B1TB2216

Graduation Thesis

Learning visual attribute from image and text

Maharjan Sumit

March 2, 2015

Department of Information and Intelligent Systems
Tohoku University

Learning visual attribute from image and text[*]

Maharjan Sumit

Abstract

Visual attributes are the words describing appearance properties of an object. They have an interesting property in that they are linguistic entities and yet bear a strong connection to visual information. In other words, it would be impossible to learn a meaningful representation of visual attributes only from language context.

In this paper, we make a preliminary study on the approach to learn the meaning of visual attributes from both image and text. We collect a large scale dataset of images and texts from the real-world on-line marketplace. We then attempt to learn a grounded representation of automatically generated attributes from the dataset using Canonical Correlation Analysis (CCA) projecting both image and text representation into a common subspace. This encodes both visual and semantic meaning of a word. Through empirical study, we show how grounded learning changes the meaning of the attribute word through the multimodal grounding.

Keywords:

natural language processing, distributional semantics, machine learning, grounding, computer vision, multimodal

[*] Graduation Thesis, Department of Information and Intelligent Systems, Tohoku University, B1TB2216, March 2, 2015.

Contents

List of Figures

List of Tables

1 Introduction

One of the biggest challenges in the field of Natural Language Processing is proper representation of meaning of words. Recent years have seen a surge in the use of vector space models [12] [3] which are based on distributional hypothesis (similar words appear in similar linguistic contexts). These distributional models approximate vectors that keep track of the patterns of co-occurence of the word in the text corpora, so that the degree of semantic similarity between two or more words can be compared.

These models have been successful in many natural language applications but still have several limitations. They cannot distinguish between synonyms and antonyms since both of them tend to appear in similar linguistic contexts. They represent the meaning of a word entirely in terms of connections to other words. However, meaning of a word has a strong connection with some component in the real world and hence distributional models cannot connect language to its actual meaning in the real world.

I own a really ***fat*** cat.
I own a really ***over-sized*** cat.

I own a really ***skinny*** cat.

Figure 1: DSMs cannot distinguish between synonyms(fat and oversized) and anytonyms(fat and skinny) since all of them(fat, oversized and skinny) tend to appear in similar linguistic context.

Visual attributes are the example of linguistic entities that bear a strong connection to visual information. Visual attributes are the words describing appearance properties of an object. For example, one might use *gray* or *brown* and *furry* to describe a cat. Visual attributes have been studied in the computer vision community [4], where the main focus has been in the automatic recognition of attributes from an image. Hence, it would be impossible to learn a meaningful representation of visual attributes only from language context.

In this paper, we make a preliminary study on the approach to learn the meaning of visual attributes from both image and text. We collect a large scale dataset of images and texts from the real-world on-line marketplace. We then attempt to learn a grounded representation of attributes generated automatically from the dataset using Canonical Correlation Analysis (CCA). CCA allows us to project both image and text representation into a common subspace, which encodes both visual and semantic meaning of a word. Through empirical study, we show how grounded learning changes the meaning of the attribute word through the multi-modal grounding.

Grounded representation can be expected to generate better vector models since they encode information from not only linguistic but also visual data. Using the differences in the appearance property it can help to clarify the differences between synonyms and antonyms. Since they relate text with image, it can be used for image retrival and image descriptions tasks. Also, the grounded representation can capture the visual similarity across different semantic contexts; e.g., distance between *metallic* in terms of material and *gray* in terms of color.

