Captioning Images Using Different Styles

Alexander Mathews
The Australian National University
Acton, Canberra, ACT
alex.mathews@anu.edu.au

ABSTRACT

I develop techniques that can be used to incorporate stylistic objectives into existing image captioning systems. Style is generally a very tricky concept to define, thus I concentrate on two specific components of style. First I develop a technique for predicting how people will name visual objects. I demonstrate that this technique could be used to generate captions with human like naming conventions. Full details are available in a recent publication [16]. Second I outline a system for generating sentences which express a strong positive or negative sentiment. Finally I present two possible future directions which are aimed at modelling style more generally. These are learning to imitate an individuals captioning style and generating a diverse set of captions for a single image.

Categories and Subject Descriptors

I.2 [**Artificial Intelligence**]: Vision and Scene Understanding; I.2 [**Artificial Intelligence**]: Natural Language Processing—*Language generation*

Keywords

caption generation; image description; object naming; sentiment

1. INTRODUCTION

Many of the photographs available on the Internet are not annotated with textual descriptions this makes them less easily accessible, to search engines and people with vision impairments. To alleviate this problem automatic annotations methods are needed. One area that has been receiving a lot of attention recently is the generation of natural language sentences to describe images.

Producing natural language captions for general web images is a very challenging problem; one which is strongly linked to the general image understanding problem. A system that can generate captions would ideally incorporate both a working model for natural language and a strong vision model capable of reliably detecting thousands of concepts. Unfortunately no reliable models for either of these tasks exist. Nevertheless, a range of approaches to automatic caption generation have been put forward.

The focus of my research is on improving systems for generating natural language captions for images. Unlike other researchers working on caption generation my aim is to improve the stylistic properties of captions, rather than focus exclusively on visual relevance. Although visual relevance is important, there are always many different ways to describe the same image. My goal is to select between these possible ways in order to optimise a set of stylistic objectives.

Writing style is strongly effected by the goal, whether it is to inform, persuade or entertain. When people write captions they typically have a goal in mind such as to describe the content of the image clearly. The clarity of a description can be strongly affected by the choice of words – a description dense in very specific scientific names is not suitable for a general audience. A caption author may also wish to encourage the reader to respond to the image in a particular way.

I focus my efforts on these two stylistic goals. First I investigate, and try to predict, how people name visual objects. Through this exploration I gain insights into generating captions with human like naming conventions. The result is a choice of names which are more easily understood by a general audience. Next I tailor a caption generation approach to express strong sentiment. The goal is to produce captions that encourage viewers to respond either positively or negatively towards an image. In my future work section I explore a couple of promising directions for allowing richer and more complex style domains to be used.

2. RELATED WORK

My thesis involves three main areas, caption generation as a whole, naming visual objects and sentiment in image captions. Caption generation is the framework in which my research fits, while the other two areas are the stylistic domains I have chosen to focus on.

2.1 Caption Generation

Over the last few years the challenging task of automatically generating natural language captions for images has grown in popularity; though it is still a very difficult problem. In fact two necessary components, visual concept detection and Natural Language Generation (NLG) are both unsolved research problems in their own right. A variety of attempts have been made to build complete caption generation systems, these tend to fall into one of three categories:

MM'15, October 26-30, 2015, Brisbane, Australia
ACM 978-1-4503-3459-4/15/10.
http://dx.doi.org/10.1145/2733373.2807998

