Probabilistic Ensemble Fusion for Multimodal Word Sense Disambiguation

Yang Peng*, Daisy Zhe Wang*, Ishan Patwa*, Dihong Gong* and Chunsheng Victor Fang†
*University of Florida, Gainesville, FL
Email: {ypeng, daisyw, ipatwa, dihong}@cise.ufl.edu
†Awake Networks, Mountain View, CA
Email: vicfcs@gmail.com

***Abstract*—With the advent of abundant multimedia data on the Internet, there have been research efforts on multimodal machine learning to utilize data from different modalities. Current approaches mostly focus on developing models to fuse low-level features from multiple modalities and learn unified representation from different modalities. But most related work failed to justify why we should use multimodal data and multimodal fusion, and few of them leveraged the complementary relation among different modalities.**

In this paper, we first identify the correlative and complementary relations among multiple modalities. Then we propose a probabilistic ensemble fusion model to capture the complementary relation between two modalities (images and text). Experimental results on the UIUC-ISD dataset show our ensemble approach outperforms approaches using only single modality. Word sense disambiguation (WSD) is the use case we studied to demonstrate the effectiveness of our probabilistic ensemble fusion model.

***Keywords*-Multimodal Fusion, Probabilistic Ensemble Fusion, Word Sense Disambiguation**

I. Introduction

Words in natural languages tend to have multiple senses, for example, the word "crane" may refer to a type of bird or a type of machine. The problem of determining which sense of a word is used in a sentence is called word sense disambiguation (WSD). WSD was first formulated as a distinct computational task during the early days of machine translation in the 1940s, making it one of the oldest problems in computational linguistics. Different kinds of methods have been introduced to solve WSD, including knowledge-based approaches, statistical models and other machine learning techniques [1], [2]. While most of the existing approaches exploit only textual information, very limited research efforts have been conducted on multimodal data for word sense disambiguation [25], [26]. In this paper, WSD is used as a classification task to demonstrate the effectiveness of our probabilistic ensemble fusion model.

With abundant multimedia data on the Internet from 2000s, researchers are starting to develop multimodal machine learning models to integrate data of multiple modalities, including text, images, audios and videos. Current approaches mostly focus on developing a unified representation model for multiple modalities and then employ existing classification methods on the unified representation [17], [23], [24], [25], [26]. Previous work [17] conducted correlative matrix analysis on textual and visual features, which motivated us to discover the *correlative* relation among multiple modalities. In addition to correlative relation, we also discoverd the *complementary* relation among multiple modalities. These two properties of multimodal data explain why multimodal machine learning could achieve higher accuracy than single-modality approaches, as discussed in Section 3.

In this paper, we propose a probabilistic ensemble fusion model to combine the results of text-only classification and image-only classification to achieve better accuracy than text-only and image-only classification. Our ensemble fusion model is designed to capture the complementary relation between images and text. Different ensemble approaches, including the linear rule, the maximum rule, and logistic regression, were used to combine the results from classification methods using single-modality data.

Our main contributions focus on:

- First, we identified the correlative and complementary relations among different modalities and discovered that by utilizing these relations we can achieve better accuracy than approaches only using single-modality data.
- Second, we designed a probabilistic ensemble fusion model with different ensemble approaches to achieve better accuracy than image-only and text-only classification methods. The probabilistic ensemble fusion model is designed to capture the complementary relation among different modalities, which has been ignored by most previous work.

Overview Section 2 talks about the related work for word sense disambiguation and multimodal machine learning. Section 3 discusses the correlative and complementary relations among different modalities. Our probabilistic ensemble fusion model is introduced in Section 4. Experimental results, which can demonstrate the effectiveness of our model, are explained in Section 5. The conclusion and next steps of our project are shown in Section 6.

This work was done when Chunsheng Victor Fang was affiliated with Pivotal Software Inc..

