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

Evaluation of Audio-Visual Alignments in Visually Grounded Speech Models

Khazar Khorrami[1], *Okko Räsänen*[1,2]

[1]Unit of Computing Sciences, Tampere University, Finland
[2]Department of Signal Processing and Acoustics, Aalto University, Finland
khazar.khorrami@tuni.fi, okko.rasanen@tuni.fi

Abstract

Systems that can find correspondences between multiple modalities, such as between speech and images, have great potential to solve different recognition and data analysis tasks in an unsupervised manner. This work studies multimodal learning in the context of visually grounded speech (VGS) models, and focuses on their recently demonstrated capability to extract spatiotemporal alignments between spoken words and the corresponding visual objects without ever been explicitly trained for object localization or word recognition. As the main contributions, we formalize the alignment problem in terms of an audiovisual alignment tensor that is based on earlier VGS work, introduce systematic metrics for evaluating model performance in aligning visual objects and spoken words, and propose a new VGS model variant for the alignment task utilizing cross-modal attention layer. We test our model and a previously proposed model in the alignment task using SPEECH-COCO captions coupled with MSCOCO images. We compare the alignment performance using our proposed evaluation metrics to the semantic retrieval task commonly used to evaluate VGS models. We show that cross-modal attention layer not only helps the model to achieve higher semantic cross-modal retrieval performance, but also leads to substantial improvements in the alignment performance between image object and spoken words.
Index Terms: cross-modal learning, audio-visual alignment, visual object localization, word segmentation

1. Introduction

Utilization of statistical dependencies between different modalities has the potential to replace or supplement supervised learning in many tasks, as the data streams can potentially be used as weak supervision for each other. As an example of such multimodal systems, so-called visually grounded speech (VGS) models have been recently proposed [1–6]. They are unsupervised audiovisual algorithms that can learn shared semantic concepts between visual and speech data, given a series of unlabeled images and utterances describing them (e.g., [1, 4]). As a result, the models can be used to search for semantically similar content in audio and images. VGS models are also interesting for modeling of infant language learning, which is essentially an unsupervised multimodal learning process with learners having access to both auditory and visual input (see, e.g., [7, 8]).

Although VGS models typically operate based on high-dimensional semantic embeddings from which spatial and temporal characteristics of the data have been abstracted away, the architecture of these models can also be modified to estimate alignments between visual objects and the corresponding spoken words in the input data. For instance, [3] introduced a latent audiovisual tensor that assigns explicit alignment scores on each spatial position and time-step for input pairs of images and utterances. One could also attempt to extract the object-word-alignments from other latent (e.g., audiovisual attention) layers of the models. This makes VGS models promising for unsupervised decoding (segmentation) of image and audio data into their constituent units in an unsupervised manner, and opens up a largely unexplored strategy to learn structural properties of initially unorganized and unlabeled data (e.g., visual object categorization or unsupervised language learning).

Localization of objects within images using text-based queries has been investigated in different lines of research, such as in object detection given a category label [9] and in natural language object retrieval [10–15]. Given a query word, these models either directly try to predict the locations of object bounding boxes in images [12, 14], or by ranking alternative image regions in terms of the potential presence of the relevant object [15]. Even though the majority of existing multi-modal works have focused on phrase-based object detection, the multimodal learning and alignment task can also be viewed as symmetric with respect to both input modalities, i.e., learned objects in the visual input could also be used to segment correct words in the other modality, and vice versa. Moreover, object localization has been mainly evaluated using intersection over union (IoU) between the predicted and true bounding boxes of visual objects. IoU uses hard decisions on what pixels are detected, potentially followed by another thresholding to determine if a specific object is detected or not (see, e.g., [15]).

In [3], the proposed VGS model was evaluated using a number of heuristic approaches, including measuring the capability of the audiovisual tensor to conduct speech-based object localization, to form meaningful audio-visual pattern clusters, and in building an object-word concept dictionary. Although functional, these evaluation strategies are variable, dependent on the characteristics of the selected small test set, and require manual analysis of the results. Hence, in order to enable systematic comparison of alternative models and facilitate further unsupervised multimodal alignment system development, standardized definition of the alignment problem and associated evaluation metrics would be desirable.

