Semantic Retrieval of Personal Photos using a Deep Autoencoder Fusing Visual Features with Speech Annotations Represented as Word/Paragraph Vectors

Hung-tsung Lu[1], *Yuan-ming Liou*[2], *Hung-yi Lee*[2], *and Lin-shan Lee*[12]

[1]Graduate Institute of Computer Science and Information Engineering, National Taiwan University
[2]Graduate Institute of Communication Engineering, National Taiwan University

r03922011@ntu.edu.tw, r02942070@ntu.edu.tw

Abstract

It is very attractive for the user to retrieve photos from a huge collection using high-level personal queries (e.g. "uncle Bill's house"), but technically very challenging. Previous works proposed a set of approaches toward the goal assuming only 30% of the photos are annotated by sparse spoken descriptions when the photos are taken. In this paper, to promote the interaction between different types of features, we use the continuous space word representations to train a paragraph vector model for the speech annotation, and then fuse the paragraph vector with the visual features produced by deep Convolutional Neural Network (CNN) using a Deep AutoEncoder (DAE). The retrieval framework therefore combines the word vectors and paragraph vectors of the speech annotations, the CNN-based visual features, and the DAE-based fused visual/speech features in a three-stage process including a two-layer random walk. The retrieval performance was significantly improved in the preliminary experiments.

Index Terms: image retrieval, speech annotation, word representation, paragraph vector, convolutional neural network, deep autoencoder, random walk, fused features

1. Introduction

With the popularity of digital cameras and smart phones, many people saved huge collections of personal photos, but found it challenging to browse across the collection to find a desired photo. Users usually prefer to use personal words as queries to look for photos (e.g. who, where, when, what (objects/events), such as "uncle Bill's house" or "wedding ceremony"). This makes the very successful content-based image retrieval [1, 2] less useful here, because it requires an example photo as the query. The huge number of annotated photos over the Internet can be useful in identifying photos of publicly known objects (such as "White House") [3, 4], but not necessarily for the personal photo descriptions considered here. Manual annotation of each individual photo is certainly useful, but not attractive at all. This led to the idea of annotating some photos with speech [5, 6], and this task seems to be simply the spoken document retrieval [7, 8, 9].

A major issue in spoken document retrieval is that the query and its relevant documents may use different set of words. Latent topics or factor analysis can handle this issue to some extent, with probabilistic latent semantic analysis (PLSA) and non-negative matrix factorization (NMF) as two typical examples [10, 11]. But PLSA and NMF may not be able to solve the problem here, because the query and the labels for related photos may be in several different categories (e.g. some photos by where and some by who, while the query by event) or use different sets of words, and the latent relationships among different terms, specially in different categories, very possibly cannot be trained with very sparse personal annotations. This led to the concept of using image features jointly with speech annotations [12]. Related photos may be linked by image features if annotated very differently, or even not annotated at all.

In recent works, we proposed to fuse local image features (e.g. visual words by clustering low level image features [13, 14]) and global image concepts (e.g. "people" or "outdoor" by Columbia374 detector [15]) with the sparse, free-form, and spontaneously spoken annotations [16]. We further enhanced the sparse voice annotations by finding semantically/syntactically related words with continuous space word representations, and modeled the relationships among photos and their labels with NMF [17]. In addition, we reinforced the retrieval process considering the different types of features with two-layer mutually reinforced random walk [18, 19].

We propose a completely new framework for the task in this paper. We estimate paragraph vectors for the speech annotation, extract more precise image semantic features using deep convolutional neural network (CNN), and fuse the speech and visual features together with a deep autoencoder (DAE). These different types of features are properly used in a three-stage retrieval process. Significant improvements in performance were observed in preliminary experiments.

2. Proposed Approach

2.1. Overview of the Proposed Approach

As shown in Figure 1, in the preparation phase on the left, we first have a deep convolutional neural network (CNN) model trained with an image database as at the upper left corner. This CNN model is able to project each photo on a 1000-dimensional semantic space and construct a 1000-dim posterior probability vector to be taken as the visual features of the photos. On the lower left of Figure 1, we also train a continuous space word representation model and a paragraph vector model with a large text corpus, based on which the word vectors and paragraph vectors for the speech annotations are also obtained as speech features. In the middle of Figure 1, we train a deep autoencoder (DAE) model to fuse the visual and speech features, which can transform any input photo into the visual/speech fused semantic representation no matter it is annotated or not.