978-1-5090-0379-2/15 $31.00 © 2015 IEEE
DOI 10.1109/ISM.2015.35

II. Related Work

In this section, we will briefly discuss the research work related to word sense disambiguation and multimodal machine learning. One of the unsupervised WSD algorithms, Yarowsky algorithm [16] is exploited as a text classification approach in our ensemble fusion model. A few research papers on multimodal machine learning are also reviewed. Most of the previous work ignored the complementary relation among multiple modalities, which motivated us to propose the ensemble fusion model to capture the complementary relationship.

A. Word Sense Disambiguation

Word sense disambiguation (WSD) first appeared in the 1940s as a subtask of machine translation. Over the years, there have been numerous research efforts to address this challenging problem. The most prominent approaches can be categorized as supervised, unsupervised, and knowledge-based approaches [1], [2].

For supervised approaches, the earliest efforts were decision list [3] in 1980s and decision tree [4] in 1990s. Many supervised statistical algorithms have been employed for WSD, including Naive Bayesian, Neural Networks, and Support Vector Machines [1], [2]. However, it is unrealistic to manually label a very large collection of textual data, which is the major limitation of the supervised approaches.

Unsupervised approaches [1], [2], [5], [6], [7], [8], on the other hand, do not require a large labeled dataset, which enables them to overcome the knowledge acquisition bottleneck, i.e. the lack of large data collections with manual annotations. Words carrying the same sense usually have similar contexts, i.e. nearby words in a small neighborhood. Thus the contexts can be used to induce senses for a specific word through clustering with appropriate similarity measure. But unsupervised approaches have a major disadvantage that they do not exploit any knowledge inventory or dictionary of real-world senses. One of the famous unsupervised algorithms, Yarowsky algorithm [16] is used inside our probabilistic ensemble fusion model and will be explained in detail in Section 4.

Knowledge-based methods, which utilize knowledge resources (e.g. dictionaries, ontologies, etc), provide a better trade-off between disambiguation accuracy and computational costs than supervised and unsupervised methods. There are several different kinds of knowledge-based techniques, including the overlap of sense definitions [9], [10], selectional restrictions [11], [12], [13], and structural approaches (semantic similarity measures and graph-based methods) [14], [15]. Although the disambiguation accuracy of knowledge-based methods is not as good as their supervised alternatives, they do not require large amount of manually labeled data for supervised classifier training. With the help of rich structured information from knowledge bases, these approaches usually have a much larger coverage than the unsupervised WSD methods.

B. Multimodal Machine Learning

With abundant multimedia data, including text, images, audios and videos, appearing on the Internet from 2000s, researchers are starting to working on multimodal data fusion and cross-modal machine learning to solve real-world problems, such as classification and information retrieval.

Rasiwasia et.al. [17] proposed several state-of-the-art approaches to achieve cross-modal information retrieval. The first approach was correlation matching, which aimed to map the different feature spaces for images and text to the same feature space based on correlation analysis of these two spaces. The second approach was semantic matching, which represented images and text with the same semantic concepts using multi-class logistic regression. More previous work on multimodal information retrieval can be found in [18], [19], [20], [21], [22]. With deep learning taking over the machine learning community in recent years, there have been efforts in exploiting deep learning for multimodal learning [23], [24]. In [23], J. Ngiam et.al. proposed the bimodal deep Boltzmann machine and the bimodal deep autoencoder for multimodal fusion and cross-modal learning.

For word sense disambiguation, there have been several research projects on using images and text to improve disambiguation accuracy [25], [26]. In [25], May et.al. combined the image space and text space directly and applied a modified version of Yarowsky algorithm [16] on the combined space to solve WSD. But this naive combination of two spaces ignored the correlation between the image space and text space, which might lead to poor accuracy. In [26], Saenko et.al assumed the features of one modality are independent of sense given the other modality, then used LDA to model the probability distributions of senses given images and text separately, and combined these two distributions using a sum rule. Although the linear rule in our model and the sum rule in [26] may look similar, the ideologies and motivations behind them are quite different. The goal of the sum rule in [26] was to model the joint probability distribution of senses given both images and text under the independence assumption, while our goal of the linear rule approach is to capture the complementary relationship between images and text in the ensemble fusion framework, where text classification and image classification are conducted first and then the linear rule is used to combine the results of them to achieve higher accuracy.

