prior work. Most of prior work uses a single global feature vector to encode the visual information, some notable exceptions being (Paraskevopoulos et al., 2020) and (Srinivasan et al., 2020c). The method in (Paraskevopoulos et al., 2020) works on video sequences and extracts a feature vector for each video frame, while in (Srinivasan et al., 2020b) the authors extract ResNet features for $K = 16$ object proposals obtained from a detection network. Compared to the latter approach, our approach does not require a pretrained detection module and hence is simpler and can be trained with less supervision.

3.3 Fusion mechanisms

Our fusion techniques combine the speech and visual embeddings (as produced by each of the two encoders) before feeding them into the decoder. In the following we assume that speech embeddings have dimension $T \times D_{\mathbf{a}}$, while visual embeddings have dimension $K \times D_{\mathbf{v}}$ (the first axis, of length K, can correspond to a list of boxes in an image or a list of frames in a video). We experiment with two fusion approaches, either along the embedding dimension (emb) or along the sequence dimensions (seq); these two variants are illustrated in Figure 1. The choice of fusion is also influenced by the visual encoder: if we represent the visual input with a single feature vector ($K = 1$) it is possible to concatenate along the embedding dimension, while if we use a list of visual features ($K > 1$) then the concatenation along the sequence dimension is more suitable.

Fusing along the embedding dimension (emb). In this case we fuse the speech and visual features along the dimension of the embeddings. More precisely, we first project the two inputs into a common subspace and replicate the visual embedding T times, then we concatenate the two representations and, finally, project the output to have dimension $D_{\mathbf{a}}$. In this case, the fusion procedure outputs a matrix of the same size as the input speech matrix, $T \times D_{\mathbf{a}}$. Retaining the original dimension has a number of advantages: it allows us to maintain the same decoder size as in the unimodal case (enabling transfer learning and fairer comparisons) and to use residual connections (from speech to the fused features), which are known to help learning.

Fusing along the sequence dimension (seq). When the embeddings of the two input modalities are both sequences, it makes sense to concatenate the visual and speech features along the sequence dimension (temporal for speech and patch-wise for the image). As the decoder attends along sequential dimension of the input, this operation will become more expensive after the fusion. However, the seq fusion has the advantage of being more flexible than the emb variant, since the decoder has the option of pooling separately the audio and visual features, without mixing the two.

Relation to previous work. Many of the previous approaches were based on recurrent networks and the common ways of incorporating the visual context were (i) to set the first decoded "word" as the visual embedding (Sun et al., 2016; Moriya and Jones, 2018), or (ii) to initialize the hidden state of the recurrence with the visual embedding

(Caglayan et al., 2019). Another method, encountered especially for adapting acoustic features, was visual adaptive training (Miao and Metze, 2016; Palaskar et al., 2018), which amounts to applying a linear transformation parameterized by the visual encoding. While concatenation of features was previously employed (Miao and Metze, 2016; Palaskar et al., 2018) it was not used in the context of Transformer architectures. When the visual information is a sequence, attention-based methods are a popular choice (Srinivasan et al., 2020b; Paraskevopoulos et al., 2020; Ghorbani et al., 2021). All these methods pool independently across the audio and visual streams, whereas in our case the seq method pools over both of them simultaneously. The methods in (Srinivasan et al., 2020b; Ghorbani et al., 2021) attend to the visual sequence based on the previously decoded word (as we do), while (Paraskevopoulos et al., 2020) pools based on the audio. The main distinction between (Ghorbani et al., 2021) and (Srinivasan et al., 2020b) is that the former simply concatenates the two pooled representations, while the latter predicts which of the two modalities (visual or audio) should be preferred through a second, hierarchical attention layer.

4 Experimental setup

In this section we present the experimental setup, including the multimodal datasets (§4.1) and additional implementation details (§4.2).

4.1 Datasets

We carry out the evaluation on three datasets that contain the three desired modalities (audio, visual, language).

Flickr8K (Hodosh et al., 2013; Harwath and Glass, 2015) consists of 8K images, each described by five captions. The original dataset (Hodosh et al., 2013) contained only the visual and language modalities, and it was later extended with audio recordings of the read captions by (Harwath and Glass, 2015).

How2 (Sanabria et al., 2018) contains instructional videos downloaded from YouTube and comes with additional shot information and transcriptions. We use the 300h variant, which totals around 13.5K videos (190K shots). The dataset consists of pre-extracted audio and visual features, but, in order to use pretrained models, we had to use the original videos; the raw data was kindly provided by the authors, upon request.

