Fine-Grained Grounding for Multimodal Speech Recognition

Tejas Srinivasan
Language Technologies Institute
Carnegie Mellon University
tsriniva@andrew.cmu.edu

Ramon Sanabria
CSTR, ILCC
University of Edinburgh
r.sanabria@ed.ac.uk

Florian Metze
Language Technologies Institute
Carnegie Mellon University
fmetze@andrew.cmu.edu

Desmond Elliott
Department of Computer Science
University of Copenhagen
de@di.ku.dk

Abstract

Multimodal automatic speech recognition systems integrate information from images to improve speech recognition quality, by grounding the speech in the visual context. While visual signals have been shown to be useful for recovering entities that have been masked in the audio, these models should be capable of recovering a broader range of word types. Existing systems rely on global visual features that represent the entire image, but localizing the relevant regions of the image will make it possible to recover a larger set of words, such as adjectives and verbs. In this paper, we propose a model that uses finer-grained visual information from different parts of the image, using automatic object proposals. In experiments on the Flickr8K Audio Captions Corpus, we find that our model improves over approaches that use global visual features, that the proposals enable the model to recover entities and other related words, such as adjectives, and that improvements are due to the model's ability to localize the correct proposals.[1]

1 Introduction

Multimodal language processing is inspired by evidence that conceptual representations in humans are distributed across modality-specific systems (Barsalou, 2003). In recent years, researchers have developed deep learning models that combine visual, linguistic, and auditory modalities for a variety of multimodal tasks, such as automatic image captioning (Vinyals et al., 2015), visual question-answering (Antol et al., 2015), and image–speech retrieval (Harwath and Glass, 2015), *inter-alia*.

In multimodal automatic speech recognition (ASR), there have been efforts to integrate visual context into acoustic models (Miao and Metze, 2016) and sequence-to-sequence models (Palaskar et al., 2018; Sanabria et al., 2018; Caglayan et al., 2019). However, it is not clear if the visual context actually improves ASR or if it helps to regularize the model (Caglayan et al., 2019). Srinivasan et al. (2020) recently showed that *global* visual context (a single feature vector representing the entire image) is useful when the visually depictable linguistic inputs are masked, *i.e.*, masking the speech that refer to entities. This experimental methodology, inspired by Caglayan et al. (2019), creates a *systematic gap* in the speech signal that can be resolved by leveraging the visual context; for example, when the audio drops during online distance-based learning or video calls with family and friends.

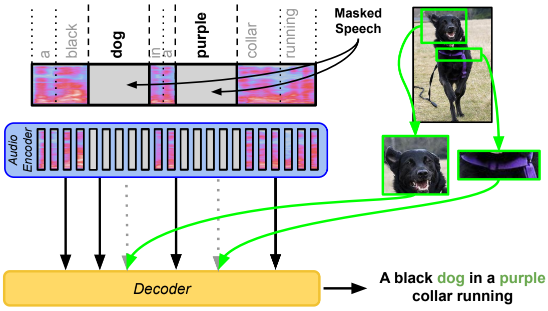

Figure 1: Our multimodal speech recognition model transcribes masked speech using visual features extracted from object proposals.

We present a model for multimodal ASR that learns to integrate visual features from object proposals (Ren et al., 2015), rather than image-level features, which has previously proven to be useful for image captioning and VQA (Anderson et al., 2018). Object proposals are rectangular image regions that are expected to contain objects. The novelty of our model is that when it encounters masked audio, it grounds (Harnad, 1990) the missing speech to different regions of the image. Our model learns separate attention distributions (Bahdanau et al., 2016) for each modality and combines

[1] The code is available at https://github.com/tejas1995/MultimodalASR

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 2667–2677
November 16 - 20, 2020.

them using a hierarchical attention mechanism in the decoder (Libovický and Helcl, 2017). This approach to integrating visual context from object proposals allows the model to better learn the relationship between speech and depicted colours, entities, and (to some extent) cardinals.

In experiments on the Flickr8K Audio Captions corpus (Harwath and Glass, 2015), we find that our model is much better at recovering masked speech than previous work. We also find that our model is right for the right reasons. In Section 4.1, we perform an object localization analysis, finding that 44% – 49% of the maximally attended object proposals, and 80% – 83% of the top-5 attended proposals, overlap with the ground-truth bounding box annotations. This shows that our model is verifiably leveraging the visual context.

The main contributions of this paper are:

- A new model for multimodal ASR that integrates visual features from automatically detected object proposals (Ren et al., 2015), which make it possible for the speech to be directly grounded into regions of the image.
- We propose a method for forcing the model to leverage the visual context by masking a broad range of words in the speech input during training, as opposed to only masking entities (Srinivasan et al., 2020).
- We define an object localization evaluation for multimodal ASR to show when models attend to the expected regions of the image when integrating visual context.

2 Methodology

2.1 Problem Formulation

ASR is the task of transcribing a speech sequence $\mathbf{x_{1 \ldots S}}$ into a sequence of words $\mathbf{y_{1 \ldots T}}$, where $\mathbf{S}$ and $\mathbf{T}$ are the lengths of the speech and word sequence, respectively. In multimodal ASR, there is an additional visual context $\mathbf{v}$, which can be used to improve the speech transcription. In this paper, the visual context is given by a static natural image and is literally described by the speech sequence.