The previous research work, as discussed above, mostly focused on developing unified representation models from text and images based on correlation between images and text, and then use classification techniques on top of unified representation models to solve different tasks. On the other hand, our probabilistic ensemble fusion model is designed

to capture the complementary relation between text and images.

III. Why Multimodal Fusion Works

In this section, the correlative and complementary relations among multiple modalities are explained in detail. In order to simplify the scenario, only two modalities, images and text, are discussed in this section. We also explain how our probabilistic ensemble fusion approach can capture the complementary relation to achieve higher accuracy than single-modality approaches.

A. Correlative Relation

The correlative relation between text and images means images and textual sentences of the same documents tend to contain semantic information describing the same objects or concepts. For example, the image and textual sentence in Figure 1(a) both refer to the sense "bass (fish)", while the image and sentence in Figure 1(b) both describe the sense "bass (instrument)".

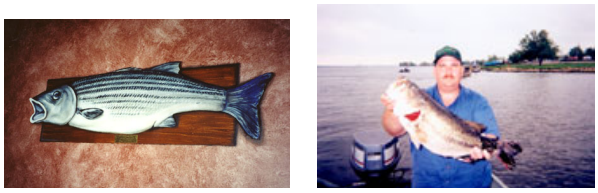
(a) "fish of florida: rock sea bass" (b) "l.a. kidwell musical instruments - product (bass 006)"

Figure 1: Examples selected from UIUC-ISD dataset [29] for keyword "bass". The left figure shows a document carrying sense "bass (fish)" and the right figure shows another document carrying sense "bass (instrument)".

Because images and text for the same documents have this semantic correlative relation, they tend to be correlated in the feature spaces as well. Then it is possible to conduct correlation analysis on textual features and visual features to construct a unified feature space to represent multimodal documents. Previous research papers [17], [23], [24] exploit the correlative relation to develop a unified representation model for multimodal documents, although most of them did not identify this relation explicitly. Since there exists a "semantic gap" [19] between semantic concepts and image features, the performance of these approaches utilizing the correlative relation, is highly dependent on the image features, textual features and the correlative analysis methods, as well as the nature of the data (e.g. whether the correlative relation exists in majority of the documents inside the dataset).

B. Complementary Relation

Images and text are complementary to each other by containing different semantic information. In our word sense disambiguation case, textual sentences contain more useful and relevant information for classification in some documents, while images contain more useful information in other documents. For example, in Figure 2(a), the sentence "portfolio 2" contains little information to disambiguate senses for "bass", while the image depicts the "bass fish" object. In Figure 2(b), the image is rather complex and shows a lot of foreground objects, including a person, a boat, a fish, the lake and the trees, while the textual sentence contains cues which can be directly used to disambiguate, such as "fishing", "lake" and "catch".



(a) "portfolio 2" (b) "lake fork fishing guide, bass guide - guarantees bass catch"

Figure 2: Examples selected from UIUC-ISD dataset [29] for sense "bass (fish)".

Image classification and text classification are also complementary to each other. For some documents text classification generates correct results while for others, image classification generates correct results. The reasons are twofold: first, the content in images and text are complementary to each other; second, text classification usually has higher confidence of its classification results than image classification, while it frequently fails to classify a lot of unseen documents. In other words, text classification has higher confidence and smaller coverage, while image classification has lower confidence and larger coverage. This observation motivated us to propose a probabilistic ensemble fusion model to combine the results of text classification and image classification, which is explained in Section 4. "coverage" here means the percentage of documents a classification approach can effectively classify. For example, if the Yarowsky classifier cannot determine which sense the keyword "bass" carries in an unseen sentence, we say the Yarowsky classifier cannot effectively classify this document. More details on "coverage" and the complementary relation between image classification and text classification will be explained in Section 5.

