Align or attend? Toward more efficient and accurate spoken word discovery using speech-to-image retrieval

— ECE 590 SIP presentation

Liming Wang

Nov.16th, 2020

Multimodal Word Discovery (MWD)

Figure: From babyblue.com

Multimodal Word Discovery (MWD)

"Which describes which?"

- Given:
  - Spoken caption: $\mathbf{x} = x_1, \cdots, x_T$
  - Image regions: $\mathbf{y} = y_1, \cdots, y_L$
- Find: which spoken frames describes which visual region

Maximum likelihood estimation (MLE)

$$\max_\theta p(\mathbf{x}, \mathbf{y}|\theta) = \max_\theta \sum_{\mathbf{A}} p(\mathbf{x}, \mathbf{y}, \mathbf{A}|\theta)$$

$$\mathbf{A}^* = \text{argmax}_{\mathbf{A}}\, p(\mathbf{A}|\mathbf{x}, \mathbf{y}, \theta),$$

where $A_{ti} = 1$ if word t and region i are aligned.

Two-step approach: MWD via speech-to-image retrieval

Step 1

Sentence-level matching (speech-to-image retrieval Harwath and Glass (2015)):

$$\max_{\theta} p(\mathbf{M}|\mathbf{x}^{(1:N)}, \mathbf{y}^{(1:N)}, \theta) = \prod_{n,m:M_{nm}=1} \frac{p(\mathbf{x}^{(n)}, \mathbf{y}^{(m)}|\theta)}{\sum_{n'} p(\mathbf{x}^{(n)}, \mathbf{y}^{(n')}|\theta)},$$

where $M_{nm} = 1$ if caption n and image m are matched.

Step 2

Word-level matching (MWD):

$$\mathbf{A}^* = \text{argmax}_{\mathbf{A}}\, p(\mathbf{A}|\mathbf{x}, \mathbf{y}, \theta).$$

DAVEnet: State-of-the-art MWD system

Origin

First proposed by Harwath et al. (2018)

Assumptions

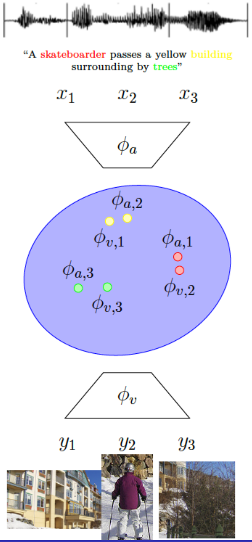
1. Dominant (soft) alignment assumption:

$$p^{\text{DAVEnet}}(\mathbf{x}, \mathbf{y}|\theta) := \exp\left(\sum_{t,i} A_{ti} s(x_t, y_i)\right)$$

2. Common space assumption:

$$s(x_t, y_i) = \frac{\phi_a(x_t)^\top \phi_v(y_i)}{\|\phi_a(x_t)\|\|\phi_v(y_i)\|},$$

where $\phi_a(\cdot), \phi_v(\cdot)$ are learned by two **DNNs**



Does DAVEnet always learn good word-level representation?

Analysis: MLE of DAVEnet

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

where maximum is achieved if, for $\mathtt{svd}(\mathbf{A}) = \mathbf{U},\mathbf{\Sigma},\mathbf{V}$:

$$\mathbf{\Phi}_a\mathbf{U} = \mathbf{\Phi}_v\mathbf{V}$$

Good sentence embedding $\neq$ good word embedding

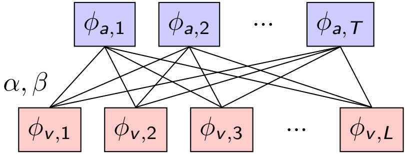
:

$$\begin{aligned}\mathbf{A}\text{ independent of }\phi_v,\phi_a &\implies \phi_v^*,\phi_a^*\text{ independent of }x_t\text{ and }y_i\\ &\implies \text{bad word-level representation}\end{aligned}$$

Fix the common space: Attention mechanisms

Intuition

- Need to make **A** variable of ϕ_a, ϕ_v
- May still fail to learn good word embedding with variable **A** (e.g., constant $\phi_v(y_i)$)



Cosine attention

$$\alpha_{ti} = \frac{\exp\left(s(x_t, y_i)\right)}{\sum_t \exp\left(s(x_t, y_i)\right)}$$

$$\beta_{ti} = \frac{\exp\left(s(x_t, y_i)\right)}{\sum_i \exp\left(s(x_t, y_i)\right)}$$

