Joint Image and Text Representation for Aesthetics Analysis

Ye Zhou[1], Xin Lu[2], Junping Zhang[1]* , James Z. Wang[3]

[1]Fudan University, China [2]Adobe Systems Inc., USA [3]The Pennsylvania State University, USA

ABSTRACT

Image aesthetics assessment is essential to multimedia applications such as image retrieval, and personalized image search and recommendation. Primarily relying on visual information and manually-supplied ratings, previous studies in this area have not adequately utilized higher-level semantic information. We incorporate additional textual phrases from user comments to jointly represent image aesthetics utilizing multimodal Deep Boltzmann Machine. Given an image, without requiring any associated user comments, the proposed algorithm automatically infers the joint representation and predicts the aesthetics category of the image. We construct the AVA-Comments dataset to systematically evaluate the performance of the proposed algorithm. Experimental results indicate that the proposed joint representation improves the performance of aesthetics assessment on the benchmarking AVA dataset, comparing with only visual features.

Keywords

Deep Boltzmann Machine, Image Aesthetics Assessment, Multimodal Analysis

1. INTRODUCTION

Image aesthetics assessment [2] aims at predicting the aesthetic value of images as consistent with a human population's perception as possible. It can be useful in recommendation systems by suggesting aesthetically appealing images to photographers. It can also serve as a ranking cue for both personal and cloud-based image collections.

In previous studies [3, 13], image aesthetics is commonly represented by an average rating or a distribution of ratings by human raters. Image aesthetics assessment is then formulated as a classification or regression problem. Under this framework, some effective aesthetics representations are proposed, such as manually-designed features [2, 4, 9, 1, 3, 13], generic visual features [10], and automatically learned features [8]. These approaches, including the more recent deep learning approaches [8, 17, 16], have generated good quality aesthetics ratings or categories (*e.g.*, high vs. low aesthetics).

Ratings alone, however, are limited as a representation of visual aesthetics. Many image sharing sites, *e.g.*, Flickr, Photo.net, and Instagram, support user comments on images, allowing explanations of ratings. User comments usually introduce rationale as to how and why users rate the aesthetics of an image. For example, many factors can result in high ratings on an image. Comments such as "interesting subject", "vivid colors", or "a fine pose" are much more informative than a rating. Similarly, comments such as "too small" and "blurry" explain why low ratings occur. Motivated by this observation, this work leverages comments in addition to images and their aesthetics ratings to study aesthetics assessment.

In this work, we first systematically examine effective visual and textual extraction approaches for image aesthetics assessment. We then propose a joint representation learning approach, building upon the most effective visual and textual feature extraction approaches found through this process. We jointly examine both the visual and textual information of images during training. In the testing stage, we only use the visual information of an image to derive the aesthetics prediction.

Forming a joint representation for image aesthetics assessment is nontrivial because aesthetics-related phrases are unstructured and can include a wide range of concepts. Our contribution is an approach for learning from both images and text for aesthetics assessment. We utilize multimodal Deep Boltzmann Machine (DBM) [15] to encode both images and text. We analyze and demonstrate the power of this joint representation, comparing with using only visual information.

As we were conducting our study, we generated a dataset that the research community can use. The AVA-Comments dataset contains all user comments extracted from the provided page links in the AVA dataset, and enables the extraction of aesthetics information from user comments. It complements the benchmark AVA dataset for aesthetics assessment.

*Correspondence should be addressed to Junping Zhang, School of Computer Science, Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, Shanghai, China.
Emails: yezhou14@fudan.edu.cn, xinl@adobe.com, jpzhang@fudan.edu.cn, jwang@ist.psu.edu.