We investigate the utility of the additional visual context in *noisy scenarios*, where words are randomly masked in the speech sequence. We expect that when the audio is clean, the audio context should be sufficient for transcription. However, when segments of the audio signal are masked, a multimodal ASR model will use the visual context to recover the missing word(s) in the speech.

2.2 ASR Models

Unimodal ASR Our UNIMODAL model is a word-level (Palaskar and Metze, 2018) sequence-to-sequence model with attention (Bahdanau et al., 2016; Chan et al., 2016). The model takes as input a sequence $\mathbf{x_{1 \ldots S}}$ (as described in Section 3.2) which is passed through the encoder. The encoder consists of 6 bidirectional LSTM layers (Schuster and Paliwal, 1997; Hochreiter and Schmidhuber, 1997) with temporal sub-sampling (Chan et al., 2016) in the middle two layers. The decoder is a two-layer conditional GRU (Cho et al., 2014) which computes attention over the encoder states $\mathbf{E}$.

$$\mathbf{E} = \text{Encoder}(\mathbf{x_{1 \ldots S}}) \tag{1}$$

$$\mathbf{h_t^{dec1}} = \text{GRU}_1(\mathbf{y_{t-1}}, \mathbf{h_{t-1}^{dec1}}) \tag{2}$$

$$\mathbf{z_t} = \text{Attention}(\mathbf{E}, \mathbf{h_t^{dec1}}) \tag{3}$$

$$\mathbf{h_t^{dec2}} = \text{GRU}_2(\mathbf{z_t}, \mathbf{h_{t-1}^{dec2}}) \tag{4}$$

Multimodal ASR with Global Visual Features The baseline multimodal ASR model uses global visual features $\mathbf{v}$ extracted from the entire image, which are incorporated into the ASR decoder. We add a hierarchical attention layer (Libovický and Helcl, 2017) that adaptively weights the features from the speech encoder context vector $\mathbf{z_t}$ (Eqn. 3) and the visual feature vector $\mathbf{v}$. The hierarchical context vector $\mathbf{z_t^{hier}}$ is the input to the second layer of the ASR decoder (Eqn. 4):

$$\mathbf{z_t^{hier}} = \text{Attention}(\{\mathbf{z_t}, \mathbf{v}\}, \mathbf{h_t^{dec1}}) \tag{5}$$

$$\mathbf{h_t^{dec2}} = \text{GRU}_2(\mathbf{z_t^{hier}}, \mathbf{h_{t-1}^{dec2}}) \tag{6}$$

By conditioning the hierarchical attention on the output of the first decoder layer, it learns modality-specific attention weights α_a and α_v that form a probability distribution. α_a and α_v effectively control the importance of the audio and visual modalities for decoding at a given timestep. We expect that when the audio is clean, α_a will be higher, since clean audio is usually sufficient to transcribe a word. When the audio signal is masked, however, we expect that α_v will increase if the model effectively uses the visual context in the absence of information from the audio signal. We refer to this model as Multimodal ASR with Global Features (MAG), because it utilizes global visual features.

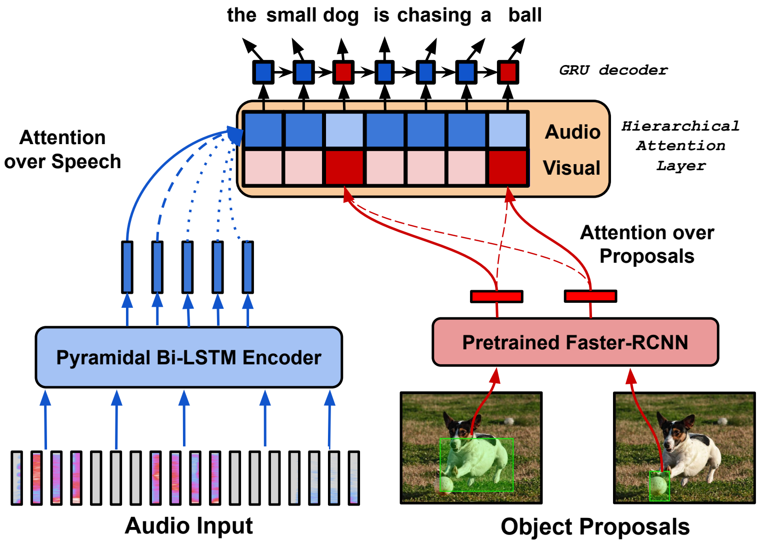

Figure 2: Multimodal ASR with Object Proposals combines attention over object proposals with attention over the audio encoding in hierarchical attention layer to correctly recover masked words in the audio input.

Multimodal ASR with Object Proposals Our proposed model, Multimodal ASR with Object Proposals (MAOP), utilizes visual features from a set of object proposals, instead of the full image. The intuition behind this is that looking at object proposals can help the model localize the most important visual information at a given timestep. Identifying the relevant object proposal(s), rather than looking at the complete image, can ease the burden of transcription on the decoder. For example, it is easier for the decoder to generate a color adjective to describe an object if it extracts visual features directly from the relevant object proposal, rather than from a global visual feature vector.

