ALIGN OR ATTEND? TOWARD MORE EFFICIENT AND ACCURATE SPOKEN WORD DISCOVERY USING SPEECH-TO-IMAGE RETRIEVAL

Liming Wang[1], *Xinsheng Wang*[2,3], *Mark Hasegawa-Johnson*[1], *Odette Scharenborg*[3], *Najim Dehak*[4]

[1]Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign,
[2]School of Software Engineering, Xian Jiaotong University,
[3]Multimedia Computing Group, Delft University of Technology,
[4]Center for Language and Speech Processing, Johns Hopkins University

ABSTRACT

Multimodal word discovery (MWD) is often treated as a byproduct of the speech-to-image retrieval problem. However, our theoretical analysis shows that some kind of alignment/attention mechanism is crucial for a MWD system to learn meaningful word-level representation. We verify our theory by conducting retrieval and word discovery experiments on MSCOCO and Flickr8k, and empirically demonstrate that both neural MT with self-attention and statistical MT achieve word discovery scores that are superior to those of a state-of-the-art neural retrieval system, outperforming it by 2% and 5% alignment F1 scores respectively.

Index Terms— Multimodal learning, spoken term discovery, language acquisition, low-resource speech technology

1. INTRODUCTION

Multimodal word discovery (MWD) is a form of distant supervision in which a learner, provided with nothing except a set of images and their spoken descriptions, tries to learn the words corresponding to each visible object. Early MWD systems modeled each word using sequential replay memory [1] or hidden Markov models (HMM, [2]), but the most successful recent systems have treated MWD as a by-product of the *speech-to-image retrieval* problem [3]. A speech-to-image retrieval model is trained to retrieve images from a large database in response to a spoken query; the pretrained retrieval system can then discover word-like units by searching for visible regions of interest (ROI) and spoken segments with the highest similarity scores. The retrieval-based approach pioneered in [3] has been used in a number of recent systems for multimodal word discovery, e.g., based on similarity between audio and visual ROI embeddings [4, 5, 6, 7]. Three more recent papers seek to cluster the audio embeddings in order to recognize repeated words across multiple utterances [4, 8, 9]. In [4], dot-product similarity profiles between the two modalities are clustered through a combination of Bayesian and spectral clustering. More recently [9] employed a vector-quantized variational autoencoder [10] to learn the discrete representation of speech directly during training of the retrieval system. Retrieval-based word discovery systems are able to learn words from a much larger vocabulary than earlier HMM-based systems with reasonably high precision, but with low recall: retrieval-based systems are trained to represent the meaning of a sentence globally, but do not necessarily therefore learn the alignment of meaning to particular words [6].

Inspired by statistical machine translation (SMT) [11] and neural machine translation (NMT) [12], Wang and Hasegawa-Johnson [13] proposed a DNN-HMM-DNN multimodal alignment system that exploits the repetition of similar concepts in multiple images in order to learn an ROI-to-word alignment matrix (in analogy to SMT) or attention matrix (in analogy to NMT) that is complementary to the retrieval-based approach. Indeed, to the extent that performance can be compared, the retrieval-based and translation-based approaches have complementary characteristics; retrieval-based word discovery has high precision but low recall [4], while translation-based word discovery systems have higher recall but lower precision [13].

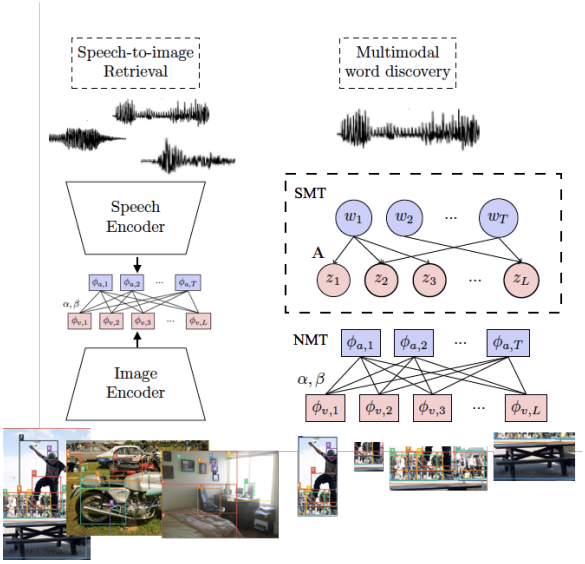

Fig. 1: Model architectures of the speech-to-image retrieval and word discovery systems

In this paper, we analyze and visualize the tradeoff between speech-to-image retrieval performance and word discovery performance. With a better understanding of the relation between the two tasks, more efficient and accurate MWD systems are proposed as a step toward the eventual goal of joint word discovery and retrieval learning. Further, we propose to use alignment F1 used in evaluating MT systems to better assess the quality of word-level representation learned by the discovery system.