In the retrieval phase on the right of Figure 1, given a query in text form, in stage 1 we first compare the query with all word arcs in the lattices of the speech annotations based on the similarity in continuous space word representations to select some initial photos. In stage 2, the results obtained in stage 1 are extended by selecting more photos based on the visual CNN

features and visual/speech fused DAE features. Two-layer random walk is finally performed in stage 3 to re-rank the retrieved photos to give the final photo list.

2.2. Visual Features: Convolutional Neural Network

Deep convolutional neural networks (CNN) have been found very useful in many image-based complex tasks. In this work, we use the CNN model trained with the ImageNet database [20, 21] to extract the visual features for each photo. ImageNet is an image database organized according to the WordNet [22] hierarchy (currently including only the nouns), where each meaningful concept in WordNet is possibly described by multiple words or phrases, and each node of the hierarchy is depicted by hundreds or thousands of images.

We simply feed the personal photos into the well-trained CNN model available in the Caffe website [23, 24] to project each of them onto the 1000-dim pre-defined semantic space after softmax function to extract the visual features of the photos. The 1000-dim visual features before softmax function in CNN are also used to train the deep autoencoder fusing visual/speech features.

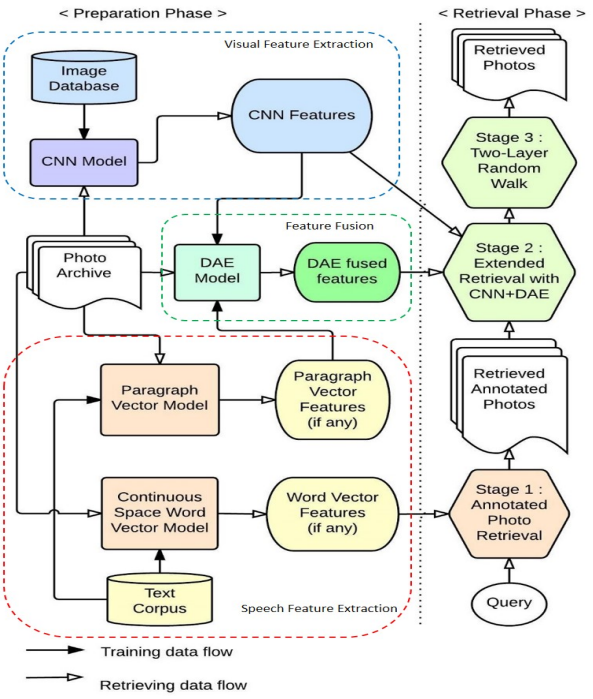

Figure 1: *The proposed approach*

2.3. Speech Features: Word Vector and Paragraph Vector

The speech annotation is the key information here because it provides the most important personal semantic concepts such as "uncle Bill's house" or "Mary's wedding ceremony". But these speech annotations can be very spontaneous under varying acoustic conditions and may include out-of-vocabulary (OOV) words. The one-best recognition accuracy can be low. We therefore represent each utterance by a lattice, and find the continuous space word vector for every word item on the arcs of the lattices.

2.3.1. Continuous Space Word Representation

Many different models were developed for representing words as vectors in continuous space [25, 26, 27, 28, 29]. Recurrent neural network language model (RNNLM) as in Figure 2(a) used the hidden layer at the previous time, $h(t-1)$, with a recurrent structure to take into account the previous context [28]. Continuous bag-of-words model (CBOW) as in Figure 2(b) learned to predict the present word $w(t)$ based on the preceding and following words such as $w(t-2), w(t-1), w(t+1), w(t+2)$ via a projection layer without non-linear elements. Continuous Skip-gram model as in Figure 2(c) is very similar to CBOW, but with the layers reversed. The word representation can be obtained from the transformation for the projection layer of CBOW or Skip-gram model [29]. It was shown that CBOW and continuous Skip-gram models are better than RNNLM for both syntactic and semantic tasks.