Concretely, for every image $\mathbf{I}$, we extract $\mathbf{N}$ object proposals $\mathbf{p_{1 \ldots N}}$, where each object proposal is a rectangular patch of the image that expected to contain an object. For each object proposal $\mathbf{p_j}$, we extract visual features $\mathbf{v_j}$ for that patch, in an identical manner to how they were extracted for the entire image. At every decoding timestep, the model estimates an attention distribution over the object proposal features $\mathbf{v_{1 \ldots N}}$, which gives a weighted visual representation vector $\mathbf{v_t^{att}}$. Finally, the decoder has a hierarchical attention mechanism that attends over the encoder context $\mathbf{z_t}$ and the visual representation $\mathbf{v_t^{att}}$.

$$\mathbf{v_t^{att}} = \text{Attention}(\mathbf{v_{1 \ldots N}}, \mathbf{h_t^{dec1}}) \quad (7)$$

$$\mathbf{z_t^{hier}} = \text{Attention}(\{\mathbf{z_t}, \mathbf{v_t^{att}}\}, \mathbf{h_t^{dec1}}) \quad (8)$$

$$\mathbf{h_t^{dec2}} = \text{GRU}_2(\mathbf{z_t^{hier}}, \mathbf{h_{t-1}^{dec2}}) \quad (9)$$

We want $\mathbf{v_t^{att}}$ to be representative of the most important object proposal(s) at that decoding timestep. This hierarchical attention allows the model to both identify which parts of the visual and speech context are relevant for the current decoding timestep, as well as which modality is more important. Figure 2 illustrates the structure of our MAOP model.

2.3 Audio Masking

Previous work has shown that the audio signal needs to be degraded during training in order to utilize the visual context (Srinivasan et al., 2019). We simulate a degradation of the audio signal during training by randomly masking words with silence. This approach extends Srinivasan et al. (2020), where they masked a fixed set of words corresponding to entities, *i.e.*, objects and places. The justification for random word masking, as opposed to entity masking, is that noise in audio signals is unlikely to systematically occur when someone is speaking about an entity. Instead, multimodal ASR models should be responsive to missing audio across the linguistic spectrum.

In real-world settings, the rate at which the speech is dropped is highly variable. Therefore, we train models with an augmented version of the dataset: for each audio utterance, we create four masked audio samples, where words are masked with 0%, 20%, 40% and 60% probability. Note that the text transcript ($\mathbf{y_{1 \ldots T}}$) and image ($\mathbf{v}$) remain intact. This approach to augmenting the dataset will result in models that can adapt to different amounts of corruption in the audio signal during evaluation.

3 Experimental Setup

3.1 Dataset

We perform experiments on the Flickr 8K Audio Caption Corpus (Harwath and Glass, 2015, FACC), which contains 40K spoken captions (total 65 hours of speech) corresponding to 8K natural images from the Flickr8K dataset (Hodosh et al., 2015). The augmented dataset that we use for training and testing (Section 2.3) consists of 160K spoken captions: each caption in the original dataset has four corresponding captions in the augmented dataset.

In addition to the FACC dataset, we use the SpeechCOCO dataset (Havard et al., 2017) to pre-train our models. SpeechCOCO contains over 600 hours of synthesised speech paired with images, as opposed to natural speech in the FACC dataset.

3.2 Acoustic Features

We extract 43-dimensional filter bank features from 16kHz raw speech signals. In order to mask the audio, we first extract word-audio alignments from a pre-trained Gaussian Mixture model-HMM model trained on the Wall-Street Journal Corpus, and expand the start and end timing marks by 25% of the segment duration to account for misalignments. We mask words in the audio by replacing word segments with 0.5 seconds of silence.

3.3 Global Visual Features

MAG uses a single "global" feature vector extracted from each image. We extract visual features from ResNet-50 CNN (He et al., 2016) pre-trained on ImageNet. We extract 2048-dim average-pooled features, and project these to 256-dim through a learned linear layer: $\mathbf{v} = \mathbf{W} \cdot \text{CNN}(\mathbf{img})$

3.4 Object Proposal Features

MAOP uses multiple image features extracted from object proposals. We extract object proposals using a Faster-RCNN object detection model (Ren et al., 2015) with a ResNet-101 CNN backbone (He et al., 2016). We use an implementation[2] that is pre-trained on Visual Genome dataset (Krishna et al., 2017). We extract a feature vector for each proposal $\mathbf{p_j}$ from the 2048-dim average pooling layer of the CNN for $\mathbf{N} = 36$ proposals. Similar to the Global Visual Features, features for each proposal are projected to 256-dim through a learned linear layer: $\mathbf{v_j} = \mathbf{W} \cdot \text{CNN}(\mathbf{p_j})$.

3.5 Model Implementation

All models are trained using Adam optimizer (Kingma and Ba, 2014), with a learning rate of 0.0004, decay of 0.5 and batch size of 36. The encoder and decoder GRU both have 256 hidden units. The embedding dimension for the decoder is also 256, and the input and output decoder embeddings are tied (Press and Wolf, 2017). The norm of the gradient is clipped with a threshold of 1 (Pascanu et al., 2012). UNIMODAL has 8.3M parameters, while MAG and MAOP have 9.1M parameters each.