2. PROBLEM FORMULATION

In the problem of multimodal word discovery, a learner is given a set of N (image,utterance) pairs. Each utterance has frame-level features $x = (x_1, \cdots, x_T)$, and each image has ROI-level feature vectors $\mathbf{y} = (y_1, \cdots, y_L)$ that depict the visual content in $\mathbf{x}$. The learner tries to find the most likely alignment of speech frames to each image region, denoted as $\mathbf{A} \in [0,1]^{T\times L}$, where $a_{ti} = 1$ if frame t is part of the description of image region i. Here we define *visual concepts* to be the discrete class labels of each ROI, $\mathbf{z} = (z_1, \cdots, z_L)$, chosen from a predefined set of class labels $z_\ell \in \{1, \ldots, K\}$ with a stochastic relationship to the spoken language. Since both $\mathbf{A}$ and $\mathbf{z}$ are latent, one learning strategy is to maximize the likelihood of the observations $(\mathbf{x}, \mathbf{y})$:

$$\max_\theta p(\mathbf{x}, \mathbf{y}|\theta) = \max_\theta \sum_{\mathbf{z}} \sum_{\mathbf{A}\in\{0,1\}^{T\times L}} p(\mathbf{x}, \mathbf{y}, \mathbf{A}, \mathbf{z}|\theta) \quad (1)$$

Alternatively, the learner may instead reformulate the unsupervised learning problem as a binary classification problem, which can be learned more easily using a classical supervised learning approach. Suppose we label the n-th utterance-image pair as $(\mathbf{x}^{(n)}, y^{(n)}), n = 1, \cdots, N$, the model then tries to maximize the likelihood that $\mathbf{x}^{(n)}$ and $\mathbf{y}^{(n)}$ form a pair. The model then maximizes the speech-to-image retrieval probability:

$$\max_\theta p(\mathbf{Y}|\mathbf{X}, \theta) = \max_\theta \prod_{n=1}^{N} \frac{p(x^{(n)}, y^{(n)}|\theta)}{\sum_{m=1}^{N} p(x^{(n)}, y^{(m)}|\theta)}, \quad (2)$$

or the image-to-speech retrieval probability, defined similarly but with a different denominator. A typical approach for speech-to-image retrieval is to assume the joint distribution $p(x^{(n)}, y^{(n)}|\theta)$ to be of the form $p(x, y|\theta) \propto \exp(s(\mathbf{x}, \mathbf{y}))$, where $s(\mathbf{x}, \mathbf{y})$ is an average similarity score, computed as

$$s(\mathbf{x}, \mathbf{y}) = \gamma \sum_{i=1}^{L} \sum_{t=1}^{T} \alpha_{ti}(x_t, y_i)\phi_a(x_t)^\top \phi_v(y_i) + (1-\gamma) \sum_{t=1}^{T} \sum_{i=1}^{L} \beta_{it}(y_i, x_t)\phi_a(x_t)^\top \phi_v(y_i). \quad (3)$$

The functions $\phi_a(\cdot) \in \mathbb{R}^D$ and $\phi_v(\cdot) \in \mathbb{R}^D$ are two neural networks called the *speech encoder* and *visual encoder* respectively used to map the speech and visual features to the joint embedding space, usually constrained to be unit-norm. The weighted average $\mathbf{A} = \gamma\alpha + (1-\gamma)\beta^\top \in [0,1]^{T\times L}$ can be viewed as the soft alignment between the speech frames and image region i, with properties that $\gamma \in [0,1]$, $\sum_{t=1}^{T} \alpha_{ti} = 1$, and $\sum_{i=1}^{L} \beta_{it} = 1$. For example, [4] proposes a MISA (max-image, summed-audio) alignment, in which $\gamma = 0$, $\beta_{it} = 1$ if $i = \text{argmax}_t \, \phi_a(x_t)^\top \phi_v(y_i)$, and $\beta_{it} = 0$ otherwise. Alternatively, α_{ti} and β_{it} may be computed as functions of x_t and y_i, inspired by the attention-weighting of NMT [12, 14, 13]. By maximizing Eq. (3), the model is essentially solving the following optimization problem:

$$\max_{\phi_a, \phi_v} s(\mathbf{x}, \mathbf{y}) = \max_{\substack{\|\phi_a(x_t)\|_2=1\forall t, \\ \|\phi_v(y_i)\|_2=1\forall i}} \text{Tr}\left(\mathbf{\Phi}_a \mathbf{A} \mathbf{\Phi}_v^\top\right) \quad (4)$$

where

$$\mathbf{\Phi}_a := [\phi_a(x_1), \cdots, \phi_a(x_T)] \in \mathbb{R}^{D\times T}, \quad (5)$$

$$\mathbf{\Phi}_v := [\phi_v(y_1), \cdots, \phi_v(y_L)] \in \mathbb{R}^{D\times L} \quad (6)$$

Let the singular value decomposition of $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top, \mathbf{U} \in \mathbb{R}^{T\times K}, \mathbf{V} \in \mathbb{R}^{L\times K}$, we can show that the objective satisfies:

$$\max_{\substack{\|\mathbf{\Phi}_{a,t}\|_2=1\forall t \\ \|\mathbf{\Phi}_{v,i}\|_2=1\forall i}} s(\mathbf{x}, \mathbf{y}) = \max_{\substack{\|\tilde{\mathbf{\Phi}}_{a,k}\|_2=1 \\ \|\tilde{\mathbf{\Phi}}_{v,k}\|_2=1,\forall k}} \text{Tr}\left(\tilde{\mathbf{\Phi}}_a \mathbf{\Sigma} \tilde{\mathbf{\Phi}}_v^\top\right) \quad (7)$$

$$\leq \max_{\substack{\|\tilde{\mathbf{\Phi}}_{a,k}\|_2=1 \\ \|\tilde{\mathbf{\Phi}}_{v,k}\|_2=1,\forall k}} \|\Sigma^{1/2}\tilde{\mathbf{\Phi}}_a^\top\|_2 \|\Sigma^{1/2}\tilde{\mathbf{\Phi}}_v^\top\|_2 \quad (8)$$

by the Cauchy-Schwartz inequality, where $\tilde{\mathbf{\Phi}}_a = \mathbf{\Phi}_a\mathbf{U}$, $\tilde{\mathbf{\Phi}}_v = \mathbf{\Phi}_v\mathbf{V}$, and the maximum is achieved if and only if $\tilde{\mathbf{\Phi}}_a = \tilde{\mathbf{\Phi}}_v$. If A is independent of $\mathbf{x}$ and $\mathbf{y}$, then the optimal Φ_a depends only on the *average* of the embedding vectors and thus does not contain word-level information. A special case of this will be the SISA score function used in [4], where $\mathbf{A} = \frac{1}{TL}\mathbf{1}_{T\times L}$ and the model can simply assign the same embedding to every frame to achieve maximum likelihood: $\bar{\phi_a} = \frac{1}{T}\sum_{t=1}^{T} \phi_a(x_t) = \bar{\phi_v} = \frac{1}{L}\sum_{i=1}^{L} \beta_{it}\phi_v(y_i)$. If the goal of the system is information retrieval, then this is an acceptable outcome: averaging embeddings across the utterance is acceptable if it boosts the match between the image and the audio. If the goal is word discovery, however, this is an unacceptable outcome. Word discovery requires a meaningful alignment matrix $\mathbf{A}$, therefore word discovery requires that some sort of attention mechanism be used to represent the dependence of $\mathbf{A}$ on $\mathbf{x}$ and $\mathbf{y}$.