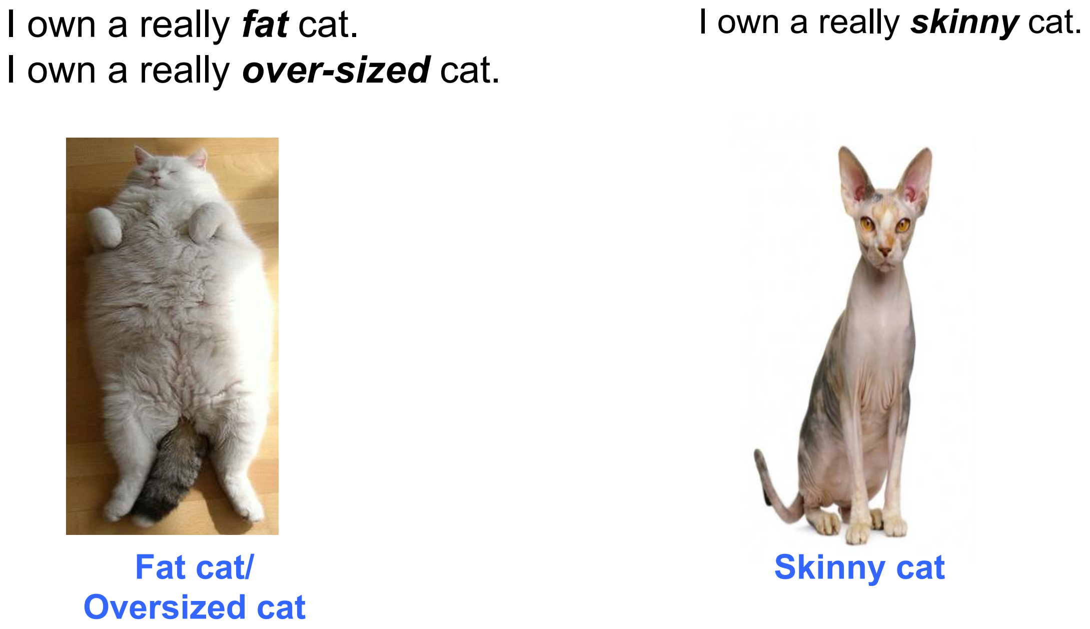

I own a really ***fat*** cat.
I own a really ***over-sized*** cat.
I own a really ***skinny*** cat.
Fat cat/
Oversized cat
Skinny cat

Figure 2: Including visual information can help to distinguish synonyms and antonyms as they tend to refer to different types of images.

Our contribution in the paper is summarized in the following points.

- Large-scale dataset of image and text rich in attribute description, collected from a real-world on-line market Etsy.
- Preliminary empirical evaluation of grounded learning by CCA.

2 Related Work

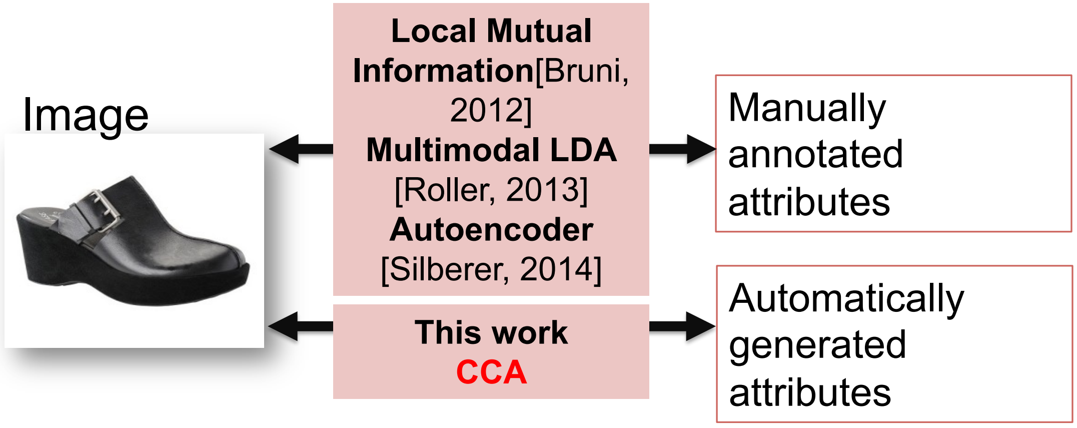


Image
Local Mutual Information[Bruni, 2012]
Multimodal LDA [Roller, 2013]
Autoencoder [Silberer, 2014]
This work
CCA
Manually annotated attributes
Automatically generated attributes

Figure 3: Previous works performed multimodal grounding on manually amnnotated attributes using different methods while in this paper we perform grounding on automatically generated attributes using CCA.

2.1 Visual attributes

Visual attributes have been getting attention in the computer vision community as a mean of categorization [4, 1, 15], zero-shot learning [8], or sentence generation from an image [13, 19]. Previous works mainly focused on the recognition of the attributes from an image. In this paper, we are interested in learning the grounded representation of attributes from both semantic and visual meaning.