Models are trained using the *nmtpytorch* framework (Caglayan et al., 2017). We first pre-train our models on the SpeechCOCO dataset, which is also Augmented with masked speech. For every model described in Section 2, we train models on FACC using several checkpoints from the SpeechCOCO pre-training, and choose the model with the best development WER on the Augmented development set. This pre-training step, inspired by Ilharco et al. (2019), was crucial to ensure stable training of our models on the FACC dataset. Models take ≈ 5-6 hours to train on the FACC dataset.

3.6 Evaluation Metrics

Our model development (and the associated results) is conducted on the development set of the Flickr8K Audio Captions Corpus; the rest of our analysis is conducted on the test set. We report **Word Error Rate** (WER) for all our models, and for datasets with masked audio, we compute **Recovery Rate** (RR) (Srinivasan et al., 2020), which measures the percentage of masked words in the dataset that are correctly recovered in the transcription:

$$\text{RR} = \frac{|\text{correctly transcribed masked words}|}{|\text{masked words in dataset}|}$$

In addition, we calculate the contribution of the visual signal when decoding each word in the Multimodal ASR models by inspecting the attention weights of the audio and visual modalities in the hierarchical attention layer. We introduce a new metric to quantify this: **Grounding Rate**. Grounding Rate measures the percentage of correctly recovered words which had a higher visual attention weight than normal (quantified by $\mathbb{E}[\alpha_v]$). $\mathbb{E}[\alpha_v]$ is computed as the average of α_v over all decoding timesteps in the Augmented development set:

$$\text{GR} = \frac{|\text{recovered words where } \alpha_v > \mathbb{E}[\alpha_v]|}{|\text{correctly recovered masked words}|}$$

[2] https://github.com/peteanderson80/bottom-up-attention

	↑ Recovery Rate (%)				↓ Word Error Rate (%)				
Masking Percentage	Aug.	20%	40%	60%	Aug.	0% (Clean)	20%	40%	60%
UNIMODAL	29.2	37.4	31.4	25.0	33.8	**13.6**	**25.9**	40.2	56.8
MAG	33.5	40.1	34.9	30.4	33.3	13.8	26.1	39.8	54.8
MAOP	**36.3**	**41.5**	**37.3**	**33.2**	**32.8**	14.1	26.1	**39.1**	**53.6**

(a) Recovery Rate (RR) and Word Error Rate (WER) of the ASR models on the FACC development set.

		Nouns	Places	Adjectives	Colors	Verbs	Adverbs	Cardinals
RR(%)	UNIMODAL	37.6	29.3	27.4	27.2	27.6	28.4	56.1
	MAG	48.2	39.5	30.1	29.9	**29.3**	**30.6**	56.7
	MAOP	**52.4**	**42.4**	**38.0**	**46.0**	29.0	26.7	**57.4**
GR (%)	MAG	88.1	88.1	68.5	66.3	50.2	25.5	**90.9**
	MAOP	**91.5**	**91.8**	**87.0**	**92.0**	**58.7**	**28.5**	87.9

(b) Comparison of Recovery Rate (RR) and Grounding Rate (GR) of our ASR models on different word categories.

Table 1: Results on the Flickr8K Audio Captions development set.

It has been noted that attention does not always provide a perfect explanation for an observed phenomenon (Jain and Wallace, 2019; Serrano and Smith, 2019). In this paper, we examine attention to determine whether the weights align with our intuition of how the masked words are recovered, i.e. does the model recover words using the visual modality and the correct object proposal? We also use the attention distribution to conduct a quantitative object localization analysis in Section 4.1.

4 Results and Analysis

In Table 1a, we summarize the performance of our three ASR models - UNIMODAL, MAG and MAOP. We examine performance on the Augmented development set, which is constructed similarly to our training set described in Section 2.3, consisting of samples with 0%, 20%, 40% and 60% of words masked. We also evaluate the models on datasets constructed at each individual masking level (i.e. individual datasets where words are masked with 20%, 40%, 60% probability).

First, we find that the multimodal ASR models outperform the UNIMODAL model in terms of recovery rate, and that the difference increases as the masking rate increases from 20% to 60%. The Word Error Rate of the UNIMODAL model is slightly lower than the multimodal models for clean data, but these models perform much better than UNIMODAL with higher speech masking rates. Furthermore, the MAOP model that operates over object proposals substantially outperforms the MAG model, which uses a single global visual vector, on both metrics and at all masking levels.

We now turn our focus to analysing which types of words are best recovered by our multimodal models. We conduct this analysis across seven categories: five syntactic (nouns, verbs, adjectives, adverbs and cardinals) and two semantic (places and colors).[3] For each category, we create a new test set where we mask all occurrences of words belonging to that category.

In Table 1b, we report the recovery rate for our models on the different word categories. We see that MAG and MAOP are good at recovering entities (nouns and places) as well as their properties (adjectives and colors), but they perform similarly to UNIMODAL for other types. Furthermore, we see that while MAOP outperforms MAG on almost all word categories, the improvements on adjectives and colors are most significant. This shows that using object proposals gives the model a more fine-grained view of the entities and their attributes.