Instead of mapping the feature vectors to a joint embedding space, an alternative for the joint likelihood function is to use a Bayesian network based on SMT [11], as done in [13]. SMT-style alignment explicitly models the co-occurrence patterns between discrete units discovered in the two modalities, namely, the *image concepts* $z_i \in \{1, \ldots, K\}$ and *word types* $w_t \in \{1, \ldots, W\}$. In this work, we assume that speech frames can be segmented accurately into words with an unsupervised algorithm, such as those in [15], though segmentation can also be inferred multi-modally. Let us further assume a many-to-one, bag-of-words translation probability: the probability of a visual concept given a word type is independent of the sequential order in either modality. Under these assumptions, Eq. (1) can be simplified to the following conditional likelihood:

$$\max_\theta p(\mathbf{y}|\mathbf{x}, \theta) \propto \max_{\Psi_a, \Psi_v, \mathbf{P}, \mathbf{A}} \text{Tr}\left(\mathbf{\Psi}_a^\top \mathbf{P} \mathbf{\Psi}_v \mathbf{A}\right) \quad (9)$$

$$\text{s.t. } \psi_{a,t} \in \Delta_W, \psi_{v,z} \in \Delta_L, \mathbf{P}_w \in \Delta_K, \forall t, i, w \quad (10)$$

where Δ_d is the probability simplex of dimension d, $\mathbf{\Psi}_a \in [0,1]^{W\times T}$ and $\mathbf{\Psi}_v \in [0,1]^{K\times L}$ are two classifiers with probabilities over the latent word types (image ROI) given the acoustic features (visual concepts) respectively, and $\mathbf{P}$ is the translation probability from a word type to an image concept. Notice with $\mathbf{\Psi}_a, \mathbf{\Psi}_v, \mathbf{P}$ fixed, this objective can be rewritten as:

$$\max_{\mathbf{A}_t\in\Delta_L,\forall t} \text{Tr}\left(\mathbf{\Psi}_a^\top \mathbf{P} \mathbf{\Psi}_v \mathbf{A}\right) \leq \sum_{\ell=1}^{L} \|\left(\mathbf{\Psi}_v^\top \mathbf{P}^\top \mathbf{\Psi}_a\right)_\ell\|_\infty, \quad (11)$$

with equality iff $\mathbf{A}_{t\ell} = 1$ for $\ell = \text{argmax}_\ell \left(\mathbf{\Psi}_v \mathbf{P}^\top \mathbf{\Psi}_a^\top\right)_\ell$ and $\mathbf{A}_{t\ell} = 0$ otherwise. Therefore, as long as the latent word/concept classifiers are sufficiently accurate, it can be shown that the SMT is a consistent estimator when learning many-to-one relations between spoken words and image regions. In this work, we trained end-to-end using an exact EM-algorithm described in [13]. We can then plug Eq. (9) into Eq. (2) to find the match likelihood for speech-to-image retrieval. To find the optimal alignment and image concepts, we used a two-stage decoding procedure:

$$A^* = \underset{\mathbf{A}}{\text{argmax}}\, p(\mathbf{A}|\mathbf{x}, \mathbf{y}), \quad z^* = \underset{\mathbf{z}}{\text{argmax}}\, p(\mathbf{z}|\mathbf{A}, \mathbf{x}, \mathbf{y}). \quad (12)$$

	Data	S2I @1	@5	@10	I2S @1	@5	@10
MISA+ TDNN [4]	COCO	12	38	57	12	41	59
Cosine+ TDNN	COCO	**13**	**42**	**60**	**14**	**43**	**61**
Additive+ TDNN	COCO	9	31	48	10	35	53
Normalized+ TDNN	COCO	10	32	48	9	33	48
Cosine+ LSTM	COCO	10	30	45	11	32	45
Cosine+ Transformer	COCO	5	17	26	4	16	24
SMT+ TDNN	COCO	3	13	20	0.1	0.5	1
TDNN (phones)	COCO	32	66	79	32	66	79
	Flickr	17	42	55	18	39	51
SMT (phones)	COCO	7	24	36	4	16	28
	Flickr	7	19	29	3	11	19

Table 1: Speech-to-image (S2I) and image-to-speech (I2S) retrieval performance of various systems: Recall @ 1, 5, 10. Inputs are either phones or audio. Encoders are trained for speech-to-image retrieval using either a TDNN [4], an LSTM, or a Transformer. Alignment is either MISA, SMT, cosine, additive, or normalized.