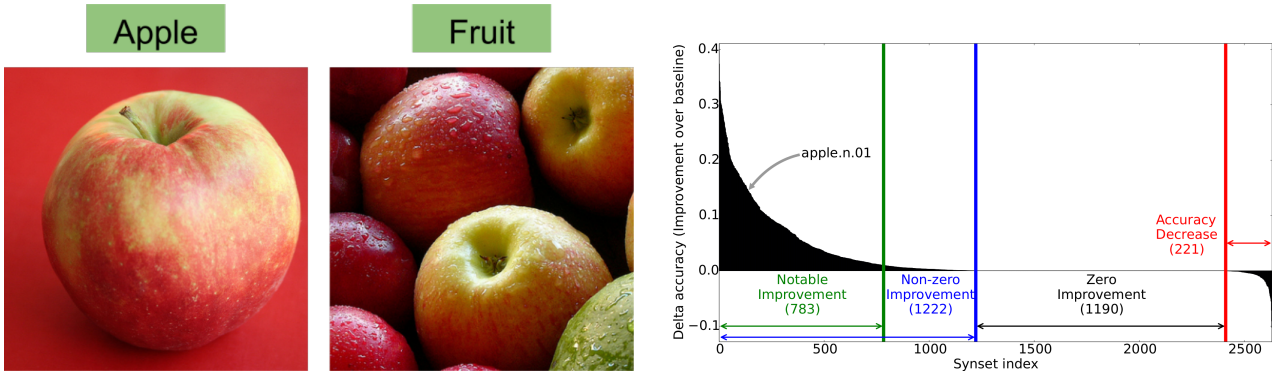

Figure 1: (Left)An example of the same object being named differently because of the visual context. (Right)The per-concept accuracy improvement of my context enriched method over a baseline which selects the most frequent name for each concept.

detection and generation, caption retrieval and generative deep networks.

Detection and generation approaches [27, 13, 20] employ a pipeline of independently trained object and attribute detectors; they typically follow a similar formula. First a fixed set of object and attribute detectors are used to identify visual concepts, these are then mapped into some intermediate representation, where smoothing is applied to enhance coherency. The smoothed intermediate representation is then realised as natural language text by a NLG system. A major problem with these approaches is that they use rigid intermediate representations and rely heavily on template systems for NLG; resulting in text which sounds very formulaic and un-natural.

Caption retrieval approaches [18, 8, 14] are designed to address the narrow scope and formulaic language that affects detection and generation approaches. Intuitively, transferring captions, or parts of captions [14], from other images should produce very natural image descriptions because most of the human imposed structure is kept in place. The challenge for caption retrieval approaches is ensuring that the relevance of the caption to the image.

Generative deep networks for natural language caption generation have recently become very popular [5, 10, 11, 15, 26]; quickly becoming the state of the art for caption generation. Generative deep networks consist of at least two components, a vision component, which is typically a Convolution Neural Network (CNN) [12], and a language component which is typically a Recurrent Neural Network (RNN). These two components are trained simultaneously using back-propagation.

In the related works mentioned above the primary aims were to improve the visual relevance or linguistic fluency of generated captions. Whereas, the aim of my thesis is to incorporate stylistic objectives into existing caption generation approaches.

2.2 Naming Visual Objects

A model for how people name objects was introduced in the psychology literature by Rosch [22]; this model revolved around the idea of basic-level categories. The basic-level category represents the ideal trade-off between low in-category visual variability and high between-class visual variability. Work by Rosch [21] notes that there are contextual effects on both the level of abstraction used to name an object and even if it will be named. Chaigneau et al. [2] demonstrate, using adult subjects, that situational information changes the way subjects categorize unfamiliar objects into familiar categories.

Recently there has been interest in developing machine learning techniques to choose the most appropriate name to give to a visual object. One approach taken by Deng et al. [4] is to optimise the accuracy-specificity trade off by using a semantic hierarchy to select the appropriate name. This technique does not take into account how people actually describe objects. Ordonez et al. [17] present a model predicting the labels people will actually use to name objects. Their model uses a text based component which trades off name frequency with linguistic proximity and a visual component that assesses the visual saliency of names to the image. This model does not capture the important affects of visual context. In section 3.1 I present my model which both incorporates visual context and works on a much larger scale than previous approaches.

2.3 Sentiment in Image Captions

Identifying the sentiment conveyed by both text and images is an active area of research. In contrast automatically generating text which conveys a strong sentiment has not been explored in great detail, though it is an emerging subfield of Natural Language Generation (NLG) called affective NLG.

Many techniques for text based sentiment analysis rely on aggregating known sentiment estimates for individual phrases. Two well know sources for phrase sentiments are SentiWordNet [6] and SentiStrength [25]; though many approaches learn phrase sentiments directly from the target domain. More recently authors such as Socher et al. [24] have shown how to jointly learn phrase sentiments and composition rules, allowing constructs such as negations to be captured.

Research towards identifying the sentiment in images has mostly been focused on using low level visual features such as color and contrast [9]. More recently Borth et al. [1] have made use of visual concept detectors to identify the sentiments expressed in images. They trained concept detectors for over 1200 adjective noun pairs identified as expressing a strong sentiment.