MM '16, October 15-19, 2016, Amsterdam, Netherlands

© 2016 ACM. ISBN 978-1-4503-3603-1/16/10…$15.00

DOI: http://dx.doi.org/10.1145/2964284.2967223

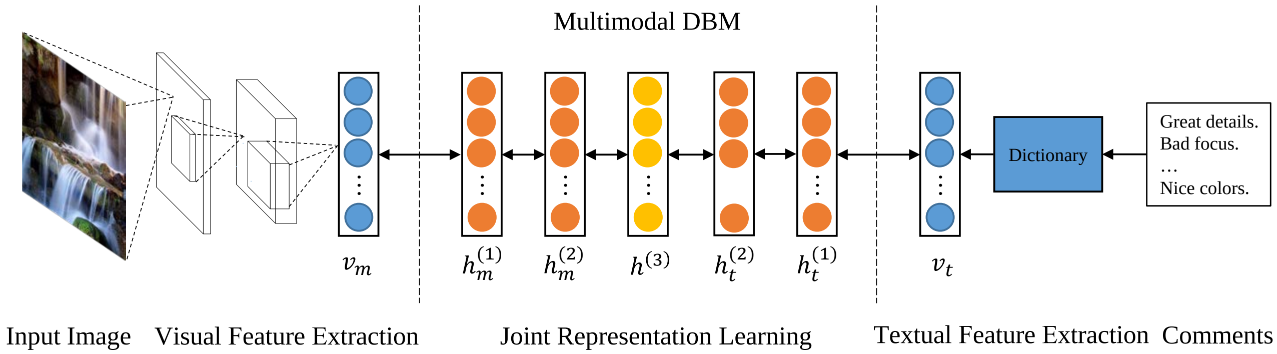

Figure 1: Approach Overview. Our approach consists of three modules: visual feature extraction, textual feature extraction, and joint representation learning.

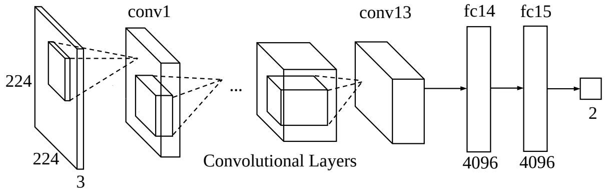

Figure 2: The architecture of the fine-tuned VGG-16 CNN model.

2. THE METHOD

We now provide technical details. As shown in Fig. 1, during training, we utilize a multimodal DBM to encode a joint representation with both visual and textual information, where visual information is extracted from deep neural network and textual information is collected from user comments. Importantly, test images are not associated with user comments. In testing, we leverage the learned representation to approximate the missing textual features and predict the aesthetic value of a given image.

2.1 Visual Feature Extraction

We demonstrate that pretrained VGG-16 model [14] on ImageNet can achieve reasonable performance for image aesthetics assessment, and a better performance is achieved by fine-tuning the VGG-16 model on an image aesthetics dataset. The detailed experimental results are discussed in Section 3.2.

We show the network structure during fine-tuning in Figure 2. It is noticeable that to deal with image aesthetics assessment as a binary classification problem, the last softmax layer with 1,000 classes on the ImageNet dataset is replaced by a 2-node softmax layer. The learning rate during fine-tuning process is lower than training from scratch, which avoids ruining the weights of the convolutional layer.

In testing, we adapt the multiview test performed in [6], and extracted 10 image patches from each testing image, which include the four corners and the central patch of the test input, and their horizontal flips.

We take the output of the fc15 layer (4,096-dimensional feature vector, illustrated in Fig. 2) of the fine-tuned network as the visual representation, and feature vectors from 10 patches are averaged to represent each image.

2.2 Textual Feature Extraction

Extracting aesthetics related information from user comments is not easy because user comments are unstructured and include discussions both about or beyond the content of the image. To examine the performance of different textual features on aesthetics classification, we build classifiers between the textual features and the aesthetics labels. Experimental results are shown in Section 3.3.

One of the most commonly used textual features is the word counts of a dictionary extracted from the original text. For the aesthetics assessment task, we extract a dictionary containing aesthetics-related words and phrases from user comments. In particular, we learn weights for all uni-grams, bi-grams and tri-grams occurred in user comments using a SVM variant with Naïve Bayes log-count ratios [18]. According to the experimental results discussed in [11], Naïve Bayes performs well on the snippets datasets and SVM is better at full-length reviews. A combination of Naïve Bayes like feature and SVM makes the model more robust.