	Alignment Recall	Alignment Precision	Alignment F1
SMT+TDNN	60	30	40
SMT+Transformer	21.8	43	29
SMT (phones)	37.9	19	25.5
NMT+TDNN	54.9	27.8	36.9
NMT+Transformer	**62.7**	**31.8**	**42.2**

Table 2: Word discovery performance of various systems on MSCOCO. NMT systems use cosine-similarity attention

3. EXPERIMENTS

3.1. Dataset

Two datasets commonly used in cross-modal learning tasks, Flickr8k [16] and MSCOCO [17], are adopted in this paper. Their corresponding spoken captions are given by [18] and [19] respectively. In the original spoken databases [18, 19], each image in both datasets is paired with five spoken captions. To make the training process easy, only one spoken caption is randomly selected from the 5 paired captions for each image. Flickr8k is split according to [20], with 1000 images in the evaluation set. The evaluation set of MSCOCO also consists of 1000 images that were randomly chosen from the MSCOCO 2014 validation set. For both datasets, the reference caption is filtered to remove all but the most frequent 2000 word types, not including stop words.

3.2. Feature extraction

To obtain the image features from different ROI, a Faster-RCNN [21] pre-trained on ImageNet [22] and Visual Genome [23] by [24] is adopted. This pre-trained Faster-RCNN predicts the possible regions and gives the corresponding confidence score for each region. Here, we extract 10 feature vectors from the penultimate layer of the Faster-RCNN of the 10 ROIs with top confidence scores predicted by the Faster-RCNN. Their bounding boxes are also obtained, which allows us map the discovered words to the original images.

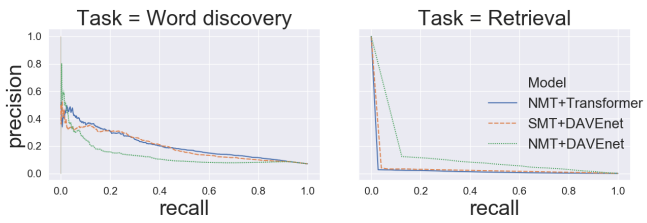

Fig. 2: Alignment and Retrieval precision-recall curves for various models

In the retrieval systems, the speech inputs are represented by mel-frequency filter bank (Fbank) features [25] calculated with 25ms hamming window and 10ms skip. We experiment with three different speech encoders all trained for the speech-to-image retrieval task, namely, the TDNN-based speech encoder from the distributed implementation of the state-of-the-art model DAVEnet [4], a three-layer LSTM encoder [26], and a three-layer, single-head Transformer encoder trained using ESPnet [27]. The embedding dimensions of the models are set to be 1024 except for the LSTM encoder, whose embedding dimension is 1000. The number of parameters for the TDNN, LSTM and Transformer are $15,965,570$, $56,370,001$ and $37,793,792$ respectively.

The speech and image outputs from the retrieval systems (ϕ_a and ϕ_v respectively) are used as inputs to the word discovery systems. Before the alignment step, the SMT system averages embedding vectors of the speech encoders within each spoken segment and compresses them to 300 dimensions using principal component analysis (PCA). $\psi_a(\cdot)$ and $\psi_v(\cdot)$ are set to be softmax distributions with Gaussian kernels with 400 latent word types and 80 latent image concepts respectively for COCO.

3.3. Evaluation metrics

To evaluate the retrieval performance in the retrieval experiments, commonly used evaluation metrics, i.e., Recall@1, 5, 10 scores, are adopted. The *alignment F1* is adopted to evaluate the performance in the word discovery task. An alignment is defined as the association of a spoken segment (in the phone-level case, a sequence of consecutive phones) to an image ROI. The alignment F1 is the harmonic mean of alignment recall and precision. The alignment recall is the percentage of alignments discovered by the system from the set of correct alignments and the alignment precision is the percentage of correct alignments in the set of alignments discovered by the system.

In the evaluation of word discovery performance, we assume ground truth word boundaries to be known to systems with raw audio as inputs and unknown to systems with phone-level transcriptions. For SMT with phone labels, we instead use an unsupervised segmentation system [28] to segment the phone sequence into word-like units. Since the ground truth ROI bounding box is inaccessible during training, direct comparison of the gold alignments with the predicted alignments is impossible. Instead, we align a spoken word with the ground truth bounding box having the highest intersection-over-union (IoU) score with its aligned, predicted box.

(a) audio-level TDNN+NMT (b) audio-level TDNN+SMT (c) phone-level SMT

Fig. 3: Word discovery results of different systems on the image-caption pair "a woman eating a piece of pastry in a market area." The texts are not available in the first two figures during training and are shown for ease of understanding.