Localized Narratives (Pont-Tuset et al., 2020) is a recently introduced dataset, that extends four popular image datasets (Flickr30K (Young et al., 2014), COCO (Lin et al., 2014), ADE20K (Zhou et al., 2019), Open Images (Kuznetsova et al., 2020)) with new captions, audio recordings and mouse traces (which locate the spoken words in the image). Compared to the original datasets, the captions are richer and the audio component is challenging due to noisy recording conditions and accented speech. We use this dataset to carry out an ablation study. In order to be able to perform such extensive studies we (i) use only the Flickr30K part, (ii) segment the audio into sentences (based on the provided transcripts), (iii) remove utterances longer than 15 seconds, (iv) subsample half of the utterances. This procedure is applied on all three splits (training, validation, testing) and yields around 32K, 1K, 1K samples, respectively.

4.2 Implementation details

The implementation is based on the ESPnet framework (Watanabe et al., 2018) and our code is available online.[1]

The speech recognition component is a Transformer architecture, which is pretrained on the LibriSpeech dataset (Panayotov et al., 2015). The system outputs tokens from a vocabulary with 5000 elements, which was obtained by subword segmentation using an unigram language model (Kudo, 2018); we don't use an external language model. For finetuning, we train for 50 epochs for the smaller datasets (Flickr8K and Localized Narratives) and 30 epochs for the larger dataset (How2). For optimization, the learning rate is warmed up linearly from 3.2×10^{-8} to 8×10^{-4} over 25K batches, after which it is decreased as a function of $1/s^2$ in the step number s. At test time, we ensemble the ten best models by averaging their weights; this technique gives small, but consistent improvements over predicting with only the best model.

The visual encoder is a ResNet architecture with either 18 or 50 layers, pretrained on the ImageNet dataset (Russakovsky et al., 2015), yielding 512 or 2048-dimensional embeddings, respectively. The input image is rescaled to 224×224 pixels and standardized using the ImageNet statistics. We perform image data augmentation by random horizon-

[1] We will provide the link after the anonymization period has passed.

tal flipping. The How2 dataset contains video, but since our visual embedding works on images, we use only the middle frame. As the videos are shots and hence stable in terms of viewpoint change, we expect a single frame to encode enough information,

For the fuse seq variant we use two gMLP layers, as this choice gave slightly better results than the two other variants that we have experimented with (zero and one layer).

5 Experiments

This section presents the empirical evaluation of the proposed methodology. First, in §5.1 we compare our best unimodal and multimodal systems to baseline and state-of-the-art approaches. Second, in §5.2 we present an ablation study over the main individual contributions: transfer learning and data augmentations.

5.1 Main results

Table 1 presents speech recognition performance for four of our systems: two unimodal variants (a pretrained ASR, used as a baseline, and its finetuned counterpart based on adapting the pretrained method on each dataset) and two multimodal variants (both trained by finetuning all components, but differing in the fusion techniques, emb or seq, as described in §3.3). Both multimodal methods use the SpecAugment data augmentation and a ResNet with 50 layers as the visual encoder. We contrast our approaches to state-of-the-art methods. Previous studies evaluate usually on a single dataset, for example, Flickr8K (Sun et al., 2016; Srinivasan et al., 2020b) or How2 (Paraskevopoulos et al., 2020; Ghorbani et al., 2021), while we report performance on three datasets: the two aforementioned ones and Localized Narratives (on which we are the first to report multimodal speech recognition performance).

We observe that the pretrained method already improves over previous work on Flickr8K, although its results are poorer on How2 and Localized Narratives due to data mismatch. However, by finetuning, the speech-only unimodal system significantly outperforms the current state-of-the-art, yielding relative improvements of 72% and 31% on Flickr8K and How2, respectively. The results for the multimodal systems, which include the visual information, are better than the unimodal results in the case of How2 and Localized Narratives dataset; for Flickr8K it is difficult to improve presumably because it is a clean dataset for which the ASR already works well and many of its errors are not visually grounded. Among the two fusing methods the results are mixed, the fusion along the embedding dimension, emb, being the better method on two of the three datasets.

5.2 Ablation

In this subsection we carry two main ablation studies to better understand the impact of data augmentation and the importance of transferring representations.