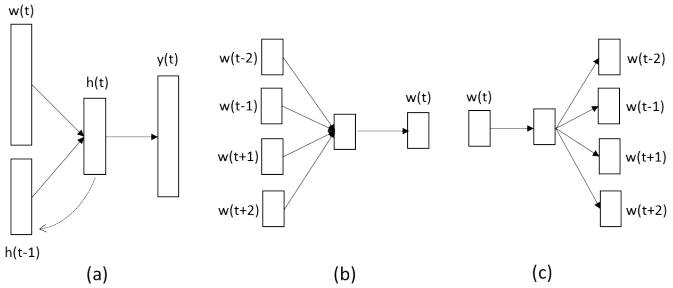

Figure 2: *Neural networks for modeling word representations: (a) RNNLM (b) CBOW (c) Skip-gram*

2.3.2. Paragraph Vector

The paragraph vector was proposed recently to represent each paragraph (or document) by a vector which is trained with the word vectors for the words in the paragraph when predicting the following words in the paragraph [30]. Such paragraph vectors may have the potential to overcome the weaknesses of models based on bag-of-words by considering the word ordering, representing the missing information from the current context, and somehow characterizing the topics of the paragraph. Empirical results showed that paragraph vectors outperformed models based on bag-of-words as well as other techniques for text representations, and achieved new state-of-the-art results on several text classification and sentiment analysis tasks [30].

In the paragraph vector learning framework as shown in Figure 3, every paragraph with a single paragraph id is mapped to a unique vector (such as $\boldsymbol{v}_0$), represented by a column in the paragraph representation matrix $\boldsymbol{D}$, and every word (such as word 1, word 2, word 3) is also mapped to a unique vector (such as $\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3$), represented by a column in the word representation matrix $\boldsymbol{M}$. The paragraph vector and word vectors are averaged or concatenated to predict the next word (such as word 4) in a context of fixed-length sampled by a sliding window over the paragraph. The paragraph vector is shared across all contexts generated from the same paragraph but not across paragraphs. The word vector matrix $\boldsymbol{M}$ is shared across all paragraphs, so the vectors for the same word are the same for all paragraphs. In this work, we represent each speech annotation utterance as a lattice, and pick up the top N possible paths on it to form a paragraph which shares the same paragraph id.

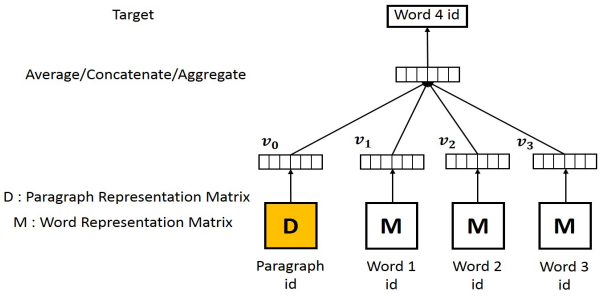

Figure 3: *The framework for learning paragraph vectors.*

2.4. Fused Features: Deep Autoencoder (DAE)

Because only 30% of the photos are annotated by spoken descriptions while all photos have CNN visual features, we first train a Deep AutoEncoder (DAE) [31, 32] using visual features for several epochs, and then use the paragraph vectors for the speech annotations to enhance the DAE model. In this way, we properly fuse the visual and speech features for the photos. As shown in Figure 4, the DAE is trained first with the visual features only (denoted as $\boldsymbol{x}$, and normalized to [0, 1]) as the training input at the bottom right, while the training target on the top right being the same $\boldsymbol{x}$. This gives the right path shown with solid lines, with the output $\hat{\boldsymbol{x}}$ on the top right close to $\boldsymbol{x}$. In the feed-forward stage as in (1) (2) (3) below, we can calculate the distortion function $E(\hat{x}, x)$ for each visual feature set $\boldsymbol{x}$ which can be a mean square error or some other error function in the output layer. The errors will then be back-propagated and the parameters updated by stochastic gradient decent (SGD) algorithm.