For MSCOCO, due to the lack of phrase-level ground truth alignment, we only evaluate on visual words, which are words that describe one of the 80 class names for MSCOCO. To find the visual words, we augment the class names with their plural forms and the related words most commonly mentioned in the captions, such as "man" and "girl" for the `person` class. Phone-level models tend to align each ROI to a long sequence of phones (a phrase) rather than to individual words, which is often the correct behavior, but fails to be detected using a standard IoU criterion. In order to detect correct alignments of phrases to ROI, therefore, we report a correct alignment if the plurality of phones within the phrase are aligned to their correct bounding box.

3.4. Implementation Details

The models are trained with masked margin softmax loss [7], which is a special case of Eq. (2) and observed to work better than triplet loss. Stochastic gradient descent (SGD) is used with a starting learning rate set to be 10^{-6} for the transformer to avoid gradient explosion and 10^{-5} for all other systems. We also train two types of alignment models based on NMT and SMT respectively. The NMT system uses an attention mechanism on top of the speech encoder to learn the soft alignment in Eq. (3). We experiment with three types of attention mechanisms: cosine-similarity (dot product) attention, additive attention and normalized attention [29].

3.5. Results

The retrieval results are shown in Table 2. The cosine-similarity attention mechanism is shown to improve the retrieval performance of TDNN and performs the best among all the speech-level systems. Among the attention mechanisms, we found the cosine-similarity attention not only the most accurate but also the most efficient, as it does not involve additional training parameters. Further, we observe that TDNN performs better than both BiLSTM and Transformer, suggesting that local context is more important than the global context for retrieval. Lastly, the SMT system does not perform as well as the NMT, as it does not use any contextual information beyond word level, which may be beneficial to learn better word-level representation. SMT performance is particularly bad for image-to-speech (I2S) retrieval, suggesting that long-term audio context may be more important for I2S than S2I.

The word discovery results are shown in Table 2. For the SMT+TDNN and NMT+TDNN model, we use the embedding vectors from the cosine-similarity attention system as it performs the best during retrieval. To better understand the relation between word discovery and retrieval, we also plot the precision-recall curves for the two tasks side by side in Fig. 2 and provide an example of bounding box alignments for NMT+TDNN, SMT+TDNN and the phone-level SMT in Fig. 3. Quantitatively, the NMT+Transformer outperforms the SMT+TDNN and the NMT+TDNN systems by 2% and 5% respectively, suggesting that a better alignment mechanism is indeed beneficial for word-level representation learning. However, combining SMT with self attention is not effective for word discovery as it degrades the word discovery performance compared to SMT+TDNN. Also, we found that while the phone-level model tends to perform better than audio-level in retrieval, the noise in segmentation makes it perform worse in word discovery than the audio-level model. From Fig. 2, while NMT+TDNN outperforms the other two systems in retrieval, it underperforms them in word discovery, showing a discrepancy between learning to retrieve images and learning to discover word-like units. The gap in performance is also explained qualitatively by Fig. 3. We can see that the NMT+TDNN system is able to align the store name to the word "market," but aligns all the other regions to the word "piece", showing a lack of word-level knowledge. The SMT-based discovery system is able to correctly identify objects such as the woman in front annotated by the red bounding box and the pastry near her mouth. Further, the SMT model is able to align the person inside the orange box to the word "woman," even though she is not mentioned in the caption, due to the bag-of-word nature of the SMT model. The SMT model also aligns the objects sold in the shop to the word "pastry," apparently because they are visually similar to pastry. Similarly to the audio-level SMT, the phone-level SMT model is able to align the semantically correlated regions to segments of the corresponding words for both "market" and "woman," which suggests that segmentation error is the main cause of its low performance.

4. CONCLUSION

We have studied theoretically and empirically the tradeoff between speech-to-image retrieval and word discovery. We demonstrate that a speech embedding learned using a TDNN gives the highest speech-to-image retrieval scores, but that embedding learned using a self-attention Transformer model gives higher scores for word discovery. In both cases, accuracy is boosted by using an NMT-based attention mechanism with self-attention layers, which helps the retrieval model to learn better alignments for visual words. From our results, we believe a joint retrieval-discovery is important for developing bet-

ter word discovery systems.