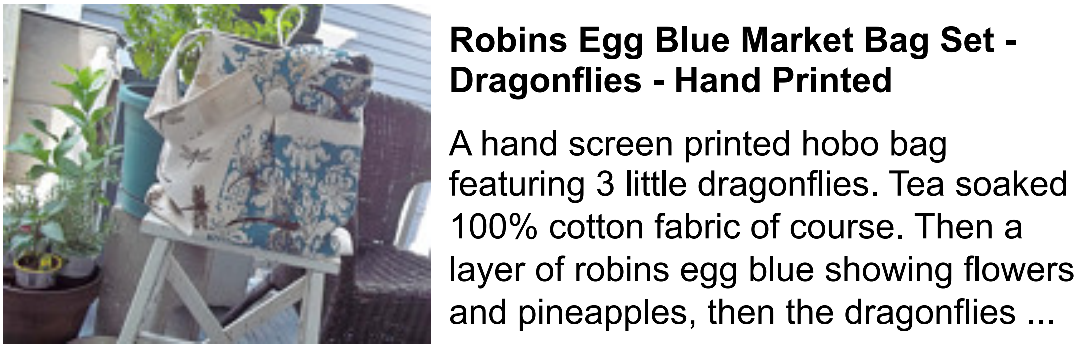
2.2 Image and text corpus

There has been a continuous effort in building a larger corpus of image and text in the research community, such as SBU1M [14], Flickr30k [5], or Microsoft COCO [10]. All of them contain a large collections of pairs of image and one or more sentences. However, these existing datasets are not necessarily designed to study visual attributes. On the other hand, attribute-focused datasets in the vision community [16, 18] do not come with text descriptions. Visual attributes are usually manually annotated by looking at the image according to some mentioned guidelines. This task is costly and can generate only limited amount of data while deciding the guidelines is a difficult task. Furthermore, there is not much information on what type of words are the most appropriate visual attributes. Hence, the visual attributes generated by manual annotation is limited in number.

2.3 Multimodal grounding

Recently there is a surge of interest in learning grounded representation of lexical meaning by various approaches, including Local Mutual Information[2], Multimodal LDA [17], or Autoencoder [20], since the early work using CCA [9]. All these previous works use manually annotated visual attributes.

In this paper, we build a large corpus of image and text focused on visual attributes using an on-line market. Instead of manually annotating visual attributes on each of the images, we automatically extract the attributes from the corpus and make a preliminary study on the grounded learning using CCA.



Robins Egg Blue Market Bag Set - Dragonflies - Hand Printed

A hand screen printed hobo bag featuring 3 little dragonflies. Tea soaked 100% cotton fabric of course. Then a layer of robins egg blue showing flowers and pineapples, then the dragonflies ...

Figure 4: Example item from Etsy. Each item consists of a title, an image, and a description about the product. Notice the visual description about the product, such as color, pattern, or material.

3 Etsy Dataset

In this paper, we have collected data from Etsy, an on-line marketplace of handmade crafts. From the website, we have crawled 1,216,512 items that spans across various product categories including art, fabric, gadget accessories, or bath products. Each item consists of a title, an image of 170 × 135 pixels, a product description, and other metadata such as price. Figure 4 shows an example of an item.

Etsy has a favorable property for us to study visual attributes. As sellers try to describe a product in detail to make the product sound more attractive, the seller description tends to include detailed information about product appearance rich in attributes, as well as other information such as size, usage, delivery method, etc.

4 Attribute Representation

Given a pair of image and bag-of-attributes, we wish to learn the grounded representation of the attribute words. Our strategy is first to process image and text part separately, then combine these two vectors using CCA.

4.1 Image Representation

In this paper, we represent the image by a simple color histogram. Given an image, we calculate 16-bin histogram for red, green, and blue channels and concatenate them to form a 48 dimensional vector. We sticked to the simple color histogram in this paper for visualization purpose since using higher-level image features do not necessarily produce an easily interpretable visualization.

4.2 Text Representation

Visual attributes could be various linguistic entities, from adjectives, nouns, to verbs. In this paper, we restrict attribute-vocabulary to manually-annotated 248 adjectives for simplification purpose. Note that it is straightforward to expand the range of attributes.

