Study of Unsupervised Learning of Visual Objects with Spoken Words based on Siamese Network

Hanjuan Zhang

Sun Yat-sen University, China

Keywords: speech and image; convolutional networks; multimodal learning; deep learning.

Abstract: Humans learn to understand spoken languages and recognize images by mostly unsupervised learning. It should also be possible for machines to jointly learn spoken languages and visual object. By abandon the fully connected layer and trying different effective networks like vgg19 and resnet, the existing neural network processing directly on the image pixels and audio waveform can be improved. My models also do not rely on any labels or any common supervision. The best model trained on parallel speech and images achieves a precision of over 70% on its top ten retrievals and almost 50% on its top five retrievals. Flickr8k dataset and audios dataset converted from Flickr8k text are used to train the models.

1. Introduction

There has recently been a huge interest in jointly learning the speech and visual object, especially concerning speech waveform and image models that can learn from unlabelled speech paired with visual object. Some current works focus on bring together vision and audio using large amounts of transcribed speech data. This is not similar to how we learn things as children. However, we can always hear ambient sounds while focusing on finding the sounds that have real meaning. This make prompted the development of the models that can handle weaker or noisy forms of sounds and video[1-5]. In my work, the goal is to find relevant images given a spoken description without any intermediate supervisions. But the problem could be much harder to solve for the speech audio is much shorter and the number of categories is much larger.

Most approaches map the image and the speech into a same feature space, using dot product[2], cross-entropy loss[6] and so on. Retrieval-based model in [2] has a precision of over 30% on its top ten retrievals. [7] using the same idea in [2], but using recurrent highway network as speech encoder. Networks in [8] using a combination of Resnet and a multilayer GRU to encode the audio caption. Both shows that encodings of form and meaning emerge and evolve in hidden layers of stacked RNNs processing grounded speech. Although effective, such models have failed to represent the connection between features of the image and the speech. My model is based on the model in [9], which discard the fully connected layer and only retain the convolutional layers up through conv5 from VGG16 network. This model can provide good accuracy for image classification task. Also, each location within the map possessing a receptive field that can be related directly back to the input, which enables the recovery of spatial activation maps for a given target class, and can be used for object localization. My primary aim here is to find out a more effective model to train the paired image and speech waveform. Using a newly generated dataset and some models that performed well in image recognition, my work present an analysis of an updated version of the model of [9], and compare it to alternative models for the semantic retrieval. What makes my model unique is that my train models process directly on the image pixels and audio waveform, instead of some speech transcriptions, and without any labels.

2. Image to Speech Retrieval

The multi-modal encode the images and their corresponding speech waveforms to a common embedding space. The goal is to make the value of the distance between the paired images and

speech waveforms low and the distance value between mis-paired images and speech waveforms high. My model consists of two part, including an image model and an audio model.

2.1 Speech Retrieval Dataset

Instead of using the existing dataset like Places Audio Caption dataset [2][9][10], the Flickr8k image corpus are used as input for training image model and Flickr8k text corpus are used to generate the input of audio model. The Flickr8k dataset [11] consisting of approximately 8,000 images that are each paired with five or more different captions which provide clear descriptions of the salient entities and events. The captions are writing by annotators to describe the scenes, situations, event and wntities. The images in the dataset were chosen from different Flickr dataset and contains a variety of scenes and situations. The terminal commands "say" in the macOS system is used, that can convert txt files to m4a files, and iTunes, that can convert m4a files to wav files, to transform Flickr8k text corpus into corresponding speech waveforms. Then the dataset is augmented by pairing every image and audio, resulting in a grand total of over 30,000 paired images and audios for training. Due to the lack of computing resources, only additional 100 image/caption pairs are used for validation.