Given this background, our present goal is to formally define the problem of multimodal alignment between speech and images, here studied in the context of VGS models, and to define evaluation metrics accord with the problem definition. We start from audiovisual alignment tensors (derived from [3]), and define two metrics that capture complementary aspects of the alignment performance in terms of finding visual objects for spoken words (or phrases), and finding words in speech, given visual objects. We also propose a new VGS model variant for unsupervised audiovisual alignment using audiovisual attention, and show that it outperforms the models from [3] in both alignment and semantic retrieval tasks. Although we place our work in the context of VGS models, the proposed methodology is applicable to both explicit and implicit alignment models, and should generalize beyond image and speech modalities.

http://dx.doi.org/10.21437/Interspeech.2021-496

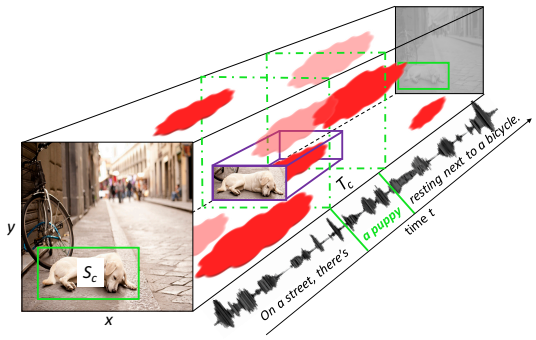

Figure 1: *An illustration of a temporospatial alignment tensor* ***T****[x ,y, t] between an image and an utterance related to the image. Red "clouds" illustrate manifolds where T has positive values. Ground-truth alignment region for [dog] is visualized with a violet box. Conceptually adapted from [3].*

2. Problem and evaluation definition

The basic goal of the alignment process is to identify the link between spoken words or phrases and the corresponding visual objects or concepts in images. Formally, we define this link using an audiovisual alignment tensor $\mathbf{T}[x, y, t] \in \mathbb{R}^3_{\geq 0}, x \in \{1, ..., N_x\}, y \in \{1, ..., N_y\}, t \in \{1, ..., T\}$, which defines the association strength between each pixel $\{x, y\}$ in the image space with each time step t in the speech signal (see Fig. 1 for illustration). $N_x \times N_y$ is the size of the image in pixels, and T is the duration of the utterance in terms of 10-ms signal frames. In $\mathbf{T}[x, y, t]$, zeros stand for no association and the larger the (positive) value, the stronger the model's confidence that the pixels and timestamps are related to each other. The goal of an alignment model is to derive this tensor, given an input image and an utterance. In practice, the tensor may be a final output of a model (e.g., using supervised training) or extracted as a latent representation of a VGS model (as in [3] and all models in our present experiments).

Evaluation of the alignment is based on ground-truth-knowledge on visual objects that are related to specific acoustic words or phrases, basically word/phrase timestamps together with object pixel masks. Given a concept ("class") c, we define $\{x, y\} \in S_c$ as the set of pixels corresponding to visual object o_c, and $\{t_{\text{onset}}, ..., t_{\text{offset}}\} \in T_c$ as the set of time-frames corresponding to the related word w_c. Together they define a subset of $\mathbf{T}[x, y, t]$ corresponding to the ground-truth alignment.

We propose two new primary metrics to evaluate how well $\mathbf{T}$ relates to the ground-truth: *alignment score (AS)* and *glancing score (GS)*, both measured separately in time and space.

2.1. Alignment score

AS measures how well the model attends to the correct visual object(s) throughout a spoken word, or how well the attention across all the pixels of an object focuses on the time-instances of the corresponding word. In the case of object alignment given a spoken utterance, $AS_{\text{object}}(c) \in [0, 1]$ is calculated as follows.

First, each frame in $\mathbf{T}$ is normalized to sum up to one $\mathbf{T}'[x, y, t] = \mathbf{T}[x, y, t] / \sum_{x,y} \mathbf{T}[x, y, t]$ to measure the distribution of "attention" across the image at each time step. Then the proportion of attention on the target object is obtained by

$$AS_{\text{object}}(c) = \frac{1}{|T_c|} \sum_{S_c, T_c} \mathbf{T}'[x, y, t] \quad (1)$$

where $|T_c|$ is the number of frames in w_c. As a result, $AS_{\text{object}}(c)$ obtains a value of 0 if none of the attention is on the target object o_c during the word w_c, and 1 if all attention is located on the object o_c during the entire duration of w_c.