We represent each attribute-word by word2vec [12], which we learn from the dataset. For each item in the dataset, we first normalized text in the title and the description by removing URL, email address, or formatting string. Using the normalized texts as a corpus, we learn a 300 dimensional word2vec model.

Each item in our dataset contains multiple attributes. To represent each item using word2vec, we calculate the average vector for each item. We applied a POS tagger [11] to find adjectives in the title and the text, and kept adjectives with minimum document frequency of 100. This preprocessing left us with a bag of attribute-words for each item. Then we calculated the average word2vec from this bag to obtain a vector representation for items.

After removing items without any attribute words, we obtained 943,934 items from the Etsy dataset. Out of the 943,934 pairs of text and image vector pairs, we used 643,934 pairs for training and the remaining 300,000 pairs for testing purpose.

4.3 Canonical Correlation Analysis

Canonical correlation analysis [7, 6] is a method for exploring the relationships between two multivariate sets of variables, all measured on the same individual.

Consider two random variable $X \in R^{n_1}$ and $Y \in R^{n_2}$ both characterizing the same object, each a distinct view offering different information. So from them, we want to derive new variables $U, V \in R^m$(where $m \leq min(n_1, n_2)$) whose correlation is maximized. CCA finds $U = (U_1 \dots U_m)$ and $V = (V_1 \dots V_m)$ such that

$$U_i, V_i = \arg\max_{\phi,\psi \in R} Corr(\phi, \psi) \tag{1}$$

under the constraint that $\phi = a^T X$ and $\psi = b^T Y$ for some vectors $a \in R^{n_1}$ and $b \in R^{n_2}$ and that

$$Corr(\phi, U_i) = Corr(\phi, V_j) = 0 \tag{2}$$

for $j = 1 \dots i-1$. These new variables U and V found by CCA can be viewed as m-dimensional representations of $X \in R^{n_1}$ and $Y \in R^{n_2}$ that have incorporated our prior belief that X and Y are referring to the same object.

Since we have text vectors (consider X) and image vectors (consider Y) referring to the same item, we can generate CCA vectors U and V for the item.

Using the 300 dimensional word2vec and the 48 dimensional color histogram, we calculated 48 dimensional CCA vectors in this paper.

5 Evaluation

We qualitatively evaluate CCA representation by visualization and retrieval tasks.

5.1 Visualization

We first contrast the visualization of word2vec space and CCA space, and see how including visual information changes the distribution of words in the vector space. For visualization, we used the t-SNE algorithm [21] to reduce the dimensionality of word2vec and CCA representations, and plotted each word in our vocabulary list into 2-dimensional space. The visualization is shown in Figure 6. Different colors in the plot represent different semantic categories of the words, such as color, shape, size, etc.