$$\boldsymbol{o}^{(0)} = \boldsymbol{x} \quad , \tag{1}$$

$$\boldsymbol{o}^{(l)} = f(\boldsymbol{W}^{(l)}\boldsymbol{o}^{(l-1)} + \boldsymbol{b}^{l}) \quad , \quad 1 \leq l \leq N-1 \quad , \tag{2}$$

$$\hat{\boldsymbol{x}} = \boldsymbol{o}^{(N)} = g(\boldsymbol{W}^{(N)}\boldsymbol{o}^{(N-1)} + \boldsymbol{b}^{N}) \quad , \tag{3}$$

where $\boldsymbol{x}$ is the visual feature vector, $\boldsymbol{o}^{l}$ the values on the l-th layer, $f(.)$ and $g(.)$ the sigmoid function, and $\hat{\boldsymbol{x}}$ the corresponding output.

After finishing training the right part including all $\boldsymbol{W}$'s and $\boldsymbol{b}$'s in Figure 4, we then add the left part shown with dotted lines including $\boldsymbol{V}$'s and $\boldsymbol{c}$'s in the second stage of training using those photos with speech annotations. The input speech features $\boldsymbol{z}$ are the paragraph vectors for the speech annotations which give extra dimensions in both input and output layers, denoted as $\boldsymbol{z}$ and $\hat{\boldsymbol{z}}$ respectively (the training target is also $\boldsymbol{z}$, so the output $\hat{\boldsymbol{z}}$ is close to $\boldsymbol{z}$). After randomly initializing the extra dimensions of the weight matrices and bias vectors $\boldsymbol{V}$'s and $\boldsymbol{c}$'s, we can continue the training process in exactly the same way as in (1) (2) (3) and keep updating all parameters including $\boldsymbol{W}$'s, $\boldsymbol{b}$'s, $\boldsymbol{V}$'s and $\boldsymbol{c}$'s, except now the input is $\boldsymbol{y} = (\boldsymbol{z}, \boldsymbol{x})$ and output is $\hat{\boldsymbol{y}} = (\hat{\boldsymbol{z}}, \hat{\boldsymbol{x}})$.

With this DAE trained, we can now feed each photo (with speech annotation $\boldsymbol{z}$ or not) into the DAE model and extract its fused visual/speech features by picking up the feature vectors on the bottle-neck layer of the DAE.

2.5. Photo Retrieval Phase

With the CNN visual features, the continuous word vector and the paragraph vector for speech features, and the DAE fused features, we are ready to enter the photo retrieval phase. There are three stages here are given below.

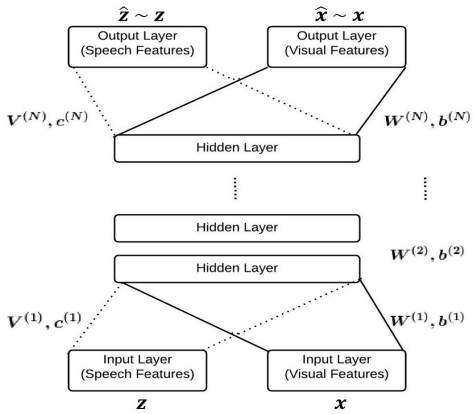

Figure 4: *Deep autoencoder model*

2.5.1. Stage 1: annotated photo retrieval by word vectors

Because speech annotations may carry explicitly or implicitly the most important key semantics about the personal photos, we represent the speech annotation for a photo as a bag-of-words, which includes all word items for all arcs in the lattice with high enough confidence scores, each represented as a continuous space word vector. Given a user query which is a word also represented as a word vector, for each annotated photo, we take the top k words having the highest cosine similarity with the query, and the average of these k cosine similarity values is taken as the relevance score for the annotated photo. In this stage, we pick up the top m photos with the highest scores as the result. Of course only those photos with speech annotations can be retrieved at this stage.

2.5.2. Stage 2: Extended retrieval with DAE fused features

All photos in the archive are fed into the DAE to extract the visual/speech fused features (for those photos without annotation, the input is $\boldsymbol{y} = (\boldsymbol{0}, \boldsymbol{x})$). We then concatenate such a DAE feature vector with the 1000-dim CNN visual vector. In this second stage, for each of the m first-pass retrieved photos, n more photos in the archive can be found based on the cosine similarity between the DAE-CNN concatenated vectors. The scores of these n photos can be set as the scores of first-pass retrieved photo they are similar to weighted by the cosine similarity values here.