Word alignment score $AS_{\text{word}}(c) \in [0, 1]$ is calculated in an analogous manner, now first normalizing each pixel of $\mathbf{T}$ to sum up to one across the utterance: $\mathbf{T}''[x, y, t] = \mathbf{T}[x, y, t] / \sum_t \mathbf{T}[x, y, t]$, and then measuring the average temporal overlap of the resulting scores and w_c across the o_c pixels:

$$AS_{\text{word}}(c) = \frac{1}{|S_c|} \sum_{S_c, T_c} \mathbf{T}''[x, y, t] \quad (2)$$

If $AS_{\text{word}}(c)$ is 0, the temporal attention at all the pixels of o_c have zero value during w_c, whereas score of 1 means that attention at *all* of the o_c pixels is always completely within the bounds of w_c and zero during other time steps.

2.2. Glancing score

The alignment score enforces each word time-step or object pixel to have consistent attention on the corresponding target in the other signal modality. However, in some use cases we may only be interested whether the model "takes a look" at the correct object once it recognizes a word, or tags the correct word in time given a partial observation of an object. As an example, a temporally causal model simulating human eye-gaze during speech comprehension could "glance" at the correct object once it recognizes the corresponding word [16]. However, such a model would not be able to maintain sustained attention on the target object before identifying the word as it unfolds in time. To evaluate such a behavior, we introduce glancing score GS.

In object glancing score $GS_{\text{object}} \in [0, 1]$, the cumulative attention map across the image during the entire word w_c is first calculated as $\mathbf{A}_c[x, y] = \sum_{t \in T_c} \mathbf{T}[x, y, t]$. This is then normalized to a spatial distribution proper $\mathbf{A}'_c[x, y] = \mathbf{A}_c[x, y] / \sum_{x,y} \mathbf{A}_c[x, y]$, and compared to ground-truth pixels to measure the proportion of attention on the target:

$$GS_{\text{object}}(c) = \sum_{\{x,y\} \in S_c} \mathbf{A}'_c[x, y] \quad (3)$$

The corresponding word glancing score $GS_{\text{word}} \in [0, 1]$ is obtained by first measuring the total object attention as a function of time by $a[t] = \sum_{\{x,y\} \in S_c} \mathbf{T}[x, y, t]$ and normalizing it to have a sum of 1 across the utterance with $a'[t] = a[t] / \sum_t a[t]$. The score is then obtained by

$$GS_{\text{word}}(c) = \sum_{t \in T_c} a'_c[t]. \quad (4)$$

In essence, GS_{object} is the proportion of attention within the spatial extent of object o_c during word w_c whenever $\mathbf{T} > 0$, but the attention does not have to be defined (positive-valued) throughout the duration of the word. In an analogous manner, GS_{word} corresponds to the proportion of attention on the target word w_c compared to other words, given the pixels of o_c, but not all pixels have to have positive attention values in $\mathbf{T}$. The more there is relative attention outside the correct targets, the lower the scores will be. If $\mathbf{A}[x, y]$ or $a[t]$ fully zero before normalization, they are replaced by a uniform distribution across their elements.

Equations above describe scoring for individual word-object pairs w_c-o_c in individual images-utterance samples. In practice, the score should be calculated for each ground-truth pair in each image-utterance-pair of the test set. Also note that confusion errors for both AS and GS can be derived from the equations simply by comparing S_{c_1} and T_{c_2}, where $c_1 \neq c_2$.

3. Model Description

Our models follow the main structure of VGS models (see, e.g., [2–4]), where two branches of neural layers embed data from

speech and image domains, respectively. Then the modality-specific vector embeddings are mapped into a shared semantic space. Input to the system consists of images paired with spoken descriptions of them. Output is a similarity score indicating the semantic relatedness of the input pairs. The network is trained using a triplet loss [1] that tries to assign higher scores to semantically related image-speech pairs compared to unrelated pairs, and with the maximum separation limited by a margin M.