2.2 Image Modeling

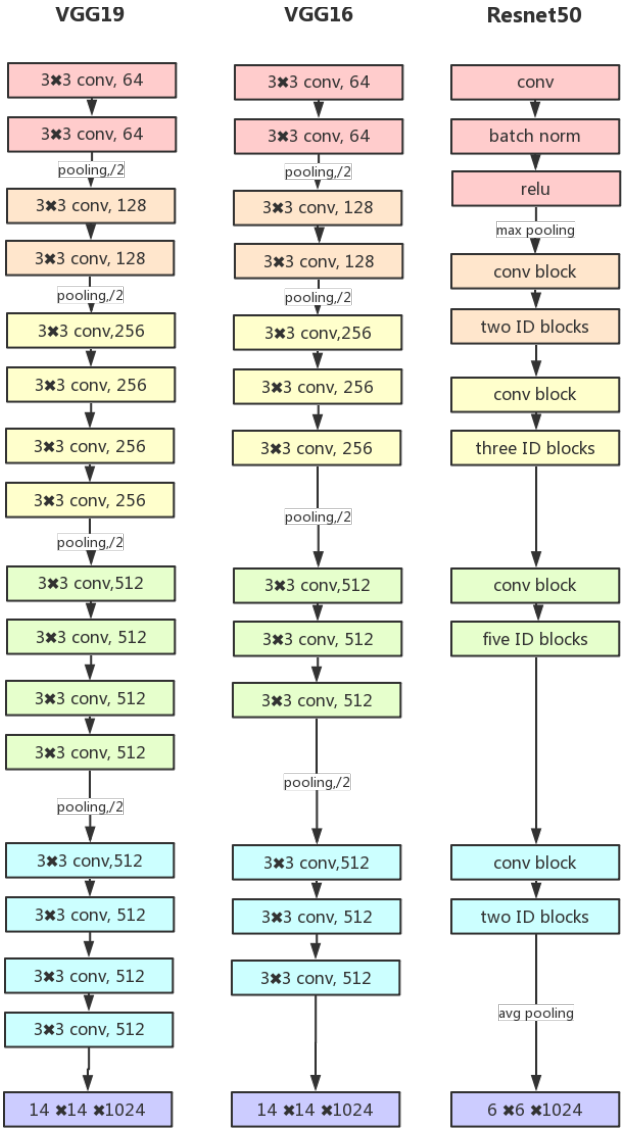

Figure 1. Example network architectures for image model. Left: the modified VGG19 model. Middle: the modified VGG16 model. Right: a resnet50 without fully connected layer.

The preparation of image data should be converting RGB to BGR and resizing every image to a 224 by 224, 3 channels image caption. Several plain/residual networks have been tested, including VGG16, VGG19 and Resnet50 in Fig.1. To provide instances for discussion, I describe three model for image models as follows.

VGG16 Network. Based on [2][6][9][10], The structure of traditional VGG16 network[12] is modified to perform an image caption. Unlike the previous work in [9], the pre-training of models is necessary and can significantly reduce the total amount of training epoches to reach the convergence. In order to make object localization and find a better connection between image and audio waveforms, I abandon the process of the fully connected layer and only retain the layers up through last convolutional(conv5) layer. With the input image of 224 by 224 pixels and a total of fourteen convolutional layers, the output of the image is a 14 by 14 feature map which location within the map can directly link back to the correspondent input. In order to map the image output with the audio output, a 1024 channel linear convolution is applied to conv5, resulting in a 14 by 14, 1024 channel feature map.

VGG19 Network. VGG19 network (Fig.1,left) is just an upgrade of VGG16 network. VGG19 have three extended layers based on VGG16. One is a 3 by 3 filters and 256 channels in conv3. One is a 3 by 3 filters and 512 channels in conv4. The third is a 3 by 3 filters and 512 channels in conv5. The network is modified by retaining only the layers up through conv5, and discarding pool5 and everything above it. The output of conv5 is a 14 by 14, 512 channels feature map. By adding an additional convolutional layer with 3 by 3 filters and 1024 channel, there will be a 14 by 14, 1024 channels feature map to match with the output from audio network.

Resnet50. With the better performance in VGG19, the tendency is to test more complicated networks. The details of conv block and identity block has shown in Fig.2. Resnet50[13] has shortcut path which can be applied as a model simplifier and provides the benefit of simple models in a complex network. If shortcut path is dominant, the layers between this shortcut are essentially ignored, reducing the complexity of the model in effect. By abandon the layers above stage 5, there will be an output of a 7 by 7, 1024 channels feature map.