Data augmentations. We evaluate the impact of the SpecAugment data augmentation technique in three scenarios: for the unimodal system and for the two multimodal variants using the two feature fusion techniques (emb and seq). For all cases, we perform finetuning of all components and for the visual-based systems we use the 50-layer ResNet.

Table 2 reveals that speech data augmentation is important for the multimodal systems, yielding improvements in five out of the six cases. These results suggest that perturbing the audio signal might help the multimodal models rely more on the visual encoder and eventually produce better results. Surprisingly, the unimodal variants have seen little benefit from data augmentation, with the exception of the results on the How2 dataset.

Transferring representations. We conduct an extensive ablation study over the impact of initialization (random or pretrained) and training procedure (fixed or finetuned weights) for each of the components of the model (audio encoder, visual encoder, decoder). These experiments are carried only on the Localized Narratives dataset and for the multimodal setting we use only the emb fusing method. Additionally, we investigate the impact of visual encoder's capacity by varying the number of layers in the ResNet architecture: 18 or 50. The results are presented in Table 3.

Rows 1–5 show the results for the unimodal system, corresponding to a standard, speech-only ASR. We observe that the pretrained variant, without any finetuning (row 1) is underperforming, most likely, due to the large mismatch in both terms of audio (speech is noisy and accented) and language data. On the other hand, ignoring the availability of pretrained representations (row 2) is also not ideal: training the network from scratch, as is customary done in previous works, produces better, but still

method	visual	fuse	Flickr8K	How2	Loc. Nar.
(Sun et al., 2016)			14.8	—	—
	✔		13.8	—	—
(Srinivasan et al., 2020b)			13.6	—	—
	✔		14.1	—	—
(Paraskevopoulos et al., 2020)			—	19.2	—
	✔		—	18.4	—
(Ghorbani et al., 2021)			—	17.7	—
	✔		—	17.2	—
pretrain			11.1	26.9	49.3
finetune			**3.8**	11.8	4.3
finetune	✔	emb	4.3	11.1	**3.9**
finetune	✔	seq	4.7	**10.8**	4.0

Table 1: Comparison to state-of-the-art approaches on the test sets of three multimodal datasets (Flickr8K, How2 and Localized Narratives) in terms of word error rate (lower values are better). Visual indicates those variants that use the visual channel as input in addition to the speech.

visual	fuse	aug.	Flickr8K	How2	Loc. Nar.
	—		**3.8**	11.8	**4.3**
	—	✔	4.2	**11.2**	4.5
✔	emb		4.8	11.8	4.1
✔	emb	✔	**4.3**	**11.1**	**3.9**
✔	seq		**4.0**	11.8	4.2
✔	seq	✔	4.7	**10.8**	**4.0**

Table 2: Evaluation of the impact of audio augmentations (aug.) on the test sets of the three multimodal datasets in terms of word error rate. All models are finetuned and the multimodal variants use the ResNet50 as visual encoder.

unsatisfactory transcriptions. Rows 3 and 4 show the results for the case when the encoder is fixed and the decoder is trained: either from scratch (row 3) or by finetuning the pretrained weights (row 4). Since the decoder of an end-to-end ASR model plays also the role of a language model, this procedure is akin to language adaptation and results in significant boosts in performance for both variants. Finally, finetuning both components (row 5) yields the best results, with a relative improvement of around 30%.

Rows 6–13 present the results for the multimodal systems, using a visual encoder with 18 (rows 6–9) or 50 layers (rows 10–13). For this set of experiments, we use only with the finetuning approach, as the results of the unimodal system have showed that this technique is superior. We also always adapt the decoder because the speech-vision fusion affects the distribution of features. Note that the fused features are projected to the same embedding dimension as the speech features, which enables sharing the pretrained decoder weights. The projection layers in the fusion layer are always trainable.

We first note that including visual information improves over the single-stream system in all scenarios: either if we keep the encoders fixed (rows 6 and 10 vs row 4) or if we finetune the encoders (rows 9 and 13 vs row 5). Second, we observe that we obtain better results as we allow for more components to be finetuned, with the last column indicating a correlation between the number of trainable parameters and the performance. The best results are achieved when finetuning all components (rows 9 and 13). Finally, increasing the capacity of the visual encoder yields similar results. We do see slight improvements for the cases when we finetune the speech encoder (row 12 vs row 8; row 13 vs row 9), potentially suggesting the coupling between the two modalities needs to be accounted also by the encoders and not only the decoders.