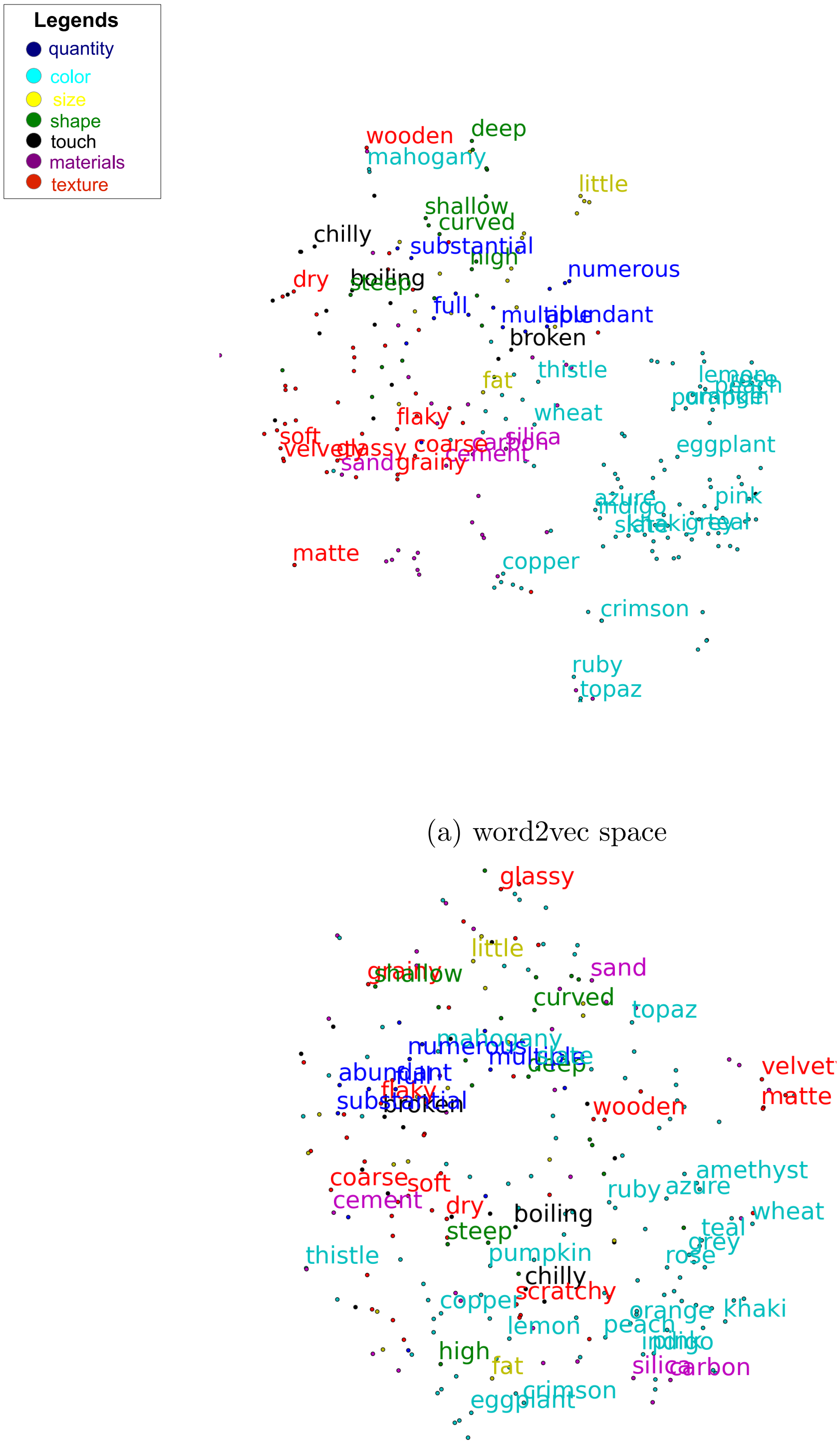
From the visualization, we can observe semantic clusters of words are much more prominent in word2vec space but in the CCA space the distance between different semantic categories are reduced. This is an expected result, because incorporating visual information will tie words in the different semantic categories (e.g., color and material) with respect to visual similarity between them.

5.2 Word-to-word retrieval

In order to further see the change in distribution of words, we calculated nearest neighbors for query words in both the word2vec and CCA space. Table 1 lists the top 10 nearest neighbors in order.

The result is consistent with the visualization; we can observe that while in the word2vec space all the nearest neighbors for *black* are colors, semantically different *metallic* and *skinny* appear in CCA space.

The word2vec model uses linguistic context to create vector models of words. Since semantically-similar words appear in similar linguistic context they tend to be closer to each other. When contextual information from image is also included (as in CCA) the words appearing in similar visual contexts move closer to each other while appearing in different visual context move further from each other.



(a) word2vec space

(b) CCA space

Figure 5: t-SNE visualization of words by semantic category in the word2vec and the CCA space