5. REFERENCES

[1] D. Roy, "A computational model of word learning from multimodal sensory input," in *Proceedings of the international conference of cognitive modeling*, Groningen, The Netherlands, 2000, pp. 1–8.

[2] S. E. Levinson, Q. Liu, C. Dodsworth, R.-S. Lin, W. Zhu, and M. Kleffner, "The role of sensorimotor function, associative memory and reinforcement learning in automatic acquisition of spoken language by an autonomous robot," in *Joint NSF DARPA Workshop on Development and Learning*, East Lansing, MI, 2000.

[3] D. Harwath, A. Torralba, and J. Glass, "Unsupervised learning of spoken language with visual context," in *Neural Information Processing Systems*, 2016.

[4] D. Harwath, A. Recasens, D. Suris, G. Chuang, A. Torralba, and J. Glass. (2018) Jointly discovering visual objects and spoken words from raw sensory input. [Online]. Available: https://arxiv.org/pdf/1804.01452.pdf

[5] W. Havard, J. Chevrot, and L. Besacier, "Models of visually grounded speech signal pay attention to nouns: a bilingual experiment on english and japanese," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[6] D. Merkx, S. Frank, and M. Ernestus, "Language learning using speech to image retrieval," in *Interspeech*, 2019.

[7] G. Ilharco, Y. Zhang, and J. Baldridge, "Large-scale representation learning from visually grounded untranscribed speech," in *The SIGNLL Conference on Computational Natural Language Learning (CoNLL)*, 2019.

[8] H. Kamper, G. Shakhnarovich, and K. Livescu, "Semantic speech retrieval with a visually grounded model of untranscribed speech," *IEEE Transaction on Audio, Speech and Language Processing*, vol. 27, pp. 89–98, 2019.

[9] D. Harwath, W.-N. Hsu, and J. Glass, "Learning hierarchical discrete linguistic units from visually-grounded speech," in *International Conference on Learning Representation*, 2020.

[10] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," in *31st Conference on Neural Information Processing Systems (NIPS 2017)*, 2017.

[11] P. F. Brown, V. J. D. Pietra, S. A. D. Pietra, and R. L. Mercer, "The mathematics of statistical machine translation: parameter estimation," *Computational Linguistics*, vol. 19, no. 2, pp. 263 – 311, 1993.

[12] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *ICLR*, 2015.

[13] L. Wang and M. Hasegawa-Johnson, "A DNN-HMM-DNN hybrid model for discovering word-like units from spoken captions and image regions," in *Interspeech*, 2020.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Neural Information Processing Systems (NIPS)*, 2017.

[15] O. J. Raŝañen, G. Doyle, and M. C. Frank, "Unsupervised word discovery from speech using automatic segmentation into syllable-like units," in *Interspeech*, 2015.

[16] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, "Collecting image annotations using Amazon's mechanical turk," in *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 2010.

[17] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollr, "Microsoft COCO: Common objects in context," in *European Conference on Computer Vision*, 2014.

[18] D. Harwath and J. Glass, "Deep multimodal semantic embeddings for speech and images," *Automatic Speech Recognition and Understanding*, 2015.

[19] W. Havard, L. Besacier, and O. Rosec, "Speech-COCO: 600k visually grounded spoken captions aligned to MSCOCO data set," in *GLU 2017 International Workshop on Grounding Language Understanding*, 2017.

[20] A. Karpathy, A. Joulin, and L. Fei-Fei, "Deep fragment embeddings for bidirectional image sentence mapping," in *Neural Information Processing Systems*, 2014.

[21] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, 2015.

[22] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[23] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International journal of computer vision*, vol. 123, no. 1, pp. 32–73, 2017.

[24] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6077–6086.

[25] S. B. Davis and P. Mermelstein, "Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[26] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, pp. 1735–1780, 1997.

[27] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "ESPnet: End-to-end speech processing toolkit," in *Proceedings of Interspeech*, 2018, pp. 2207–2211. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1456

[28] S. Goldwater, T. L. Griffiths, and M. Johnson, "Contextual dependencies in unsupervised word segmentation," in *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, 2006, p. 673680.

[29] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, "Attngan: Fine-grained text to image generation with attentional generative adversarial networks," in *CVPR*, 2018.