6 Conclusions

In this paper, we extend and build upon state-of-the-art approaches for multimodal speech recognition. We employ a Transformer ASR architecture as a baseline system in which we inject visual information through a ResNet image encoder. In contrast to the previous methods, we use pretrained representations for both the speech and visual channels. This

	audio encoder		visual encoder			decoder		WER (%)	num. trainable params ($\times 10^6$)
	init	train	init	train	network	init	train		
1	↱	🔒	—	—	—	↱	🔒	49.3	0
2	⤮	🔓	—	—	—	⤮	🔓	22.5	99.4
3	↱	🔒	—	—	—	⤮	🔓	9.1	32.9
4	↱	🔒	—	—	—	↱	🔓	6.3	32.9
5	↱	🔓	—	—	—	↱	🔓	4.3	99.4
6	↱	🔒	↱	🔒	ResNet18	↱	🔓	5.8	33.2
7	↱	🔒	↱	🔓	ResNet18	↱	🔓	5.5	44.4
8	↱	🔓	↱	🔒	ResNet18	↱	🔓	4.3	99.6
9	↱	🔓	↱	🔓	ResNet18	↱	🔓	4.2	110.8
10	↱	🔒	↱	🔒	ResNet50	↱	🔓	5.9	33.4
11	↱	🔒	↱	🔓	ResNet50	↱	🔓	5.6	56.9
12	↱	🔓	↱	🔒	ResNet50	↱	🔓	4.1	99.8
13	↱	🔓	↱	🔓	ResNet50	↱	🔓	4.1	123.3

Table 3: Transferring representations—evaluation on the test set of the Localized Narratives dataset. For each of the three components of the model (audio encoder, visual encoder, decoder), we indicate how the model's weights are initialized (either random ⤮ or shared from a pretrained model ↱) and trained (either fixed 🔒 or finetuned 🔓). For each setting we report the word error rate (WER) and the number of trainable parameters. The visual information is fused with the emb method. For these experiments, we did not use audio augmentations.

leads to substantial improvements over the state of the art on two standard multimodal datasets: Flickr8K and How2. In addition, we explore a series of fusion techniques for the speech and visual embeddings. Finally, our work also reports two ablation studies which provide important insights on the role of using speech augmentation before training the multimodal network and the individual contribution of finetuning the various components of the system.

For future work we plan to perform a detailed qualitative analysis revealing the exact benefits of introducing the visual modality as input to the ASR system. A similar avenue for further research is investigatating which parts of the inputs (audio, image, previously predicted tokens) contribute more to the output. We believe that modern tools for explainable machine learning (Kokhlikyan et al., 2020; Samek et al., 2021; Joshi et al., 2021) can help us better understand the complicated interactions that arise in the multimodal network.

References

Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. 2018. Deep audio-visual speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Ozan Caglayan, Ramon Sanabria, Shruti Palaskar, Loic Barraul, and Florian Metze. 2019. Multimodal grounding for sequence-to-sequence speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8648–8652.

Michael Fleischman and Deb Roy. 2008. Grounded language modeling for automatic speech recognition of sports video. In *Association for Computational Linguistics*, pages 121–129.

Shahram Ghorbani, Yashesh Gaur, Yu Shi, and Jinyu Li. 2021. Listen, look and deliberate: Visual context-aware speech recognition using pre-trained text-video representations. In *IEEE Spoken Language Technology Workshop*, pages 621–628.

Abhinav Gupta, Yajie Miao, Leonardo Neves, and Florian Metze. 2017. Visual features for context-aware speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5020–5024.

David Harwath and James Glass. 2015. Deep multimodal semantic embeddings for speech and images. In *Workshop on Automatic Speech Recognition and Understanding*, pages 237–244.

David Harwath, Wei-Ning Hsu, and James Glass. 2019. Learning hierarchical discrete linguistic units from visually-grounded speech. In *International Conference on Learning Representations*.

David Harwath, Adrià Recasens, Dídac Surís, Galen Chuang, Antonio Torralba, and James Glass. 2018.

Jointly discovering visual objects and spoken words from raw sensory input. In *European Conference on Computer Vision*, pages 649–665.

David Harwath, Antonio Torralba, and James Glass. 2016. Unsupervised learning of spoken language with visual context. In *Advances in Neural Information Processing Systems*, pages 1858–1866.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899.