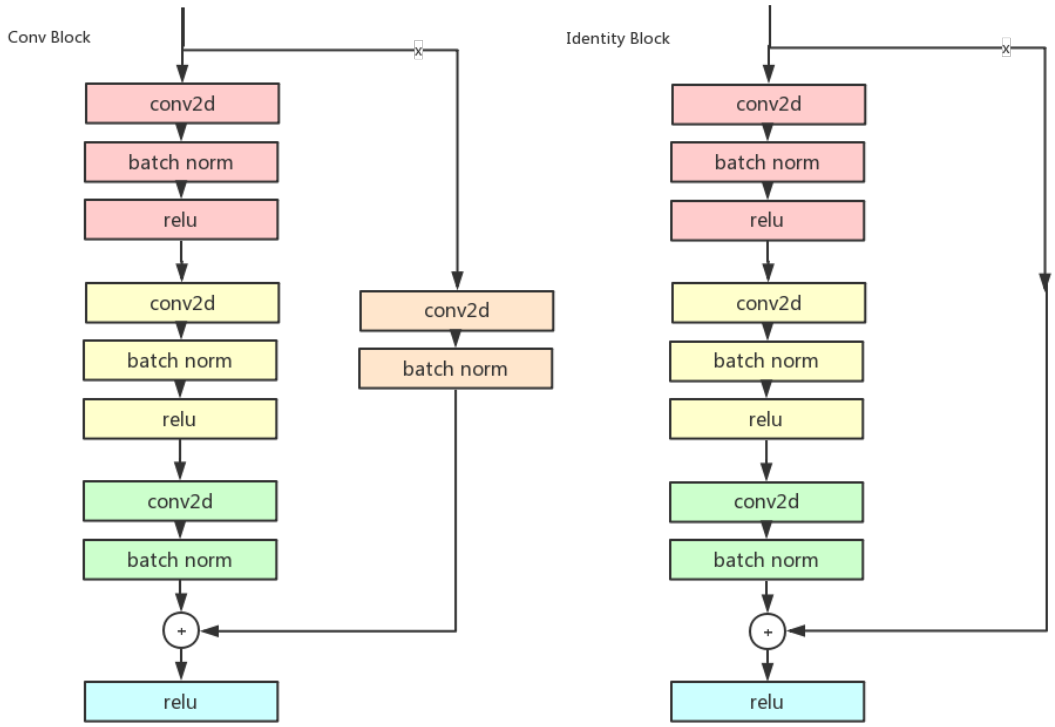

Figure 2. Left: Details of conv block with shortcut. Right: Details of identity block with shortcut.

2.3 Audio Modeling

For encoding the speech waveforms, some models tested are similar to that of [9], with different total number of layers. And some models similar to that of [8], containing multi-layer LSTM and multi-layer GRU. Mel Frequency Cepstral Coefficients (MFCCs) is used to get the acoustic feature. 10 seconds (1024 frames) of audio is required. The number of filters in the filterbank is forty. The FFT size is 551 in order to fit the dataset.