Let $\mathbf{f}^{(i)} \in \{0,1\}^{|D|}$ indicate whether an item occurs, where D is the dictionary and $|D|$ is its cardinality. $\mathbf{f}_j^{(i)} = 1$ indicates that the dictionary item D_j occurs in the ith comment, otherwise $\mathbf{f}_j^{(i)} = 0$.

Denote $\mathbf{p} = \alpha + \sum_{i:y^{(i)}=1} \mathbf{f}^{(i)}$ and $\mathbf{q} = \alpha + \sum_{i:y^{(i)}=-1} \mathbf{f}^{(i)}$ as the positive and negative count vectors, repectively. α is a smooth parameter, which is usually set to $\alpha = 1$ for word counting vectors in the experiments. The log-count ratio is computed by $\mathbf{r} = \log((\mathbf{p}/||\mathbf{p}||_1)/(\mathbf{q}/||\mathbf{q}||_1))$, accordingly, the Naïve Bayes features $\mathbf{x}^{(i)}$ is derived by $\mathbf{x}^{(i)} = \mathbf{f}^{(i)} \circ \mathbf{r}$, where $\mathbf{a} \circ \mathbf{b}$ is the element-wise product of two vectors.

We train a L2-regularized L2-loss SVM to map the Naïve Bayes features to aesthetics labels. The Naïve Bayes features are sparse and high-dimensional because the dictionary size is large (over a million). We select a small dictionary by choosing several dimensions with maximum and minimum weight in $\hat{\mathbf{r}} = \mathbf{w} \circ \mathbf{r}$, where $\mathbf{w}$ is the weight vector in SVM. Then the word counts of the small dictionary can be calculated from user comments.

2.3 Multimodal Deep Boltzmann Machine

To perform a joint representation learning with visual and textual features, a multimodal DBM model [15] is built for aesthetics assessment and high-level semantics inference.

As shown in Fig. 1, the whole model contains three parts. For visual features, a DBM model is built between input units $\mathbf{v}_m$ and hidden units $\mathbf{h}_m^{(1)}, \mathbf{h}_m^{(2)}$. The layer between $\mathbf{v}_m$ and $\mathbf{h}_m^{(1)}$ is a Gaussian-Bernoulli RBM layer. For textual features, a DBM is built between input units $\mathbf{v}_t$ and hidden units $\mathbf{h}_t^{(1)}, \mathbf{h}_t^{(2)}$. The layer between $\mathbf{v}_t$ and $\mathbf{h}_t^{(1)}$ is a Replicated Softmax layer. The two models are combined with an additional hidden layer $\mathbf{h}^{(3)}$ between $\mathbf{h}_m^{(2)}$ and $\mathbf{h}_t^{(2)}$. The joint distribution over the multimodal input is as following.

$$P(\mathbf{v}_m, \mathbf{v}_t; \boldsymbol{\theta}) = \sum_{\mathbf{h}_m^{(2)}, \mathbf{h}_t^{(2)}, \mathbf{h}^{(3)}} \Bigg(\Bigg(\sum_{\mathbf{h}_m^{(1)}} P(\mathbf{v}_m, \mathbf{h}_m^{(1)}, \mathbf{h}_m^{(2)}) \Bigg) \Bigg(\sum_{\mathbf{h}_t^{(1)}} P(\mathbf{v}_t, \mathbf{h}_t^{(1)}, \mathbf{h}_t^{(2)}) \Bigg) P(\mathbf{h}_m^{(2)}, \mathbf{h}_t^{(2)}, \mathbf{h}^{(3)}) \Bigg) .$$

Both modality can be used in the training process. But during the testing process, user comments are not present. We therefore use variational inference [15] to approximate the posterior for the joint representation $P(\mathbf{h}^{(3)}|\mathbf{v}_m)$, or the missing textual representation $P(\mathbf{v}^t|\mathbf{v}^m)$.

3. EXPERIMENTS

3.1 The AVA and AVA-Comments Datasets

We adopted one of the largest image aesthetics dataset, the AVA [12], for evaluation. The AVA contains more than 255,000 user-rated images, each has an average of 200 ratings between 1 and 10. The dataset is divided into 235,000 images for training (including 167,000 high-quality images and 68,000 low-quality ones) and 20,000 images for testing without overlapping. The aesthetics assessment is formulated as a binary classification problem. Following the criteria in [12], images with a mean rating higher than 5 are labeled as high-quality images, the others as low-quality.