Additive attention

$$\alpha_{ti} = \frac{\exp\left(\mathbf{W}_i[\phi_{a,t}; \phi_{v,i}; 1]\right)}{\sum_t \exp\left(\mathbf{W}[\phi_{a,t'}; \phi_{v,i}; 1]\right)}$$

$$\beta_{ti} = \frac{\exp\left(\mathbf{W}_i[\phi_{a,t}; \phi_{v,i}; 1]\right)}{\sum_{i'} \exp\left(\mathbf{W}_{i'}[\phi_{a,t}; \phi_{v,i}; 1]\right)}$$

Self attention

$$\alpha_{tt'}^{(m)} = \frac{\exp\left(\Phi_a^{(m)\top}\Phi_a^{(m)}\right)_{tt'}}{\sum_{t''} \exp\left(\Phi_a^{(m)\top}\Phi_a^{(m)}\right)_{tt''}}$$

Fix the common space: Change the space

DNN-HMM-DNN model by Wang and Hasegawa-Johnson (2020)

- Additional hidden variables:
  - $\mathbf{z} = [\mathbf{z}_1, \cdots, \mathbf{z}_L]$: image concept of each image region
  - $\phi = [\phi_1, \cdots, \phi_T]$: acoustic unit label of each speech segment
- Conditional likelihood:

$$p(\mathbf{y}|\mathbf{x}, \theta) = \sum_{\mathbf{z}, \mathbf{A}, \phi} p(\mathbf{z}|\mathbf{y}) p(\mathbf{A}, \phi, \mathbf{x}|\mathbf{z}, L)$$

- Learn to recognize concepts and phones with two **DNNs** ψ_a and ψ_v
- Learn to align concepts and phones with an **HMM**

DNN-HMM-DNN as learning a common probabilistic space

Rewrite the conditional likelihood using matrix operations:

$$\max_{\mathbf{A}} p(\mathbf{y}|\mathbf{x}, \mathbf{A}, \theta) = \max_{\mathbf{A}_t \in \Delta_L, \forall t} \mathrm{Tr}\left(\mathbf{\Psi}_a^\top \mathbf{P} \mathbf{\Psi}_v \mathbf{A}\right),$$

Guarantee

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.

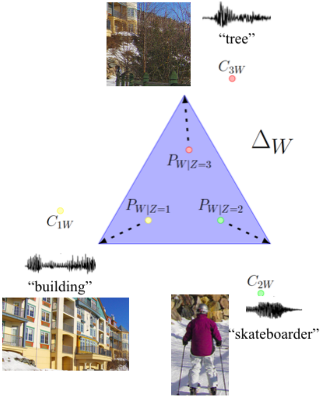


Figure: $C_{zW} := (\Psi_v^\top \mathbf{A} \Psi_a)_z$

Dataset

- **Flickr8k** (Hodosh et al. (2010)): Split according to Karpathy et al. (2014), 30000 image-caption pairs for training, 1000 images for evaluation
- **SpeechCOCO** (Havard et al. (2017)): 80 image concept classes, 80000 image-caption pairs for training, 1000 images randomly chosen from the MSCOCO 2014 validation set for evaluation
- **Preprocessing**: Filter the most frequent 2000 word types, not including stop words

Features

- **Speech features**:
    - **Retrieval**: Mel filter-bank features with 25ms window and 10ms skip step
    - **MWD**: last layer of the speech encoder averaged over each word segment, compressed to 300-d vectors with PCA
- **Image features**: 2048-d ResNet50 features for the top 10 ROIs proposed by the Faster-RCNN pretrained on Visual Genome and ImageNet

Implementation details

NMT

- **TDNN-based systems**: 5 convolution layers, 1024-d embedding, default settings of the DAVEnet implementation by (Harwath et al. (2018))
- **BiLSTM-based system**: 3 convolution layers, 1000-d embedding
- **Transformer-based system**: 3 self-attention layers, 1024-d embedding, implementation from ESPnet (Watanabe et al. (2018))
- **Loss function**: masked margin softmax loss (Ilharco et al. (2019))

SMT

- Softmax distributions with Gaussian kernels for encoders, 400 latent word types, 80 latent image concepts for SpeechCOCO; 600 latent image concepts for Flickr