Complementary and correlative relations can be both leveraged in multimodal classification tasks such as WSD to achieve high accuracy. They usually co-exist inside the same datasets, while they are probably presented in different documents. These two relations reveal the potential of using multimodal fusion to achieve higher performance than single-modality approaches, since multimodal data can either provide additional information or emphasize the same semantic information.

IV. Algorithm

In this section, the probabilistic ensemble fusion algorithm we used to combine image classification and text classification to solve WSD are explained. "ensemble fusion" means text classification and image classification are conducted on text and images separately and a fusion algorithm is used to combine the results. "probabilistic" means the results of text classification and image classification carry labels with probabilities or confidence scores and the final results of the fusion model are also probabilistic. What we expect is after fusion, we can generate labels for testing samples with higher accuracy. The details about the implementation are also described here.

A. Probabilistic Ensemble Fusion

In the probabilistic ensemble fusion model, images and text are first classified for disambiguation separately to provide confidence scores to senses for ambiguous words. Then the confidence scores are combined using different approaches, including the linear rule, the maximum rule and logistic regress classification, to generate the final confidence scores for senses.

Let's use **s** to denote the vector $(s_1, s_2, s_3, .., s_n)^T$ as the n senses of a keyword w, and **c** to denote the vector $(c_1, c_2, c_3, .., c_n)^T$ as confidence scores of the n senses of a keyword w in one document d. The process of disambiguating a testing sample in our probabilistic ensemble fusion model is shown in Figure 3.

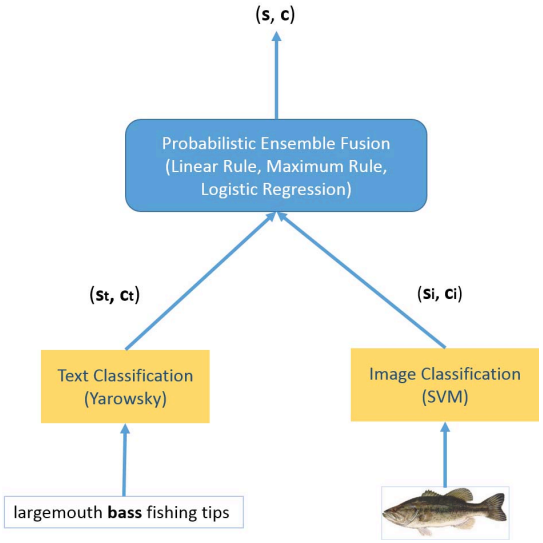


Figure 3: The probabilistic ensemble fusion model

For text classification, the original Yarowsky algorithm as stated in [16] is implemented. For image classification, we use SIFT [28] to extract local features and the bag-of-visual-words model [27] to represent images. Then, a SVM classifier is trained on the bag-of-visual-words vectors to classify images. Both the image classifier and the text classifier generate sense annotations along with confidence scores for testing samples. The image classifier and text classifier are trained on training datasets, as will be shown in Section 5.

Before we further explain these three fusion approaches, let's simplify the problem: say for one keyword w with two senses s_1 and s_2, and a document d with one image i and a textual sentence t, the image classifier generates (s_1, c_{i1}) and (s_2, c_{i2}), and the text classifier generates (s_1, c_{t1}) and (s_2, c_{t2}), where c_{i1}, c_{i2}, c_{t1} and c_{t2} denoting the confidence scores of senses s_1 and s_2 generated by image classification and text classification respectively. Confidence scores are normalized into $[0, 1]$ interval.