The user comments associated with each image are helpful for training aesthetics assessment model. For instance, images with smiling faces may not be visually appealing, but could still be rated with a high score; images with colorful content may still be rated low if they have boring content. Therefore, we crawled all the user comments for images in the AVA dataset to form the AVA-Comments dataset, where more than 1.5 million user comments were obtained from the original links. All user comments were tokenized, and all quotes and extra HTML tags such as links were removed. In our experiments, we found that user comments are particularly helpful in learning from these cases during network training. Moreover, the classification accuracy of aesthetics is improved without utilizing user comments during testing, as shown in Figs. 3 and 4. The AVA-Comments dataset is available to researchers.

3.2 Classification with Visual Features

We first evaluate the performance of the visual features used in this work. We compared the visual features discussed in Section 2.1 with pretrained VGG-16 [14] (denoted by Unmodified VGG-16), fine-tuned CaffeNet and VGG-16 presented in [17], and double-column CNN presented in [8]. For pretrained VGG-16 model, we built an L2-regularized L2-loss linear SVM to map features extracted from the layer fc15 of original VGG-16 model to aesthetics labels.

The classification results of these approaches are reported in Table 1. Our fine-tuned VGG-16 model performed the best, compared with other experimental settings. The fact that our fine-tuned model performed better than fine-tuned CaffeNet indicates that a larger capacity is required for the image aesthetics assessment task on the AVA dataset. Our fine-tuned VGG-16 is better than the result shown in [17] is because of one major setting difference: we fine-tuned VGG-16 with a larger batch size of 64 compared with 10. It means that a larger batch size may reduce the variance among mini-batches, resulting in better performance. Consequently, we use our fine-tuned VGG-16 as the visual feature extractor in the joint representation learning.

Table 1: Accuracy of Aesthetics Classification with Visual Features

Methods	Accuracy(%)
Double column CNN [8]	74.46
Unmodified VGG-16	74.12
Fine-tuned CaffeNet [17]	76.82
Fine-tuned VGG-16 [17]	77.09
Fine-tuned VGG-16 (Ours)	**78.19**

Table 2: Accuracy of Aesthetics Classification with Textual Features

Methods	Accuracy(%)
Recurrent Neural Networks [5]	79.36
Word2Vec [7]	79.62
Naïve Bayes SVM (whole dictionary)	80.90
Naïve Bayes SVM (3K small dictionary)	80.81

3.3 Classification with Textual Features

We now evaluate the performance of textual features on image aesthetics assessment. In both training and testing, user comments are available. For each image, we concatenate all user comments and generate textual input for feature extraction. We applied three commonly-adopted approaches for textual feature extraction, *i.e.*, Recurrent Neural Networks Language Model [5], Word2Vec [7], and Naïve Bayes features, on the AVA-Comments dataset.

To evaluate the performance of these feature extraction approaches, we built SVM classifier to map extracted features to aesthetics labels of images and examine the classification accuracy. Specifically, we trained a Recurrent Neural Networks Language Model [5] and a Word2Vec [7] model with the user comments. We then extracted textual features utilizing trained neural network models and built a L2-regularized L2-loss linear SVM classifier for aesthetics classification.

Meanwhile, we applied SVM on extracted Naïve Bayes features (denoted with “whole dictionary”). Since the dictionary size computed on the full training data is large (over a million), we further selected 2,000 items with the largest weights and 1,000 items with the smallest weights. These items indicate the most positive or negative phrases, forming a 3,000-item smaller dictionary. Note that we set number of positive and negative phrases according to the number of high and low quality images in the AVA dataset. A same classifier is built on word count vectors given the small dictionary and the results are shown in Table 2. The smaller dictionary is shown to keep aesthetics information to a large extent.

We present the classification results in Table 2, we found that Naïve Bayes SVM (whole dictionary) performed the best. We also found the Naïve Bayes SVM (3K small dictionary) achieved comparable performance. We thus adopted this scheme in the joint representation learning.