We also report the Grounding Rate of the multimodal models in Table 1b. When more groundable words are masked (*i.e.*, entities and adjectives), the Grounding Rate is higher, indicating that the models recover these words by using the visual modality. We also see that MAOP not only recovers more masked adjectives and colors, but also has a higher

[3] Syntactic categories' words were found by POS tagging the corpus and keeping the category's top 100 frequent words.

Grounding Rate for those categories. These results indicate that the model is using the hierarchical attention layer when it recovers groundable words.

4.1 What Are You Looking At? Analyzing the Attention Over Object Proposals

In the previous section, we showed that object proposals provide useful features for multimodal ASR models. We now turn our focus to examining whether this model is right for the right reasons.

We first investigate the attention distribution over object proposals, to determine if it is uniformly distributed over the proposals or concentrated over particular regions of interest. The object proposals are ranked for a given sample according to their attention weights, from which we compute the average proposal attention at each rank across all correctly recovered words in the Augmented development dataset. We observe that most of the proposal attention ($\approx 70\%$) is concentrated among the top 3 proposals, with 40% going to a single proposal alone. This shows that not only does the model use the visual modality, it is also able to identify a proposal that it expects to be relevant for recovering a masked word.

Given that MAOP focuses its attention distribution on one or few proposals at a time, we analyze how closely the attended object proposals match the words they are used to recover. We conduct this analysis using the ground-truth bounding box annotations from the Flickr30K Entities dataset[4] (Plummer et al., 2015) by repurposing the Intersection over Union metric (IoU) from the object detection literature (Russakovsky et al., 2015). Specifically, we compute **IoU Precision @ K** as follows:

1. For every correctly recovered word, we extract the top-K proposals at that decoding timestep.
2. We find the bounding box annotation(s) in the Flickr30K dataset for all phrases in that sentence which contain the recovered word, ignoring words that do not have a bounding box annotation.
3. From the top-K proposals and bounding boxes, we find the proposal-bounding box pair that has the highest Intersection over Union.
4. We compute IoU Precision as the percentage of samples whose `Proposal-IoU` > 0.5.

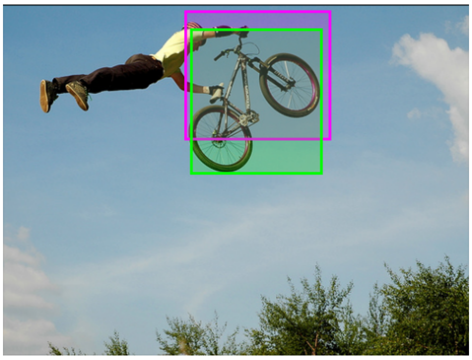

Figure 3: A **localized proposal** and **ground truth bounding box** for recovering **bike** (IoU = 0.72)

This metric computes the percentage of correctly recovered words for which the localized object proposal had a minimum IoU of 0.5 with a ground truth bounding box annotation. Figure 3 shows an example of a maximally attended proposal and a ground-truth bounding box annotation.

In Table 2, we summarize the IoU Precision @ K for our MAOP model for different values of K, for the groundable word categories.[5] We compare the IoU Precision from our top-K proposals with a Random-K baseline, where we pick K of the 36 object proposals randomly, instead of using the attention distribution. We see that our top-K proposals have a significantly higher IoU Precision than the Random-K baseline across all word types, with $\approx$ 45% of the maximally attended proposals overlapping with the ground truth bounding box, and $80 - 83\%$ of the top-5 attended proposals overlapping. The results verify that not only is MAOP focusing on a few proposals, but also the attended proposals are verifiably useful for recovering masked words.

Proposals	Nouns	Places	Adj.	Colors
Random-1	5.9	6.8	5.9	6.8
Top-1	44.7	45.3	46.3	49.4
Random-3	17.4	16.6	17.1	15.5
Top-3	71.7	68.9	70.2	71.7
Random-5	27.2	26.5	26.2	28.1
Top-5	83.6	82.3	80.4	83.2

Table 2: Intersection over Union Precision @ K (%) across four different groundable word catergories.

[4] The Flickr30K dataset is a superset of the Flickr8K dataset. For every caption, Flickr30K Entities contains bounding box annotations for the phrases within the sentence.

[5] Verbs and adverbs did not have enough ground-truth bounding box annotations in Flickr30K Entities for this analysis. Cardinals are discussed in more detail in Section 4.3.

	UNIMODAL	MAOP
Nouns	96.0	96.1
Places	90.3	89.0
Adjectives	93.6	93.1
Colors	94.8	94.2
Verbs	93.9	94.1
Adverbs	88.9	88.0
Cardinals	97.1	97.0

Table 3: Word Accuracy (%) for UNIMODAL and MAOP when transcribing clean, unmasked audio.

4.2 Performance on Clean Speech

MAOP is useful for recovering words which are masked in the speech input but we also want to know how it performs on clean speech sequences. We inspect the transcriptions on a clean, unmasked version of the test set, and calculate a **Word Accuracy (WA)** for different word categories. WA captures the percentage of words belonging to the different word categories which are correctly transcribed from the clean audio signal.