1) Linear Rule: The linear rule fusion uses a weight λ to combine the confidence scores for the two senses from images and text. The combined confidence scores for s_1 and s_2 are:

$$c_1 = \lambda \times c_{i1} + (1 - \lambda) \times c_{t1} \quad (1)$$

$$c_2 = \lambda \times c_{i2} + (1 - \lambda) \times c_{t2} \quad (2)$$

λ is calculated by dividing the accuracy of image classification by the sum of accuracy of text classification and image classification on the validation dataset:

$$\lambda = Accuracy_i / (Accuracy_i + Accuracy_t) \quad (3)$$

2) Maximum Rule: The maximum rule selects the sense s with the highest confidence score c from (s_1, c_{i1}), (s_2, c_{i2}), (s_1, c_{t1}) and (s_2, c_{t2}). For example, with $(s_1, 0.45)$ and $(s_2, 0.55)$ from image classification and $(s_1, 0.91)$ and $(s_2, 0.09)$ from text classification, we choose s_1 as the output sense for the document d according to the maximum rule, because the text classification outputs the highest confidence score 0.91 for sense s_1.

3) Logistic Regression: For logistic regression, confidence scores from two modalities, c_{i1}, c_{i2}, c_{t1} and c_{t2}, are used as features to train the logistic regression classifier on the validation dataset. Then the logistic regression classifier is used to classify the testing samples to get sense annotations with confidence scores. Logistic regression is chosen for its non-linear transformation of the confidence scores compared to the linear rule approach.

Our probabilistic ensemble model is simple but powerful. The experimental results in Section 5 demonstrate the effectiveness of our model. In addition, the model can be viewed as a general framework for multimodal fusion, where you can come up with new fusion approaches to combine the confidence scores from text classification and image classification, or new text classification and image classification methods. It also can be expanded to more modalities, such as audios and videos, beyond only images and text.

B. Implementation

The Yarowsky algorithm [16], is an unsupervised and iterative learning algorithm. It starts with a small set of seed rules to disambiguate senses and a large untagged corpus. In each iteration, the algorithm first applies known rules to untagged samples and learns a set of new rules from new tagged samples. This process is repeated until all training samples are tagged, and the learned rules are ordered by decreasing confidence scores, which are determined by the numbers of samples supporting the rules. When given an unseen testing sample, the algorithm returns the first rule in the ordered list which matches the testing sentence and the confidence score of the matched rule. The Yarowsky algorithm implementation is written in C++ and the pseudo probability distribution is implemented over Yarowsky classifier using a Python wrapper.

For image classification, OpenCV for Python is used to extract SIFT features from images, the K-Means from Python scikit-learn is used to generate visual words and multi-class SVM from Python scikit-learn is used to train a multi-class SVM classifier and generate confidence scores for labels on testing samples.

V. Experiments

Experiments were run on the UIUC-ISD dataset [29] to test the accuracy of the three fusion approaches used in our probabilistic ensemble fusion model. Results demonstrated these three fusion approaches achieved higher accuracy compared to the text-only classification and image-only classification methods.

A. Dataset

The multimodal UIUC-ISD dataset [29] is used to test the accuracy results of the Yarowsky algorithm, the SVM classifier and our ensemble fusion algorithms. There are three keywords "bass", "crane" and "squash" in the dataset. For each keyword, we selected two core senses. There are 1691 documents for "bass", 1194 documents for "crane" and 673 documents for "squash".

We have constructed a training dataset, a validation dataset and a testing dataset for each keyword. The training dataset is used to train the image and text classifiers. The validation data is used to train the logistic regression classifier and select the linear weight λ based on the accuracy of the image classification and text classification on the validation dataset. The testing dataset is used to evaluate the fusion algorithms and to demonstrate that by using multimodal fusion, we can get higher disambiguation accuracy compared to methods using single modality.

B. Results

The experimental results on the UIUC-ISD dataset are shown in Table 1. From the table, the accuracy of the three fusion methods is much higher than the image-only and text-only methods on "bass" and "crane". For "bass", the ensemble approaches improved the accuracy from 0.565 to 0.871. For "crane", the maximum rule approach improved the accuracy from 0.642 to 0.808. For "squash", because the accuracy of text-only classification is low (0.188), we cannot get much additional information from the text-only classification. Therefore the accuracy of the three fusion approaches for "squash" is quite similar to the image-only classification.