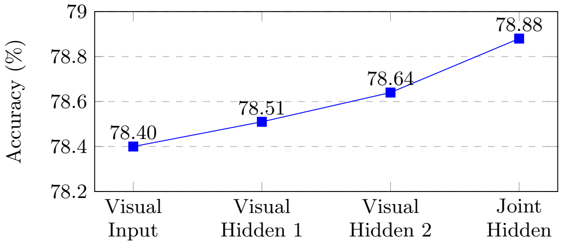

Figure 3: Classification accuracy of different layers in the multimodal DBM model with a single modality as input.

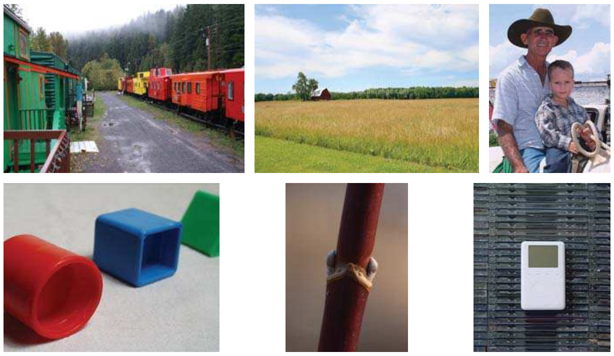

Figure 4: Test images correctly classified by the joint representation but misclassified by the visual feature. Images on the first row are high-aesthetics images and the second row are low-aesthetics images.

3.4 Classification with Joint Representation

We finally evaluate the performance of our proposed joint representation on image aesthetics assessment. The multimodal DBM model is trained as described in Section 2.3, where visual and textual features are extracted as discussed in Sections 3.2 and 3.3. The joint DBM model has three parts. The input dimension of a visual feature was 4,096, followed by a Gaussian RBM hidden layer with 3,072 nodes. The second hidden layer contained 2,048 nodes. Meanwhile, the word count vector of the 3,000-item dictionary was used as the textual input. The input dimension of a textual feature was 3,000, followed by a Replicated Softmax hidden layer of 1,536 nodes. The second hidden layer contained 1,024 nodes. To merge these two set of multimodal features, the joint hidden layer contains 3,072 nodes. The model was optimized using Contrastive Divergence (CD), and layer-wise pretraining described in [15] was performed. Finally, variational inference was used to generate the joint representation. Because we evaluate the aesthetic value of an image only based on its visual feature, the approach can be applied to situations where comments are not available.

To visualize the performance of neurons in each hidden layer in the multimodal DBM model, an L2-regularized L2-loss linear SVM that maps neurons in each layer to aesthetics labels is trained. As shown in Fig. 3, joint representation outperforms the visual feature representation, and the classification accuracy is gradually improved when utilizing representations from the input layer to the joint hidden layer.

To illustrate the difference between the visual feature representation and the joint representation of an image, we visualize several representative images that are correctly classified by the joint representation but misclassified by the visual feature in Fig. 4. The first row are high-aesthetics images and the second row are low-aesthetics ones.

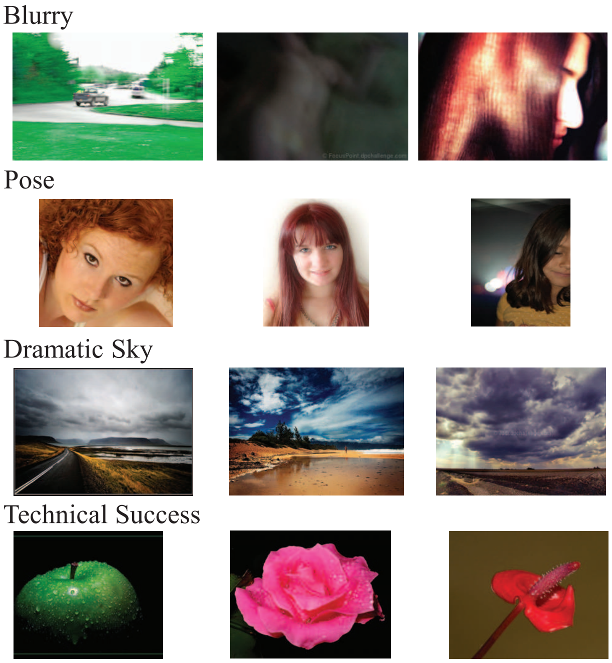

Figure 5: Representative items in the dictionary with their top-related images.