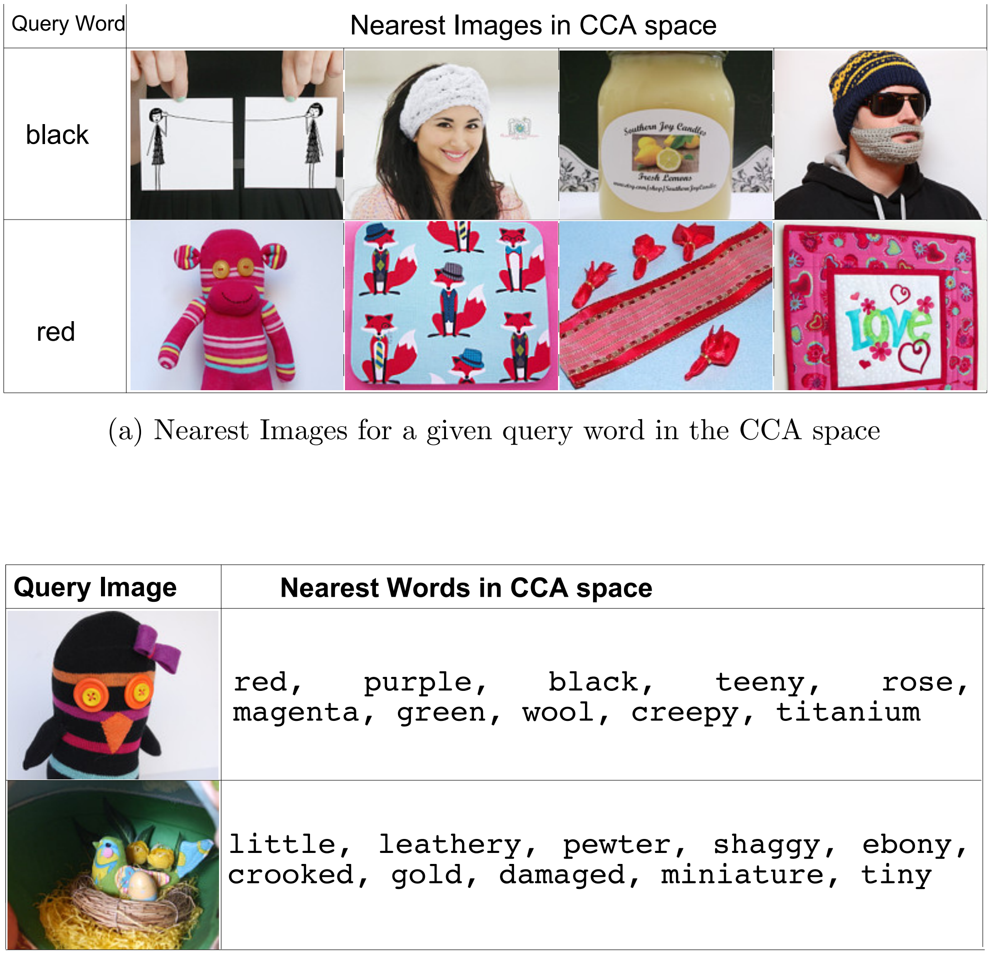
query word	**Nearest neighbors in word2vec space**	**Nearest neighbors in CCA space**
black	red, white, grey, gray, brown, blue, ivory, teal, purple, green	white, grey, blue, ivory, gray, metallic, red, skinny, brown, teal
red	purple, pink, black, teal, blue, brown, yellow, orange, green, maroon	pink, gray, purple, blue, grey, viscose, white, green, coral, beige
white	ivory, beige, red, black, brown, yellow, gray, grey, pink, blue	black, red, ivory, blue, gray, green, metallic, grey, beige, cream

Table 1: Nearest neighbors for a given query word in the word2vec and the CCA space

5.3 Cross-modal retrieval

We also demonstrate the advantage of our grounded model in the image retrieval and image description (word-retrieval) tasks. In the retrieval task, we give a query word and find the nearest image to it in CCA space of test dataset. In the description task, we give a query image and find the nearest words to the image in CCA space of test dataset. Note that this cross-modal retrieval is impossible only using either image-based representation or text-based representation. The retrieval results are shown in figure 6a, and word-retrieval results are shown in figure 6b.

We observe that image retrieval task as expected generates images that are quite closer to the given query words *black* and *pink* while the image description task generates several justifiable words in its nearest neighbors. This demonstrates that mapping of visual attribute words to some visual attribute in image has taken place.