2.5.3. Stage 3: Two-layer Random Walk Enhancement

The photo scores from stage 2 can be further enhanced by the two-layer random walk [18, 19] as in Figure 5. Each node in the lower layer represents a photo collected in stage 2, while that in the upper layer represents a photo selected in the stage 1 which has speech annotation. Let $S_U^{(0)}$, $S_L^{(0)}$ represent respectively the vectors for the initial scores for nodes in upper and lower layers evaluated at stage 1 and 2, and $S_U^{(t)}$, $S_L^{(t)}$ the enhanced version of them at the t-th iteration. The score propagation can be expressed as random walk in (4) below,

$$\begin{cases} S_U^{(t)} = & (1-\alpha)S_U^{(0)} + \alpha \cdot E_{UU}^T E_{UL} S_L^{(t-1)} & (4-1) \\ S_L^{(t)} = & (1-\alpha)S_L^{(0)} + \alpha \cdot E_{LL}^T E_{LU} S_U^{(t-1)} & (4-2) \end{cases} \tag{4}$$

where E_{UU} (based on expected term frequency on the lattice), E_{UL} (based on paragraph vectors), and E_{LL} (based on CNN features) are respectively the upper-to-upper, upper-to-lower, and lower-to-lower row-normalized cosine similarity matrices, etc. For example, in $(4-2)$ the scores of upper layer $S_U^{(t-1)}$ are weighted first by the lower-to-upper similarity E_{LU} then by the lower-to-lower similarity E_{LL}, and then contribute to the scores of the lower layer $S_L^{(t)}$. In this way the scores for the nodes are propagated and smoothed, such that nodes similar to more high-score nodes will have higher scores and so on, because the score of a node is distributed to other nodes based on the similarity.

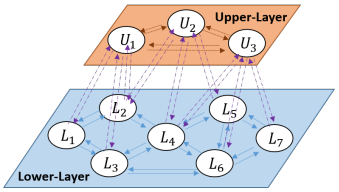

Figure 5: *A simplified example of the two-layer random walk.*

3. Experiments

3.1. Experiment Setup

The photo archive was taken from a Flickr user who has more than ten thousand photos on the web with diversified topics. We randomly selected 7777 from them to be used here. Several students generated the annotation text (primarily in Chinese) spontaneously, most indicating one or two categories of information (e.g. where or who) about the photos explicitly or implicitly, many including OOV words. The audio for these annotations were recorded by 57 students without constraints on the microphone or the acoustic conditions.

The speaker independent (SI) acoustic models were adapted by 30 utterances for each speaker to generate the speaker adapted (SA) models. A language model interpolated from two models respectively trained by news corpora and Plurk corpora was used. The recognition accuracy for the very free speech annotations was only 40.3% for words. Only 30% of the photos (2100) randomly selected out of the 7777 were allowed to have speech annotations, while the other 70% were assumed to have image features only. Another five students generated 32 queries (4 where, 4 who, 4 event and 20 object) and labeled their ground truths for evaluation. Each query is a Chinese word composed of 2 or 3 syllables. For word representation, we used a corpus of 760 million words collected from a popular Chinese based BBS forum in Taiwan covering many topics, including photo annotations, to train the RNNLM, CBOW and Skip-gram models for a 436k-word lexicon, producing 300-dimensional word vectors. The paragraph vector model was then trained based on these word vectors. The CNN model used was the one available on Caffe website [23, 24] with a subset of ImageNet [20] database of about 1.2 million training images and the 1000 predefined semantic targets were based on ImageNet Large Scale Visual Recognition Challenge 2012 [21].

3.2. Experimental Results

The results in terms of MAP@50 (mean average precision [33] evaluated for the top 50 retrieved objects) are listed in Table 1, in which Section (A) is the baseline obtained in the previous works [16, 17]. Row (a) is for NMF only, in which visual features and speech features (including word vector representations) were embedded in the matrix and fused by NMF, while two-layer random walk was applied in addition in row (b). We can see the performance was not satisfactory with NMF only, while two-layer random walk was very helpful in enhancing the output.