2.4 Joint Modeling

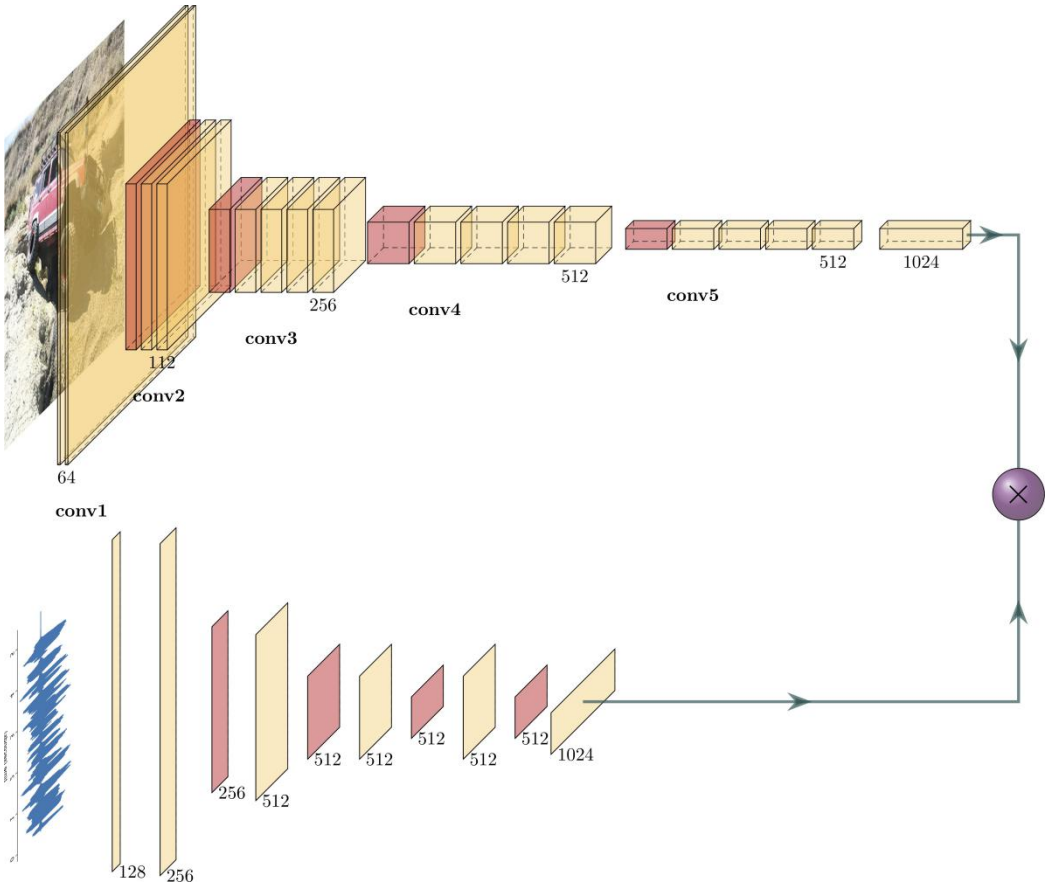

Figure 3. The structure of the multi-modal structure is a combination of a VGG19 network and a plain network and is subsequently tied together at the output level with a customized dot product node which calculates a similarity score for any given image and audio pair.

The multi-modal structure is shown in Fig.3. Let $I_{i,j,k}$ represent the output feature map of the image network with $_{i,j,k}$ stands for the three different dimension of the output image feature map respectively. Let $A_{t,k}$ represent the output feature map of the audio network with $_{t,k}$ stands for the two different dimension of the output audio map respectively. Let $M_{i,j,t}$ be the result of multiplication of two tensors with different dimensions, $M_{i,j,t} = IT \times A_{t,:}$. Let N_t be the total number of $_t$. Since S represent the similarity between a small region of the image and a small part of the audio, I would like to think that matching a small segment of the audio with all regions in the image to find the best fit can help me find out that the small segment of audio is describing which part of the image. And it could show if every segment of the audio is like a part of image and them using the average to find out if the whole audio itself is like the whole image. That is why I apply the following as the customized similarity function to calculate the distance between audio and image:

$$S = \frac{1}{N_t} \sum_{t=1}^{N_t} \max(M_{i,j,t}) \tag{1}$$

2.5 Training

In order to fulfill my goal, the model is trained to map the images and the speeches such that the similarity between the paired image and the speech is lower than the similarity between the mismatched image and the speech. This so-called hinge loss L as a function of the network parameters θ is given by:

$$L(\theta) = \sum_{i=1}^{N_b} \max(0, S_c - S_p + 1) + \max(0, S_j - S_p + 1) \quad (2)$$

Where N_b stands for the size of the minibatch. S_c stands for the similarity between the original image and the mismatched speech. S_p stands for the similarity between the truth pair. S_j stands for the similarity between the original speech and the mismatched image.

This loss function encourages a higher similarity to a matching image and speech pair than a mismatched pair. This loss function considers every mismatched pair within the minibatch. In practice, stochastic gradient descent is used with a batch size of 4, a constant decay rate of 0.5 for the 1st moment, a constant decay rate of 0.9 for the 2st moment, and an initial learning rate of 0.0001. Learning rates took a bit of tuning to get right. In the end, the models converged in less than 150 epochs, with the support of the pre-trained models. 100 pairs of image/speech are used for validation, similar to those in [9]. This method provides a single, high-level metric which captures how well the model has learned to semantically bridge the audio and visual modalities.

3. Result

The result of precision rate is shown in Table1. Precision at one (P@1) is the average precision of all highest-scoring retrieval which means the one with the highest similarity score. Precision at five (P@5) is the average precision of the top five highest-scoring retrieval. And precision at ten (P@10) is for top ten.