Results: Speech-to-image retrieval

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

Results: MWD

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

Table: Word discovery performance of various systems on MSCOCO; Results are evaluated only with words that describe one of the 80 concepts

Tradeoff between retrieval and word discovery

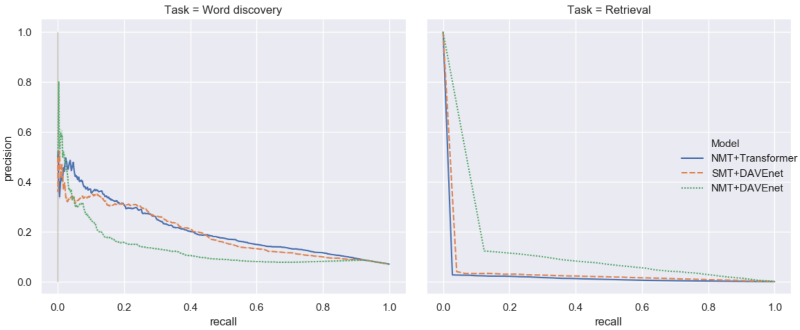
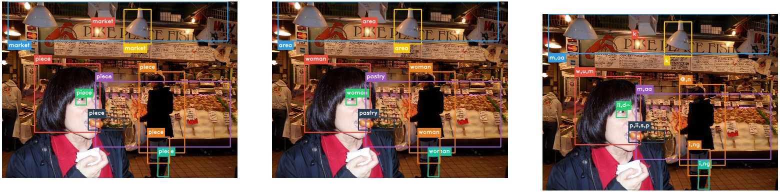


Figure: Alignment and Retrieval precision-recall curves for various models

Example



(a) audio-level DAVEnet+NMT

(b) audio-level DAVEnet+SMT

(c) phone-level SMT

Figure: 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.

Discussion

- Averaging vs. peak detection: the right approach for extracting word embedding from DAVEnet?
- Common space clustering vs. probabilistic alignment/clustering
- Discriminative training vs Maximum likelihood training

New result

Discriminative training of SMT

$$\max \mathrm{Tr}(\boldsymbol{\Psi}_a^\top \mathbf{P} \boldsymbol{\Psi}_v \mathbf{A})$$
$$\text{s.t.} \quad \mathrm{Tr}(\bar{\boldsymbol{\Psi}}^\top \mathbf{P} \boldsymbol{\Psi}_v) = 1,$$
$$\mathbf{P}_w \in \mathbb{R}^{K+}, \forall w$$

where $\bar{\boldsymbol{\Psi}} := \sum_{n=1}^N \mathbf{A} \boldsymbol{\Psi}_a^{(n)\top}$.

Solution

$$P_w^* = \frac{(\boldsymbol{\Psi}_v^\top \mathbf{A} \boldsymbol{\Psi}_\mathbf{a})_{z^* w}}{(\bar{\boldsymbol{\Psi}}^\top \boldsymbol{\Psi}_a)_{z^* z^*}} \mathbf{e}_{z^*},$$

where $z^* = \arg\max_z \frac{(\boldsymbol{\Psi}_a^\top \mathbf{A} \boldsymbol{\Psi}_\mathbf{v})_{z^* w}}{(\bar{\boldsymbol{\Psi}}^\top \boldsymbol{\Psi}_\mathbf{a})_{z^* z^*}} \approx \arg\max_z \frac{\bar{p}(z^*|w)}{\bar{p}(z^*)}$, where $\bar{p}(\cdot)$ is the empirical distribution.

Conclusion

- 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 better word discovery systems.

Future Direction

David Harwath, Galen Chuang, and James Glass. 2018. Vision as an interlingua: Learning multilingual semantic embeddings of untranscribed speech. In *IEEE International Conference on Acoustics, Speech and Signal Processing*.

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

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

M. Hodosh, P. Young, and J. Hockenmaier. 2010. Framing image description as a ranking task: data, models and evaluation metrics. In *Journal of Artificial Intelligence Research*.

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

Andrej Karpathy, Armand Joulin, and Li Fei-Fei. 2014. Deep fragment embeddings for bidirectional image sentence mapping. In *Neural Information Processing Systems*.

Liming Wang and Mark Hasegawa-Johnson. 2020. Multimodal word discovery and retrieval with spoken descriptions and image concepts. *IEEE Transaction of Audio, Speech and Language Processing*.

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