Section (B) is for the proposed approach. In row (c) for stages 1 and 2 but not including the two-layer random walk of stage 3, we see the various advanced features carrying more semantic information plus the framework of stages 1 and 2 offered much better performance (rows (c) vs (a) both without random walk). When the two-layer random walk in stage 3 was applied in addition in row (d), we see significant improvement was achieved with the random walk (rows (d) vs (c)), and the features and framework proposed here are very helpful even with the random walk enhancement (rows (d) vs (b) both with random walk). These results implied the CNN visual features provided very good global visual concepts, the paragraph vectors offered very good global spoken concepts, and the DAE features properly fused the two. All these information are jointly used in stage 2, and further mutually reinforced in the random walk of stage 3.

The results in rows (c)(d) are for a good selection of the parameters m (number of photos selected in stage 1 for each query) and n (number of extra photos selected in stage 2 for each query). The results for a few other combinations of m and n are listed in rows (e)(f)(g), all producing a total of $m(n+1) = 200$ photos to be used in the lower layer of the random walk. We see $m = 50$ and $n = 3$ in row (d) offered the best performance, while the performance degraded with smaller m. Obviously all photo ranking and selection processes in stage 2 and 3 were extended from the m photos selected in stage 1, and therefore better choice of m is critical.

Table 1: Experimental results: (A) Baselines with NMF and with 2-Layer random walk in addition, (B) Proposed approach for different choices of parameters.

	Approaches	Details	MAP@50
(A)	Baselines	(a) NMF only (with word vectors)	15.72%
		(b) NMF + 2-Layer RW	25.12%
(B)	Stage 1 + 2 only	(c) M = 50, n = 3	29.82%
	Stage 1 + 2 + 3	(d) M = 50, n = 3	**32.05%**
		(e) M = 40, n = 4	31.29%
		(f) M = 20, n = 9	27.61%
		(g) M = 10, n = 19	24.89%

4. Conclusions

This paper proposed a new framework for semantic retrieval of personal photos with sparse voice annotations and high-level personal queries. We propose several advanced features for this task including continuous space word vectors and paragraph vectors for speech features, CNN for visual features, and DAE for fused visual/speech features, to be jointly used in a 3-stage retrieval framework. In the preliminary experiments, the proposed approaches achieved significant performance improvement compared to those obtained in previous works.

5. References

[1] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, and B. Dom, "Query by image and video content: the QBIC system," IEEE Computer, Sep. 1995.

[2] Smith, John R., and Shih-Fu Chang. "VisualSEEk: a fully automated content-based image query system." Proceedings of the fourth ACM international conference on Multimedia. ACM, 1997.

[3] Naphade, Milind, et al. "Large-scale concept ontology for multimedia." MultiMedia, IEEE 13.3 (2006): 86-91.

[4] Yi-Hsuan Yang, Po-Tun Wu, Ching-Wei Lee, Kuan-Hung Lin, Winston H. Hsu, "ContextSeer: Context Search and Recommendation at Query Time for Shared Consumer Photos," ACM Multimedia 2008(full paper), Vancouver, Canada.

[5] J. Chen, T. Tan, P. Mulhem, and M. Kankanhalli, "An improved method for image retrieval using speech annotation," Proceedings of the 9th International Conference on Multi-Media Modeling 2003.

[6] Timothy J. Hazen, Brennan Sherry and Mark Adler, "Speech-based annotation and retrieval of digital photographs," Interspeech 2007.

[7] C. Chelba, J. Silva and A. Acero," Soft indexing of speech content for speech in spoken documents," Computer Speech and Language, vol. 21, no. 3, pp.458-478, July 2007.

[8] Yi-chen Pan, Hung-lin Chang and Lin-shan Lee, "Analytical comparison between position specic posterior lattices and confusion networks based on words and subword units for spoken document indexing," Automatic Speech Recognition & Understanding, pp.677-682 , Dec 2007.