However, unlike other common VGS models (e.g., [1, 4, 5]), encoder outputs of the so-called DAVEnet model [3] maintain modality-specific signal representations as a function of input speech time (1-D) and image position (2-D). The embedding vectors are then mapped ("aligned") together through matrix product, resulting a 3-D tensor $\mathbf{T}[w, h, n]$ spanning in both spatial location (w, h) and time frames (n). This audio-visual alignment tensor $\mathbf{T}$ allows the model to co-localize patterns within both modalities. The overall similarity score between input speech and images can then be obtained by taking the sum and/or maximum of $\mathbf{T}$ in time and/or spatial dimensions, resulting in three different model criteria for training: MISA (max over image, sum over audio), SIMA (sum over image, max over audio), SISA (sum over image, sum over audio) [3], whereas maxing over both dimensions appears to be difficult to train.

In our experiments, the first model variant is similar to the DAVEnet (here: CNN_0). In CNN_0, the speech encoder is stack of five convolutional layers (with layer sizes of [128, 256, 256, 512, 512]) with gradual temporal downsampling with maxpooling over time. VGG16 model [17], up to the last convolutional layer and pretrained on ImageNet data, is used as the image encoder. This is followed by one trainable 2D-convolution layer with 512 filters. Both the speech and image encoder branches are then fed to a dense linear layer, followed by L2-normalization, and then combined into $\mathbf{T}$ with matrix product.

Our CNN_{ATT} (Fig. 2) is obtained from CNN_0 by adding a cross-modal attention layer on top of the $\mathbf{T}[w, h, n]$ to help the model to attend to specific image objects and spoken words using the information from the another modality. This is inspired by the fact that the audiovisual tensor $\mathbf{T}$ introduced in [3] is similar to scoring function applied in dot product attention mechanism [18, 19]. Thus, we extended the model to have a complete attentional module by applying softmax non-linearity separately in space (for a query in time; Fig. 2 left branch) and in time (for a query in space; Fig. 2 right branch) in order to produce a distribution of weights over image space and speech frames, respectively. In parallel to the softmax, we also apply a dense layer with a sigmoid activation function to produce an extra attentional representation, as we found this to outperform the use of softmax only in the retrieval task. These attention scores are then used to produce corresponding spatially weighted representations for audio (left) and time-weighted representations for image (right), followed by average pooling to get rid of absolute positional information. In the final stage, outputs of the softmax and sigmoid layer attentions are concatenated with the original image and speech representations to produce the final speech and image embeddings. These embeddings are then L2-normalized and compared using a similarity score (dot product) in triplet-loss training, and alignments can be extracted from $\mathbf{T}$ or after taking the softmax in time or space.

We tested three alternative variations for the CNN_{ATT} architecture: CNN_{ATT}v0 with the same speech encoder as the CNN_0, CNN_{ATT}v1 with the first maxpool layer removed to obtain $\mathbf{T}$ at a higher temporal resolution of 128 frames (compared to 64 frames in other variants), and CNN_{ATT}v2 which was equal to CNN_{ATT}v0 but with less filters ([64, 128, 256, 256, 512]).

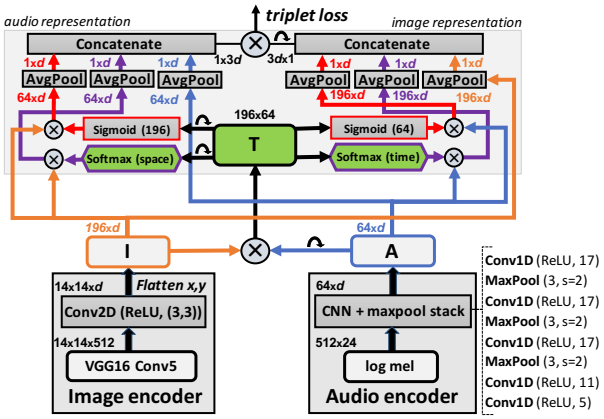

Figure 2: *CNN_{ATT}v0 architecture using a cross-modal attention block on top of DAVEnet model. 64 = number of time frames, 196 = flattened spatial coordinates (x,y), $\curvearrowright$ = transpose. The layers used in alignment evaluation are highlighted with green.*

4. Experimental setup

For our experiments, we used MSCOCO [20], which consists of images of everyday objects and their contexts, together with their speech-synthesized captions from SPEECH-COCO [21]. The dataset includes a total of 123,287 images with 91 common object categories (e.g. dog, pizza, chair) from 11 super-categories (e.g., animal, food). In SPEECH-COCO, each image is paired by five synthetic speech captions describing the scene using the object categories. In our experiments, we used the training set of MSCOCO images and their verbal descriptions for model training and validation. The original validation set of ∼40k images was used as a held-out test set for evaluation, using one randomly chosen spoken caption per image.