Query Word
Nearest Images in CCA space
black
red
Query Image
Nearest Words in CCA space
red, purple, black, teeny, rose, magenta, green, wool, creepy, titanium
little, leathery, pewter, shaggy, ebony, crooked, gold, damaged, miniature, tiny

(a) Nearest Images for a given query word in the CCA space

(b) Nearest words for a given query image in the CCA space

Figure 6: Nearest words/images for a given query word/image in the CCA space

6 Discussion

We performed multimodal grounding using weakly-supervised data obtained from the Web. We noticed several challenges in performing this task: The objects or attributes present in the image are often missing in the text while the text might contain unrelated or sometimes information that relates to completely different type of image features, such as an item available in multiple colors have each color word listed in its description. Or, the described attribute in the image might refer to significantly small section in the image with dominant background regions. Finally, the same feature describing different objects might refer to completely different image features. For example, white skin, white hair and white wall all use the same adjective *white* but refer to different type of image features. It is our future work to solve these issues.

We have not been able to demonstrate the change in the distance between synonyms and antonyms due to limited number and type of vocabulary used, in future we will investigate this process with expanded vocabulary and image features with different types of visual information.

7 Conclusion

In this paper, we introduced a new large-scale dataset of on-line market consisting of image and text that is focused on detailed item description, and presented a preliminary study of using CCA to map vectors representing text and image into a common subspace, through visualization and retrieval tasks. In the future, we would like to extend our work to larger vocabulary with explicit noise modeling in the real-world Web data and experiment with other models.

Acknowledgements

I would like to express my deepest appreciation to Professor Inui Kentaro, Professor Okatani Takayuki, Associate Professor Naoaki Okazaki, Assistant Professor Yamaguchi Kota, Masaki Saito for continuous guidance and inspiration for the research. Without their supervision and constant help, this thesis would not have been possible. I would also like to thank all the Inui-Okazaki laboratory members, for the valuable comments and constant support.

References

[1] Tamara L. Berg, Alexander C. Berg, and Jonathan Shih. Automatic attribute discovery and characterization from noisy web data. In *ECCV*, 2010.

[2] Elia Bruni, Jasper Uijlings, Marco Baroni, and Nicu Sebe. Distributional semantics with eyes: Using image analysis to improve computational representations of word meaning. In *ACM Multimedia*, pages 1219–1228, 2012.

[3] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537, 2011.

[4] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *CVPR*, 2009.

[5] Yunchao Gong, Liwei Wang, Micah Hodosh, Julia Hockenmaier, and Svetlana Lazebnik. Improving image-sentence embeddings using large weakly annotated photo collections. In *ECCV*, pages 529–545. 2014.

[6] David Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004.

[7] Harold Hotelling. Relations between two sets of variates. *Biometrika*, pages 321–377, 1936.

[8] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009.

[9] Victor Lavrenko, R Manmatha, and Jiwoon Jeon. A model for learning the semantics of pictures. In *NIPS*, 2003.

[10] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[11] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *ACL*, pages 55–60, 2014.

[12] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119, 2013.

[13] Margaret Mitchell, Xufeng Han, Jesse Dodge, Alyssa Mensch, Amit Goyal, Alex Berg, Kota Yamaguchi, Tamara Berg, Karl Stratos, and Hal Daumé III. Midge: Generating image descriptions from computer vision detections. In *EACL*, pages 747–756, 2012.

[14] Vicente Ordonez, Girish Kulkarni, and Tamara L Berg. Im2text: Describing images using 1 million captioned photographs. In *NIPS*, pages 1143–1151, 2011.

[15] Devi Parikh and Kristen Grauman. Relative attributes. In *ICCV*, 2011.

[16] Genevieve Patterson, Chen Xu, Hang Su, and James Hays. The sun attribute database: Beyond categories for deeper scene understanding. *Intl J Computer Vision*, 108(1-2):59–81, 2014.

[17] Stephen Roller and Sabine Schulte Im Walde. A multimodal lda model integrating textual, cognitive and visual modalities. *EMNLP*, 2013.

[18] Olga Russakovsky and Li Fei-Fei. Attribute learning in large-scale datasets. In *ECCV*, 2010.

[19] Amir Sadovnik, Andrew Gallagher, and Tsuhan Chen. It's not polite to point: Describing people with uncertain attributes. In *CVPR*, pages 3089–3096, 2013.

[20] Carina Silberer and Mirella Lapata. Learning grounded meaning representations with autoencoders. In *ACL*, 2014.

[21] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *The Journal of Machine Learning Research*, 9(2579-2605):85, 2008.