Table I: The accuracy of image-only, text-only, linear rule fusion, maximum rule fusion and logistic regression fusion

	Image	Text	Linear-Rule	Max-Rule	Log-Reg
bass	0.565	0.365	**0.871**	**0.871**	**0.871**
crane	0.642	0.333	0.800	**0.808**	0.775
squash	0.754	0.188	**0.768**	0.754	0.754

C. Analysis

As discussed in Section 3, text classification and image classification are complementary to each other. In this section we further explain why our ensemble fusion model works and how it can capture the complementary relation between image classification and text classification, based on the experiments conducted on UIUC-ISD dataset.

In our experiments, Yarowsky classifier works well when the testing sentences contain patterns that have been discovered in training datasets. It will generate very high confidence scores for the correct senses in most cases, for example, $(s_1, 1.0)$ and $(s_2, 0.0)$ or $(s_1, 0.95)$ and $(s_2, 0.05)$, assuming s_1 is the correct sense. However for the documents that do not contain known patterns, Yarowsky classifier would fail to disambiguate between two senses and output $(s_1, 0.0)$ and $(s_2, 0.0)$. On the other hand, the image classification is less accurate, hence generate less confident results, for example, $(s_1, 0.55)$ and $(s_2, 0.45)$ or $(s_1, 0.60)$ and $(s_2, 0.40)$, with s_1 possibly to be a wrong label.

Hence, for documents in which the Yarowsky classifier works, the results of these three fusion approaches in our ensemble fusion model would be consistent with Yarowsky since the Yarowsky outputs results with very high confidence scores. For other documents in which the Yarowsky classifier fails, the results of these three fusion approaches in the ensemble fusion model would be consistent with the image SVM classifier because Yarowsky simply returns $(s_1, 0.0)$ and $(s_2, 0.0)$ for these documents.

Therefore, our ensemble fusion model can perfectly generate more accurate results by taking advantage of both the text classification and image classification and avoiding their drawbacks. This explains why our fusion model works well when image classification and text classification are reliable to some extant, and our fusion model can still generate as good results as any single-modality classification method in cases one of the classification approaches fails, for example, "squash" in the Table 1.

Thus our conclusion is, by combining the classification results of image-only and text-only classifiers under a probabilistic ensemble fusion framework, we can achieve higher accuracy in most cases compared to methods using only single modality. Even in cases one of the single-modality classifiers has very poor performance, the fusion model can still generate results as good as the best results from any single-modality classifier.

VI. CONCLUSION

We have presented the probabilistic ensemble fusion model based on image classification and text classification to achieve better accuracy than text-only and image-only classification on the WSD problem. Our main contributions focus on two aspects. First, we identified the correlative and complementary relations among different modalities and discovered that by utilizing these relations we can achieve better results than using data only from single modality. Second, we designed a probabilistic ensemble fusion model with different ensemble approaches to achieve better accuracy than image-only and text-only classification methods.

The next steps for us are to employ sophisticated algorithms to capture both the correlation and complementation among multiple modalities, prepare large-scale multimodal datasets, and improve the performance of fusion models on various classification tasks, including word sense disambiguation.

ACKNOWLEDGEMENTS

This work was partially supported by DARPA under FA8750-12-2-0348 and a generous gift from Pivotal.

REFERENCES

[1] R. Navigli. Word Sense Disambiguation: A Survey. ACM Computing Surveys, 2009.

[2] E. Agirre and P. Edmonds. Word Sense Disambiguation, Algorithms and Applications. ISBN 978-1-4020-4808-4, Springer 2007.

[3] R.L. Rivest. Learning Decision Lists. Mach. Learn. 2, 3, 229246, 1987.