In Table 3, we observe that MAOP performs on par with UNIMODAL on all word categories. This indicates that the visual modality makes no difference when the audio is clean; however, this could be an artefact of the FACC corpus, which is composed of read speech of highly structured captions, and is thus a relatively easy dataset for ASR models. We believe that in more difficult and real-world scenarios (*e.g.*, with different accents and types of speech), MAOP could use the visual modality to improve transcription without the random word masking used in this paper.

4.3 Case Study: The Curious Case of Cardinals

MAOP is better than UNIMODAL at recovering entities and their attributes but both models perform similarly at recovering masked cardinals (see Table 1b). Interestingly, the Grounding Rate of MAOP for cardinals is high (87.9%), which shows that the model uses the visual modality, but to limited effect. One reason for this discrepancy could be that counting entities is difficult if they are not clearly distinguishable due to visual clutter (Rosenholtz et al., 2007). Another reason could be the non-uniform distribution of cardinals in the dataset: $\approx 60\%$ of the cardinals are the number *two*, leading the model to learn a biased distribution.

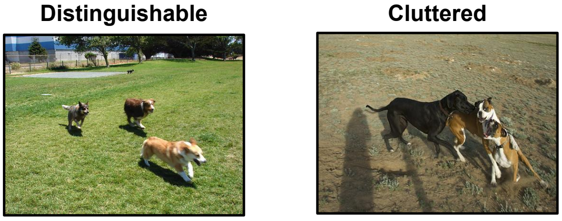

Figure 4: The images used to evaluate the ability of MAOP to count entities (Section 4.3).

As a case study, we evaluate the sensitivity of our model to visual clutter using 49 samples in the test set that contain the *cardinal-entity* phrase "*three dogs*" where the cardinal is masked in the audio. In these 49 samples, the model should be able to use the visual context to correctly recover the missing words "*three*" but the recovery rate is only 24.5%. We also chose two images from the dataset: one containing three clearly distinguishable dogs, and one containing three dogs which are hard to distinguish from each other, as shown in Figure 4. We proceed to calculate the recovery rate of the masked cardinal in these 49 examples with either the distinguishable or the cluttered image as the visual context, instead of the original images.

We find that recovery of *three* in the noun phrase *three dogs* is almost perfect using the image with the distinguishable entities (93.9%), and very low when using the cluttered image, where the entities are hard to distinguish (2.0%). This shows that MAOP is capable of counting entities when they are easy to process in the visual context. Recall that the recovery rate when the original image is provided is only 24.5%; we conducted a manual analysis of the 49 images in this case study and found that $\approx 55\%$ of them were cluttered with entities that were hard to distinguish. We leave a more thorough analysis of a broader range of object types for future work.

4.4 Qualitative Analysis

Figure 5 presents qualitative examples in which words are masked in the speech sequence and recovered in the transcription. We also visualize the object proposal with maximum attention at each step, along with a relative weight of visual modality weight α_v in the hierarchical attention layer.

In the first example, the model correctly localizes the relevant part of the image for the two masked words (*dog* and *ball*) at each step and recovers these words correctly. Moreover, α_v is relatively higher for both those words, compared to the rest

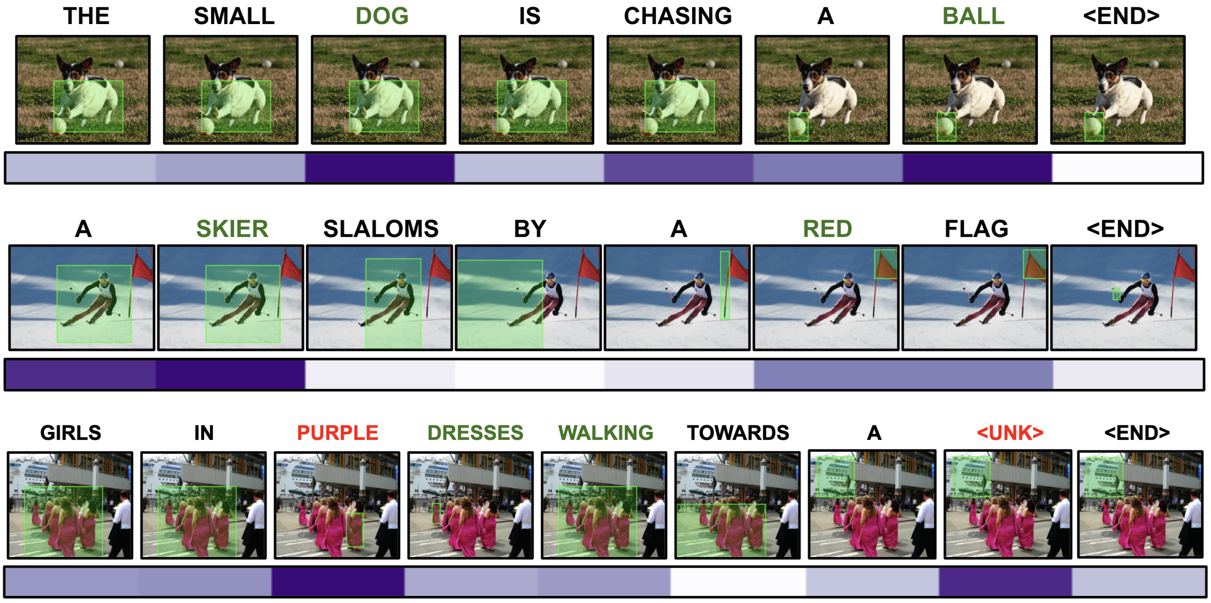

Figure 5: Proposal and hierarchical attentions for threes samples. We present the hypothesis transcription (with recovered and unrecovered masked words), along with the maximum attended proposal (highlighted image patch) and relative hierarchical visual attention (darker shade of purple indicates higher α_v) at each decoding timestep.