SPEECH-COCO contains metadata on words and their timestamps, while MSCOCO contains manually annotated pixel masks for its 80 visual object categories. However, there is no direct one-to-one mapping between image object labels (e.g., [dog]) and spoken words (e.g., "puppy"). We derived ground-truth pairings of objects and words automatically using semantic similarities from a pre-trained Word2vec model [22]. We first extracted nouns of the caption transcripts using NLTK-toolbox, and then used Word2vec cosine similarity between words and object labels as a means of identifying semantically matching pairs. Words and objects with Word2vec similarity above a threshold of 0.5 were considered as a ground-truth word-object pair w_c-o_c (a concept). Selection of this threshold was based on manual observation of the similarity histograms across the dataset, where 0.5 provided a reasonable cutoff point in the somewhat bimodal (but noisy) similarity distribution.

Model training followed the same triplet-loss protocol as in [1–4, 8]. Adam optimizer was used for both models initial learning rate of lr =10e-4 and triplet loss margin $M = 0.1$. In the speech processing channel, we applied ReLU activation after each convolutional layer, followed by batch normalization. We measured *recall@10* audiovisual semantic retrieval score [23] using the representation similarity scores to ensure that the models have learned the semantic relationships between the two modalities. For each variant, we saved the best model based on the recall score of the validation set. In both CNN_0 and CNN_{ATT} models, recall scores were measured using speech and image embedding layers before audiovisual $\mathbf{T}$ (I and A in Fig. 2). AS and GS scores were then measured for each word-object

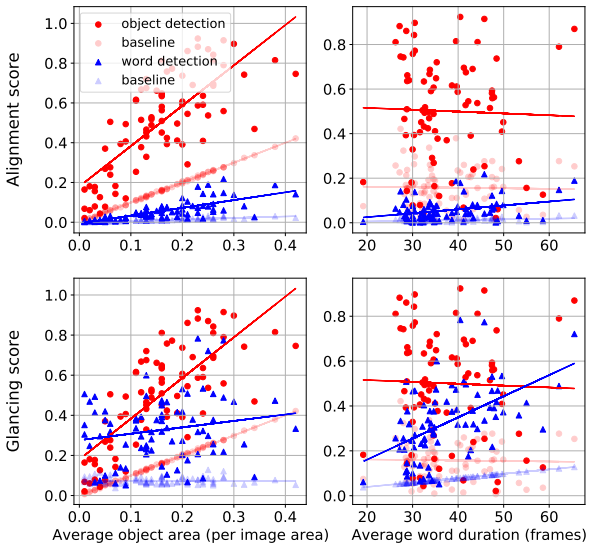

Figure 3: *Relationship of alignment score (top) and glancing score (down) with average object size and average word duration (for CNN$_{\text{ATT}}$v1 softmax). Solid lines = linear fits to the data. Shaded lines = corresponding fits to random baselines.*

pair in the test set. Total score was calculated as the average of the 80 class-specific scores, where each class-based score was computed as the average score of all samples belonging to that class. We also report baseline results for a system that assigns random values to **T** (from uniform [0, 1]).

For the CNN$_0$ and CNN$_{\text{ATT}}$, AS and GS scores were measured for the tensor **T**[*w*,*h*,*t*]. For CNN$_{\text{ATT}}$, we also measure alignments from the **T** after the attention module softmax functions (Fig. 2, green boxes), applied in spatial and temporal directions for the object and word detection tasks, respectively. For both models, we upsampled **T**[*w*,*h*,*n*] to **T**[*x*,*y*,*t*] to match with the original image size and the target 10-ms time resolution of the evaluation protocol.

5. Results

Table 1 shows the recall@10 scores obtained for the different models for both speech-to-image and image-to-speech search. The scores show that all the models learned to solve the audiovisual retrieval task, and that the performance is higher for CNN$_{\text{ATT}}$ than CNN$_0$. Results for the CNN$_{\text{ATT}}$v2 also show that the use of cross-modal attention leads to higher retrieval performance with a smaller number of parameters than in CNN$_0$. In general, CNN$_{\text{ATT}}$ models were faster to train and converged before 70 epochs. In comparison, a maximum of 150 epochs was required for the CNN$_0$ to converge.