[4] Quinlan. Programs for Machine Learning. Morgan Kaufmann, San Francisco, CA, 1993.

[5] H. Schutze. Dimensions of Meaning. In Supercomputing 1992: Proceedings of the ACM/IEEE Conference on Supercomputing.

[6] H. Schutze. Automatic Word Sense Discrimination. In Proceedings of Computational Lingustics, 1998.

[7] D. Lin. Automatic Retrieval and Clustering of Similar Words. In Proceedings of the 17th International Conference on Computational linguistics, 768774, 1998.

[8] S. Bordag. Word Sense Induction: Triplet-based Clustering and Automatic Evaluation. Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, 137144, 2006.

[9] M. Lesk. Automatic Sense Disambiguation using Machine Readable Dictionaries: How to tell a pine cone from an ice cream cone. In Proceedings of the 5th SIGDOC, 2426, 1986.

[10] S. Banerjee and T. Pederen. Extended Gloss Overlaps as a Measure of Semantic Relatedness. In Proceedings of the 18th International Joint Conference on Artificial Intelligence, 805810, 2003.

[11] D. Hindle and M. Rooth. Structural Ambiguity and Lexical Relations. Computat. Ling. 103120, 1993.

[12] S. Abney and M. Light. Hiding a Semantic Class Hierarchy in a Markov Model. In Proceedings of the ACL Workshop on Unsupervised Learning in Natural Language Processing, 18, 1999.

[13] D. Mccarthy and J. Carroll. Disambiguating Nouns, Verbs and Adjectives using Automatically Acquired Selectional Preferences. Computat. Ling., 639654, 2003.

[14] M. Sussna. Word Sense Disambiguation for Free-Text Indexing using a Massive Semantic Network. In Proceedings of the 2nd International Conference on Information and Knowledge Base Management, 6774, 1993.

[15] E. Agirre and G. Rigau. Word Sense Disambiguation using Conceptual Density. In Proceedings of the 16th International Conference on Computational Linguistics, 1622, 1996.

[16] D. Yarowsky. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. Proceeding of ACL 189-196, 1995.

[17] N. Rasiwasia, J. C. Pereira, E. Coviello, G. Doyle, G. R.G. Lanckriet, R. Levy, N. Vasconcelos. A New Approach to Cross-Modal Multimedia Retrieval. Proceedings of the international conference on Multimedia, 2010

[18] M. Paramita, M. Sanderson, and P. Clough. Diversity in photo retrieval: overview of the ImageCLEFPhoto. task 2009. CLEF working notes, 2009.

[19] R. Datta, D. Joshi, J. Li, and J. Wang. Image retrieval: Ideas, influences, and trends of the new age. ACM Computing Surveys (CSUR), 40(2):160, 2008.

[20] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and trecvid. In MIR 06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval, pages 321330, New York, NY, USA, 2006. ACM Press.

[21] C. Snoek and M. Worring. Multimodal video indexing: A review of the state-of-the-art. Multimedia Tools and Applications, 25(1):535, 2005.

[22] T. Tsikrika and J. Kludas. Overview of the wikipediaMM task at ImageCLEF 2009. In Working Notes for the CLEF 2009 Workshop, 2009.

[23] J. Ngiam, A. Khosla, et.al.. Multimodal Deep Learning. ICML 2011.

[24] Multimodal Learning with Deep Boltzmann Machines. N. Srivastava and R. Salakhutdinov. Journal of Machine Learning Research, Sept 2014.

[25] W. May, S. Fidler, et. al. Unsupervised Disambiguation of Image Captions. SemEval 2012

[26] K. Saenko and T. Darrell. Filtering Abstract Senses From Image Search Results. NIPS, 2009.

[27] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In Proceedings of ICCV, Oct 2003.

[28] D.G. Lowe, G. David. Object recognition from local scale invariant features. In proceedings of the International Conference on Computer Vision, 1999.

[29] UIUC-ISD dataset. http://vision.cs.uiuc.edu/isd/.