To examine the power of the joint representation in inferring the textual representation of test images, we picked several representative words and show their most closely related images. As shown in Fig. 5, images accurately represent the semantic meaning of words/phrases in the dictionary (*e.g.*, blurry, dramatic sky, and pose). Notably, like any learning-based methods, ours is limited when a phrase has few training examples. For instance, images related to "technical success" are often just macro photos.

4. CONCLUSIONS

This paper presented a multimodal DBM model to encode both images and their users' comments into a joint representation to improve image aesthetics assessment. Features extracted from a single modality, either images or user comments, were systematically evaluated. A joint representation was then built upon most effective visual and textual features. A large dataset with images and user comments, the AVA-Comments dataset, was built. Experiments on the AVA-Comments dataset showed that the proposed joint representation could improve image aesthetics prediction results produced by merely visual feature. The relatively small improvement indicates that the problem is challenging and requires more fundamental technological advancements.

Acknowledgments

This material is based upon work partially supported by the National Natural Science Foundation of China under Grant No. 61273299 and the National Science Foundation under Grant No. 1110970.

5. REFERENCES

[1] S. Bhattacharya, R. Sukthankar, and M. Shah. A framework for photo-quality assessment and enhancement based on visual aesthetics. In *ACM International Conference on Multimedia (MM)*, pages 271–280, 2010.

[2] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Studying aesthetics in photographic images using a computational approach. In *European Conference on Computer Vision (ECCV)*, pages III, 288–301, 2006.

[3] S. Dhar, V. Ordonez, and T. Berg. High level describable attributes for predicting aesthetics and interestingness. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1657–1664, June 2011.

[4] Y. Ke, X. Tang, and F. Jing. The design of high-level features for photo quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 419–426, 2006.

[5] S. Kombrink, T. Mikolov, M. Karafiát, and L. Burget. Recurrent neural network based language modeling in meeting recognition. In *INTERSPEECH*, pages 2877–2880, 2011.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1097–1105, 2012.

[7] Q. Le and T. Mikolov. Distributed representations of sentences and documents. In *International Conference on Machine Learning (ICML)*, pages 1188–1196, 2014.

[8] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang. RAPID: Rating pictorial aesthetics using deep learning. In *ACM International Conference on Multimedia (MM)*, pages 457–466, 2014.

[9] Y. Luo and X. Tang. Photo and video quality evaluation: Focusing on the subject. In *European Conference on Computer Vision (ECCV)*, pages 386–399, 2008.

[10] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka. Assessing the aesthetic quality of photographs using generic image descriptors. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1784–1791, 2011.

[11] G. Mesnil, T. Mikolov, M. Ranzato, and Y. Bengio. Ensemble of generative and discriminative techniques for sentiment analysis of movie reviews. *arXiv:1412.5335*, 2014.

[12] N. Murray, L. Marchesotti, and F. Perronnin. AVA: A large-scale database for aesthetic visual analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2048–2415, 2012.

[13] M. Nishiyama, T. Okabe, I. Sato, and Y. Sato. Aesthetic quality classification of photographs based on color harmony. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 33–40, 2011.

[14] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[15] N. Srivastava and R. R. Salakhutdinov. Multimodal learning with deep boltzmann machines. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2222–2230, 2012.

[16] X. Tian, Z. Dong, K. Yang, and T. Mei. Query-dependent aesthetic model with deep learning for photo quality assessment. *IEEE Transactions on Multimedia (TMM)*, 17(11):2035–2048, 2015.

[17] P. Veerina. Learning good taste: Classifying aesthetic images. Technical report, Stanford University, 2015.

[18] S. Wang and C. D. Manning. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 90–94, 2012.