[9] Ya-chao Hsieh, Yu-tsun Huang, Chien-chih Wang and Lin-shan Lee, "Improved spoken document retrieval with dynamic key term lexicon and probabilistic latent semantic analysis(PLSA)," ICASSP 2006, vol. 1, May 2006.

[10] T. Hofmann, "Probabilistic latent semantic indexing," Proc. ACM SIGIR Conf. R&D in Informational Retrieval, 1999.

[11] Lee, Daniel D., and H. Sebastian Seung. "Learning the parts of objects by non-negative matrix factorization." Nature 401.6755 (1999): 788-791.

[12] Fu, Yi-sheng, Chia-yu Wan, and Lin-shan Lee. "Latent semantic retrieval of personal photos with sparse user annotation by fused image/speech/text features." Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on. IEEE, 2009.

[13] Tirilly, Pierre, Vincent Claveau, and Patrick Gros. "Language modeling for bag-of-visual words image categorization." Proceedings of the 2008 international conference on Content-based image and video retrieval. ACM, 2008.

[14] Yang, Jun, et al. "Evaluating bag-of-visual-words representations in scene classification." Proceedings of the international workshop on Workshop on multimedia information retrieval. ACM, 2007.

[15] Yanagawa, Akira, et al. "Columbia universitys baseline detectors for 374 lscom semantic visual concepts." Columbia University ADVENT technical report (2007): 222-2006.

[16] Liou, Yuan-ming, Yi-sheng Fu, Hung-yi Lee, and Lin-shan Lee. "Semantic Retrieval of Personal Photos using Matrix Factorization and Two-layer Random Walk Fusing Sparse Speech Annotation with Visual Features," Interspeech 2014.

[17] Liou, Yuan-ming, Hung-tsung Lu, Yi-sheng Fu, Winston Hsu, and Lin-shan Lee. "Enhancing Sparse Voice Annotation for Semantic Retrieval of Personal Photos by continuous space word representations," ICASSP 2015.

[18] Cai, Xiaoyan, and Wenjie Li. "Mutually reinforced manifold-ranking based relevance propagation model for query-focused multi-document summarization." Audio, Speech, and Language Processing, IEEE Transactions on 20.5 (2012): 1597-1607.

[19] Chen, Yun-Nung, and Florian Metze. "Two-layer mutually reinforced random walk for improved multi-party meeting summarization." Spoken Language Technology Workshop (SLT), 2012 IEEE. IEEE, 2012.

[20] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database. IEEE Computer Vision and Pattern Recognition (CVPR), 2009.

[21] Olga Russakovsky*, Jia Deng*, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg and Li Fei-Fei. (* = equal contribution) ImageNet Large Scale Visual Recognition Challenge. arXiv:1409.0575, 2014.

[22] Miller, George A. "WordNet: a lexical database for English." Communications of the ACM 38.11 (1995): 39-41.

[23] Jia, Yangqing, et al. "Caffe: Convolutional architecture for fast feature embedding." Proceedings of the ACM International Conference on Multimedia. ACM, 2014.

[24] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. 2012.

[25] Hinton, Geoffrey E. "Distributed representations." (1984).

[26] Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. "Learning representations by back-propagating errors." Cognitive modeling (1988).

[27] Elman, Jeffrey L. "Finding structure in time." Cognitive science 14.2 (1990): 179-211.

[28] Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig. "Linguistic Regularities in Continuous Space Word Representations." HLT-NAACL. 2013.

[29] Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013).

[30] Le, Quoc V., and Tomas Mikolov. "Distributed representations of sentences and documents." arXiv preprint arXiv:1405.4053 (2014).

[31] Hinton, Geoffrey E., and Ruslan R. Salakhutdinov. "Reducing the dimensionality of data with neural networks." Science 313.5786 (2006): 504-507.

[32] Ngiam, Jiquan, et al. "Multimodal deep learning." Proceedings of the 28th International Conference on Machine Learning (ICML-11). 2011.

[33] Baeza-Yates, Ricardo, and Berthier Ribeiro-Neto. Modern information retrieval. Vol. 463. New York: ACM press, 1999.