Table 1. Recall scores on the set of 100 images/speech for the various models. It also shows results for the baseline models which do not have the training with (R). All the models are used with pre-trained on the image model. _5 means that the audio network is with five layers and _6 means that the audio network is with six layers. _LSTM means that the audio network is LSTM network.

Model	P@1	P@5	P@10
VGG16(R)	0.05	0.14	0.30
VGG16_5	0.07	0.47	0.76
VGG16_6	0.12	0.48	0.74
VGG19(R)	0.05	0.14	0.30
VGG19_5	0.16	0.45	0.65
VGG19_6	0.16	0.46	0.70
Resnet50(R)	0.05	0.12	0.34
Resnet50_5	0.01	0.12	0.26
Resnet50_6	0.01	0.13	0.35
VGG19_LSTM(R)	0.05	0.14	0.30
VGG19_LSTM	0.01	0.10	0.23
VGG19_GRU	-	-	-

Image and speech retrieval rate for various multi-modal network and all the baseline are shown in Table1. VGG19 with six-layer audio network performs best among all models.

VGG19 network outperforms the VGG16 network at P@1, while Resnet50 perform worse of all. In terms of absolute performance, like when doing a classification task, the network should be complex but not too complex to converge, such as VGG19. However, in terms of P@5 and P@10,

VGG16 performs best. P@10 only measures precision of the highest ranked utterances, thus, VGG16 network can be good at recall more semantic matches.

Clearly, six-layer audio network have the best performance among the other audio networks. As for the LSTM network, the network is so complicated to converge, and it is impossible to find a proper pre-trained model for it.

All the models listed, are required for pre-trained models. Without pre-trained models, it is impossible for the Siamese network to converge. While testing, the loss of models without pre-train could get stuck at 2.000 forever.

4. Conclusion

Investigating effective models to find relevant images given a spoken description without any intermediate supervisions is the goal. The audio dataset is generated by machine voice to better retrieve image and audio. With raw audio input and without any labels, the Siamese network model achieves a semantic P@1 of approximately 16%, P@5 of almost 50% and P@10 of over 70%. Although a complex model could perform worse on some samples, a complex model can just not suitable for this kind of situation or with wrong training methods.

References

[1] Chrupała, G.; Gelderloos, L.; Alishahi, A. Representations of language in a model of visually grounded speech signal 2017.

[2] Harwath, D. Unsupervised Learning of Spoken Language with Visual Context. 2016.

[3] Kamper, H.; Settle, S.; Shakhnarovich, G.; Livescu, K. Visually Grounded Learning of Keyword Prediction from Untranscribed Speech. 2017, pp. 3677–3681. doi:10.21437/Interspeech.2017-502.

[4] Synnaeve, G.; Versteegh, M.; Dupoux, E. Learning words from images and speech. In NIPS Workshop on Learning Semantics, 2014.

[5] Aytar, Y.; Vondrick, C.; Torralba, A. SoundNet: Learning Sound Representations from Unlabeled Video 2016.

[6] Kamper, H.; Shakhnarovich, G.; Livescu, K. Semantic Speech Retrieval with a Visually Grounded Model of Untranscribed Speech. IEEE/ACM Transactions on Audio, Speech, and Language Processing 2018, PP, 1–1. doi:10.1109/TASLP.2018.2872106.

[7] Chrupała, G.; Gelderloos, L.; Alishahi, A. Representations of language in a model of visually grounded speech signal. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Association for Computational Linguistics: Vancouver, Canada, 2017; pp. 613–622. doi:10.18653/v1/P17-1057.

[8] Merkx, D.; Frank, S.L.; Ernestus, M. Language learning using Speech to Image retrieval 2019.

[9] Harwath, D.; Recasens, A.; Surís, D.; Chuang, G.; Torralba, A.; Glass, J. Jointly Discovering Visual Objects and Spoken Words from Raw Sensory Input. International Journal of Computer Vision 2018. doi:10.1007/s11263-019-01205-0.

[10] Harwath, D.; Glass, J. Learning Word-Like Units from Joint Audio-Visual Analysis 2017.

[11] Hodosh, M.; Young, P.; Hockenmaier, J. Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. Journal of Artificial Intelligence Research 2013, 47, 853–899. doi: 10. 1613/jair.3994.

[12] Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv 1409.1556 2014.

[13] He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. 2016, pp. 770–778. doi:10.1109/CVPR.2016.90.