Gargi Joshi, Rahee Walambe, and Ketan Kotecha. 2021. A review on explainability in multimodal deep neural nets. *IEEE Access*.

Herman Kamper, Gregory Shakhnarovich, and Karen Livescu. 2019. Semantic speech retrieval with a visually grounded model of untranscribed speech. *Transactions on Audio, Speech and Language Processing*, 27(1):89–98.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, et al. 2020. Captum: A unified and generic model interpretability library for PyTorch. *arXiv preprint arXiv:2009.07896*.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Association for Computational Linguistics*, pages 66–75.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. 2020. The Open Images dataset v4. *International Journal of Computer Vision*, pages 1–26.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Hanxiao Liu, Zihang Dai, David R So, and Quoc V Le. 2021. Pay attention to MLPs. *arXiv preprint arXiv:2105.08050*.

Yajie Miao and Florian Metze. 2016. Open-domain audio-visual speech recognition: A deep learning approach. In *Interspeech*, pages 3414–3418.

Yasufumi Moriya and Gareth JF Jones. 2018. LSTM language model adaptation with images and titles for multimedia automatic speech recognition. In *IEEE Spoken Language Technology Workshop*, pages 219–226.

Yasufumi Moriya and Gareth JF Jones. 2019. Multimodal speaker adaptation of acoustic model and language model for ASR using speaker face embedding. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8643–8647.

Youssef Mroueh, Etienne Marcheret, and Vaibhava Goel. 2015. Deep multimodal learning for audio-visual speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2130–2134.

Niloy Mukherjee and Deb Roy. 2003. A visual context-aware multimodal system for spoken language processing. In *European Conference on Speech Communication and Technology*.

Dan Oneață and Horia Cucu. 2021. Multimodal speech recognition for unmanned aerial vehicles. *Computers & Electrical Engineering*, 90:106943.

Shruti Palaskar, Ramon Sanabria, and Florian Metze. 2018. End-to-end multimodal speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5774–5778.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5206–5210.

Georgios Paraskevopoulos, Srinivas Parthasarathy, Aparna Khare, and Shiva Sundaram. 2020. Multiresolution and multimodal speech recognition with transformers. In *Association for Computational Linguistics*.

Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. 2019. SpecAugment: A simple data augmentation method for automatic speech recognition. In *Interspeech*, pages 2613–2617.

Stavros Petridis, Themos Stafylakis, Pingchuan Ma, Feipeng Cai, Georgios Tzimiropoulos, and Maja Pantic. 2018. End-to-end audiovisual speech recognition. *CoRR*, abs/1802.06424.

Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. 2020. Connecting vision and language with localized narratives. In *European Conference on Computer Vision*, pages 647–664.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252.

Wojciech Samek, Grégoire Montavon, Sebastian Lapuschkin, Christopher J Anders, and Klaus-Robert Müller. 2021. Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, 109(3):247–278.

Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: A large-scale dataset for multimodal language understanding. In *Advances in Neural Information Processing Systems*.

Tejas Srinivasan, Ramon Sanabria, and Florian Metze. 2019. Analyzing utility of visual context in multimodal speech recognition under noisy conditions. In *The How2 Challenge: New Tasks for Vision & Language, ICML*.

Tejas Srinivasan, Ramon Sanabria, and Florian Metze. 2020a. Looking enhances listening: Recovering missing speech using images. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6304–6308.

Tejas Srinivasan, Ramon Sanabria, Florian Metze, and Desmond Elliott. 2020b. Fine-grained grounding for multimodal speech recognition. In *Empirical Methods in Natural Language Processing*.

Tejas Srinivasan, Ramon Sanabria, Florian Metze, and Desmond Elliott. 2020c. Multimodal speech recognition with unstructured audio masking. In *Workshop on Natural Language Processing Beyond Text, EMNLP*.

Felix Sun, David Harwath, and James Glass. 2016. Look, listen, and decode: Multimodal speech recognition with images. In *IEEE Spoken Language Technology Workshop*, pages 573–578.

Gabriel Synnaeve, Maarten Versteegh, and Emmanuel Dupoux. 2014. Learning words from images and speech. In *NIPS Workshop on Learning Semantics*.

Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai. 2018. ESPnet: End-to-end speech processing toolkit. In *Interspeech*, pages 2207–2211.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2019. Semantic understanding of scenes through the ADE20K dataset. *International Journal of Computer Vision*, 127(3):302–321.