of the generated sequence. The second example is similar: the model correctly localizes the objects relevant for recovering the masked words. Interestingly, the model isolates the correct proposal for the second masked word, *red*, and extracts the relevant attribute as well.

The final example shows where MAOP both succeeds and fails. The masked words *dresses* and *walking* are correctly recovered using the correct locations. However, for *purple* and *sign*, the model attends to the correct proposals, but fails to recover the words *pink* and *ship*, respectively.

We note that the object proposal attention is fairly stable across time: the same proposal is often attended to across the length of an entire phrase, rather than jumping around the image.

5 Related Work

Inspired by studies of human perception, multimodal processing is spreading into many traditional areas of research, *e.g.*, machine translation (Sulubacak et al., 2019) and ASR (Palaskar et al., 2018). It has become an important part of new areas of research such as image captioning (Bernardi et al., 2016), visual question-answering (VQA; (Antol et al., 2015)), and multimodal summarization (Palaskar et al., 2019).

The representation and integration of visual context in multimodal ASR systems is an active area of research. Previous approaches incorporate image representations either in the acoustic model (Miao and Metze, 2016), the language model (Gupta et al., 2017; Naszadi et al., 2018), or in end-to-end models (Sanabria et al., 2018). Caglayan et al. (2019) and Moriya and Jones (2018) explore different types of multimodal representations such as image-scene representations and titles of instructional videos respectively. Although all these integration methods show improvements over unimodal baselines, it is not clear when such approaches perform better, and which representations are best.

It has been argued that traditional multimodal architectures do not necessarily take advantage of image semantics in different tasks. Caglayan et al. (2019) showed that multimodal ASR models trained with *shift adaptation* (Miao and Metze, 2016)[6] use the image as a regularization signal. In a similar direction, Elliott (2018) showed that misaligment between image and text representations do not affect multimodal MT models. Ramakrishnan et al. (2018) and Grand and Belinkov (2019) showed that traditional VQA neural architectures ignore the visual context and focus on linguistic biases of the dataset. More related to our work are the studies of Srinivasan et al. (2020) and Caglayan et al. (2019), which explore how multimodal models use image information under noisy scenarios. These studies conclude that when certain nouns

[6] A linear transformation conditioned on the visual features is applied on the audio features.

are dropped from the dominant language modality, multimodal models are capable of properly using the semantics provided by the image. However, unlike this work, their explorations are limited to nouns and not expanded to other types of words.

From an image representation perspective, previous works have studied the utility of using local representations, rather than global ones for multimodal language processing tasks. For instance, Xu et al. (2015) show that, by using attention, the model can use different regions of the image while performing image captioning. More recent work shows that bounding boxes (Ren et al., 2015), a discrete variant of attention over images, improve the representation and hence the performance of different tasks such as VQA (Anderson et al., 2018), image captioning (Yin and Ordonez, 2017) and machine translation (Specia et al., 2020). In this work, we apply this methodology to multimodal ASR (see Section 3.4).

6 Conclusions

In this work, we introduce a new model for multimodal ASR that attends overs fine-grained object proposals and is capable of recovering words which are masked in the speech signal. We show that our model recovers masked words because it can accurately identify the relevant object proposal(s), and that this ability allows it to not only recover the object when it has been masked in the speech signal, but also the object's attributes.

In future work, we plan to improve our model by masking random speech segments (Park et al., 2019) rather than aligned words. If successful, this methodology would allow us to train and test our multimodal models without the need for word alignments, a current limitation of our framework. We will also experiment with more challenging speech captioning scenarios where speech ambiguities are more likely to occur (Pont-Tuset et al., 2019).

Acknowledgments

This work used the computational resources of the PSC Bridges cluster at Extreme Science and Engineering Discovery Environment (XSEDE) (Towns et al., 2014). We thank Stella Frank for discussions about whether such a model could be expected to count objects in images.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philemon Brakel, and Yoshua Bengio. 2016. End-to-end attention-based large vocabulary speech recognition. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Lawrence Barsalou. 2003. Situated simulation in the human conceptual system. *Language and Cognitive Processes*, 18(5-6):513–562.

Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research*, 55:409–442.

Ozan Caglayan, Mercedes García-Martínez, Adrien Bardet, Walid Aransa, Fethi Bougares, and Loïc Barrault. 2017. Nmtpy: A flexible toolkit for advanced neural machine translation systems. *Prague Bulletion of Math. Linguistics*.

Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. 2019. Probing the need for visual context multimodal machine translation. In *North American Chapter of the Association for Computational Linguistics (NAACL)*.

Ozan Caglayan, Ramon Sanabria, Shruti Palaskar, Loïc Barrault, and Florian Metze. 2019. Multimodal Grounding for Sequence-to-Sequence Speech Recognition. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Desmond Elliott. 2018. Adversarial evaluation of multimodal machine translation. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Gabriel Grand and Yonatan Belinkov. 2019. Adversarial regularization for visual question answering: Strengths, shortcomings, and side effects. In *North American Chapter of the Association for Computational Linguistics (NAACL)*.

Abhinav Gupta, Yajie Miao, Leonardo Neves, and Florian Metze. 2017. Visual features for context-aware speech recognition. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346.

David Harwath and James Glass. 2015. Deep multimodal semantic embeddings for speech and images. In *Automatic Speech Recognition and Understanding (ASRU)*.

William Havard, Laurent Besacier, and Olivier Rosec. 2017. Speech-coco: 600k visually grounded spoken captions aligned to mscoco data set. In *International Workshop on Grounding Language Understanding (GLU)*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition (CVPR)*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2015. Framing image description as a ranking task: Data, models and evaluation metrics (extended abstract). In *International Joint Conference on Artificial Intelligence (IJCAI)*.

Gabriel Ilharco, Yuan Zhang, and Jason Baldridge. 2019. Large-scale representation learning from visually grounded untranscribed speech. In *Computational Natural Language Learning (CoNLL)*.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. In *North American Chapter of the Association for Computational Linguistics (NAACL)*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.

Jindřich Libovickỳ and Jindřich Helcl. 2017. Attention strategies for multi-source sequence-to-sequence learning. In *Association for Computational Linguistics (ACL)*.

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

Yasufumi Moriya and Gareth J. F. Jones. 2018. LSTM language model adaptation with images and titles for multimedia automatic speech recognition. In *Spoken Language Technology Workshop (SLT)*.

Kata Naszadi, Youssef Oualil, and Dietrich Klakow. 2018. Image-sensitive language modeling for automatic speech recognition. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*.

Shruti Palaskar, Jindřich Libovickỳ, Spandana Gella, and Florian Metze. 2019. Multimodal abstractive summarization for how2 videos. In *Association for Computational Linguistics (ACL)*.

Shruti Palaskar and Florian Metze. 2018. Acoustic-to-word recognition with sequence-to-sequence models. In *Spoken Language Technology Workshop (SLT)*.

Shruti Palaskar, Ramon Sanabria, and Florian Metze. 2018. End-to-end multimodal speech recognition. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

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*.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2012. Understanding the exploding gradient problem. *CoRR, abs/1211.5063*, 2.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *International Conference on Computer Vision (ICCV)*.

Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. 2019. Connecting vision and language with localized narratives. *arXiv*, 1912.03098.

Ofir Press and Lior Wolf. 2017. Using the output embedding to improve language models. In *European Chapter of the Association for Computational Linguistics (EACL)*.

Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. 2018. Overcoming language priors in visual question answering with adversarial regularization. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In

Advances in Neural Information Processing Systems (NIPS).

Ruth Rosenholtz, Yuanzhen Li, and Lisa Nakano. 2007. Measuring visual clutter. *Journal of Vision*, 7(2):17–17.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*.

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 *Workshop on Visually Grounded Interaction and Language (ViGIL), NeurIPS*.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673.

Sofia Serrano and Noah A Smith. 2019. Is attention interpretable? In *Association for Computational Linguistics (ACL)*.

L. Specia, R. Arora, L. Barrault, O. Caglayan, A. Duarte, D. Elliott, S. Gella, N. Holzenberger, C. Lala, S. J. Lee, J. Libovicky, P. Madhyastha, F. Metze, K. Mulligan, A. Ostapenka, S. Palaskar, R. Sanabria, and J. Wang. 2020. Grounded Sequence to Sequence Transduction. *IEEE Journal of Selected Topics in Signal Processing*.

Tejas Srinivasan, Ramon Sanabria, and Florian Metze. 2019. Analyzing utility of visual context in multimodal speech recognition under noisy conditions. *arXiv preprint arXiv:1907.00477*.

Tejas Srinivasan, Ramon Sanabria, and Florian Metze. 2020. Looking enhances listening: Recovering missing speech using images. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Umut Sulubacak, Ozan Caglayan, Stig-Arne Grönroos, Aku Rouhe, Desmond Elliott, Lucia Specia, and Jörg Tiedemann. 2019. Multimodal machine translation through visuals and speech. *arXiv preprint arXiv:1911.12798*.

J. Towns, T. Cockerill, M. Dahan, I. Foster, K. Gaither, A. Grimshaw, V. Hazlewood, S. Lathrop, D. Lifka, G. D. Peterson, R. Roskies, J. Scott, and N. Wilkins-Diehr. 2014. Xsede: Accelerating scientific discovery. *Computing in Science and Engineering*, 16(05):62–74.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning (ICML)*.

Xuwang Yin and Vicente Ordonez. 2017. Obj2text: Generating visually descriptive language from object layouts. In *Empirical Methods in Natural Language Processing (EMNLP)*.