Table 2 shows the results for the object and word alignment and glancing scores together with the random baseline. As can be seen, audiovisual tensors obtained from all model variants can produce alignment and glancing scores clearly above what is obtained by chance. However, all the scores are higher for CNN$_{\text{ATT}}$. Comparing the CNN$_0$ variants, one can also observe that maxpooling across time or space (SIMA and MISA) improves from the SISA variant in the respective dimension. In fact, we also tried to train MIMA (maxpooling in time and space simultaneously), but the model did not train well (in terms of validation recall@10) even with a large number of epochs.

In terms of CNN$_{\text{ATT}}$, aligning and glancing scores increase substantially for the softmax weighted tensor compared to the preceding audiovisual tensor. In fact, among all tested variants, only the softmax tensors produced alignment and glancing scores above 0.5. This further indicates that application of cross-modal attention helps the model to focus on semantically relevant image objects and spoken words, and that the model finds the correspondences between two modalities more accurately than without attention. As can be seen, both aligning and glancing scores increase for the CNN$_{\text{ATT}}$v1 model (compared to CNN$_{\text{ATT}}$v0) with increased temporal resolution of **T**.

Table 1: *Recall@10 scores for speech-to-image and image-to-speech search tasks for the compared models. "T-dim" refers to the number of time frames in **T** for encoding 5.12 s of speech.*

Model	**parameters**	**T-dim**	**speech-to image**	**image-to speech**
CNN$_0$	11,072,128	64	-	-
SISA			0.489	0.487
MISA			0.436	0.305
SIMA			0.401	0.446
CNN$_{\text{ATT}}$ v0	10,587,540	64	**0.562**	**0.559**
CNN$_{\text{ATT}}$ v1	13,943,508	128	0.547	0.531
CNN$_{\text{ATT}}$ v2	6,401,748	64	0.536	0.544

Table 2: *Alignment and glancing scores for the model variants.*

Model	AS_{object}	AS_{word}	GS_{object}	GS_{word}
baseline	0.158	0.011	0.158	0.073
CNN$_0$				
SISA	0.200	0.016	0.200	0.102
MISA	0.270	0.020	0.267	0.132
SIMA	0.223	0.041	0.222	0.214
CNN$_{\text{ATT}}$				
v0	0.285	0.028	0.282	0.168
v1	0.292	0.038	0.295	0.227
v2	0.279	0.030	0.277	0.180
v0 softmax	0.504	0.052	0.504	0.317
v1 softmax	**0.518**	**0.076**	**0.518**	**0.446**
v2 softmax	0.501	0.056	0.501	0.327

Finally, Fig. 3 illustrates how the AS and GS depend on ground truth visual object size and word duration (for CNN$_{\text{ATT}}$v1 softmax). Both object detection (red) and word detection (blue) scores increase with increasing visual object size. They also increase more than what is expected by chance with larger targets. Fig. 3 also shows that AS and GS for word detection increase for longer words, whereas the object detection scores are not affected by the word duration. Similar patterns were observed across all the compared models.

6. Conclusions

In this paper, we proposed alignment and glancing scores as two novel metrics for evaluating performance of visually grounded speech (VGS) systems in an alignment task between visual objects and spoken words. We also introduced a VGS model variant based on cross-modal attention, and compared how different VGS models perform on the alignment task using our proposed metrics. The results show that the metrics can capture different aspects of cross-modal alignments, and that the attention-based VGS models outperform the earlier non-attentional alignment approach in both alignment and semantic retrieval tasks.

7. Acknowledgements

This research was funded by Academy of Finland grants no. 314602 and 320053. VGS model codes are available for download at https://github.com/khazarkhorrami/VGS_alignment/

8. References

[1] D. F. Harwath, A. Torralba, and J. R. Glass, "Unsupervised learning of spoken language with visual context," in *Proceedings of the Advances in Neural Information Processing Systems 29 (NIPS 2016)*, December 5–10, Barcelona, Spain, 2016, pp. 1858–1866.

[2] D. Harwath and J. R. Glass, "Learning word-like units from joint audio-visual analysis," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*, July 30–August 4, Vancouver, Canada, 2017, pp. 506–517.

[3] D. 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 *15th European Conference on Computer Vision (ECCV 2018)*, September 8–14, Munich, Germany, 2018, pp. 659–677.

[4] G. Chrupała, L. Gelderloos, and A. Alishahi, "Representations of language in a model of visually grounded speech signal," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, July 30–August 4, Vancover, Canada, 2017, pp. 613–622.

[5] D. Merkx, S. L. Frank, and M. Ernestus, "Language learning using speech to image retrieval," in *Proceedings of the 20th Annual Conference of the International Speech Communication Association (Interspeech 2019)*, September 15–19, Graz, Austria, 2019, pp. 1841–1845.

[6] M. S. Mortazavi, "Speech-image semantic alignment does not depend on any prior classification tasks," in *Proceedings of the 21th Annual Conference of the International Speech Communication Association (Interspeech 2020)*, October 25–29, Shanghai, China, 2020, pp. 3515–3519.

[7] O. Räsänen and K. Khorrami, "A computational model of early language acquisition from audiovisual experiences of young infants," in *Proceedings of 20th Annual Conference of the International Speech Communication Association (Interspeech 2019)*, Graz, Austria, September 2019, pp. 3594–1598.

[8] K. Khorrami and O. Räsänen, "Can phones, syllables, and words emerge as side-products of cross-situational audiovisual learning?-a computational investigation," *submitted for publication*. [Online]. Available: https://psyarxiv.com/37zna

[9] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, July 21–26, Honolulu, Hawaii, 2017, pp. 7263–7271.

[10] A. Karpathy, A. Joulin, and L. Fei-Fei, "Deep fragment embeddings for bidirectional image sentence mapping," in *Proceedings of the 27th International Conference on Neural Information Processing Systems – Volume 2*, 2014, pp. 1889–1897.

[11] Y. Qiao, C. Deng, and Q. Wu, "Referring expression comprehension: A survey of methods and datasets," *IEEE Transactions on Multimedia*, online early access, 2020.

[12] R. Hu, H. Xu, M. Rohrbach, J. Feng, K. Saenko, and T. Darrell, "Natural language object retrieval," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 27–30, Las Vegas, Nevada, 2016, pp. 4555–4565.

[13] R. Luo and G. Shakhnarovich, "Comprehension-guided referring expressions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, July 21–26, Honolulu, Hawaii, 2017, pp. 7102–7111.

[14] A. Rohrbach, M. Rohrbach, R. Hu, T. Darrell, and B. Schiele, "Grounding of textual phrases in images by reconstruction," in *European Conference on Computer Vision*, October 11–14, Amsterdam, Netherlands, 2016, pp. 817–834.

[15] J. Mao, J. Huang, A. Toshev, O. Camburu, A. L. Yuille, and K. Murphy, "Generation and comprehension of unambiguous object descriptions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 27–30, Las Vegas, Nevada, 2016, pp. 11–20.

[16] R. M. Golinkoff, W. Ma, L. Song, and K. Hirsh-Pasek, "Twenty-five years using the intermodal preferential looking paradigm to study language acquisition: what have we learned?" *Perspectives on Psychological Science*, vol. 8, pp. 316–339, 2013.

[17] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and F. Li, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, pp. 211–252, 2015.

[18] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, September 17–21, Lisbon, Portugal, 2015, pp. 1412–1421.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *31st Conference on Neural Information Processing Systems (NIPS 2017)*, December 4–9, Long Beach, California, 2017.

[20] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *European Conference on Computer Vision (ECVV 2014)*, September 6–12, Zurich, Switzerland, 2014, pp. 740–755.

[21] W. Havard, L. Besacier, and O. Rosec, "SPEECH-COCO: 600k visually grounded spoken captions aligned to MSCOCO data set," *arXiv pre-print*, 2017. [Online]. Available: http://arxiv.org/abs/1707.08435

[22] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proceedings of 1st International Conference on Learning Representations (ICLR 2013)*, May 2–4, Scottsdale, Arizona, 2013.

[23] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *Journal of Artificial Intelligence Research*, vol. 47, pp. 853–899, 2013.