Work on affective NLG system [23, 19, 7] show that it is possible to generate sentences which convey a strong sentiment. Of particular interest is the work of Gatti et al. [7] which presents a way to automatically modify an existing

sentence to achieve a sentiment goal. Their approach uses constituent re-ordering to emphasise or de-emphasise parts of a sentence. They also swap short phrases with similar ones that align better with the sentiment goal. Although these techniques have been shown to apply to domains such as restaurant reviews [7] and medical reports [23], the image captioning domain has not been explored. My research is the first to investigate the possibility of using affective NLG for image caption generation.

3. APPROACH

My research so far has considered two specific style domains. The first is how to choose names for visual concepts in a way that is consistent with human judgements, while the second relates to generating sentences that exhibit a strong sentiment. The work on naming visual concepts is published [16], while the research into sentence generation with sentiment is ongoing.

3.1 Choosing the Most Appropriate Names

There are often many different names that can be used to describe a visual concept, for example Figure 1 presents a visual concept, in this case an *apple*. Other names that may be used include *fruit* and *Gala apple*. Depending upon the visual context some names are more likely than others, for example when there are other fruit in the image it is more common to name the *apple* as a *fruit*. There are many situations, other than collective naming, were visual context changes how a concept is described.

I propose a three step approach to automatically determine the context dependent basic-level names of visual objects. First I use ImageNet [3], a large dataset of manually categorised images, to learn over 2600 visual concept detectors. I then fit a model for selecting a basic-level name for each of these concepts, while taking into account visual context, object importance and object appearance. Finally, I fine-tune the obtained descriptions for an image using language context – this is done via a model for ranking descriptions from different visual concepts.

The first step towards deciding on descriptive names for a visual scene is to identify concepts. I define my concept domain using the well-known ImageNet [3] database which illustrates thousands of concepts (otherwise known as synsets) with a few hundred images each. I capture visual semantics by learning a visual representation of each synset from ImageNet.

The second step is to identify a set of descriptive names which can be used to describe a visual concept. In this work, I obtain a set of descriptive names for each concept by searching up the WordNet hypernym hierarchy (i.e., expanding parent concepts). Next I learn a classifier for each concept, to infer the most likely name given the visual context. I use Convolution Neural Network features as visual context and linear classifiers to select the most likely name. The ground-truth is constructed by matching the set of possible descriptive names to human generated captions.

I apply a linear ranking function based on the RankSVM formulation to rank descriptive names using four different types of features which capture both visual saliency and linguistic context. The most important of which are the *Score* features which include visual classifier confidences and the *Word2Vec* features that capture the word context.

3.2 Generating Descriptions with Sentiment

Original: a cat on a desk next to a laptop
Sentiment: a happy cat sleeping on a desk next to an open laptop

Figure 2: An example where positive sentiment is added to a descriptive caption.

Generating captions which exhibit a strong positive or negative sentiment is challenging because it combines the difficulties of caption generation with the added challenge of incorporating sentiment in very short texts. In light of this I have based my system on a state-of-the-art generative deep network model. This deep network, by itself, does not produce captions with a strong sentiment.

The first step towards generating descriptions with strong sentiment is to construct a dataset of human generated captions with strong sentiment. Since sentiment can be difficult to impose or even identify in short texts, I focus on sentences with sentimental adjective noun pairs. Mechanical Turk workers are asked to write image captions containing one of several adjective noun pairs that have been identified as having a strong sentiment. The resulting dataset consists of captions that are both visually relevant and exhibit a strong sentiment.

The primary problem with training a generative deep network on the sentiment dataset is controlling over-fitting. In particular I want to retain the general properties of the generative deep network trained on the purely descriptive sentences, but introduce the most important aspects of the sentiment dataset. To this end I train a new model that dynamically weights the contribution of the pre-trained network and a new network. This new network incorporates additional features such as a sentiment alignment objective and domain specific visual features.

4. RESULTS TO DATE

Here I present a summary of my previously published results on selecting the most appropriate name for a visual concept.

I evaluated my method, which takes into account visual context, against a baseline that always selects the most common name for a visual concept. In Figure 1 I show the results for cross-validation on a dataset consisting of 760000 image caption pairs collected from Flickr [18]. Among the 2,633 visual concepts my method improves upon the *Most frequent name* baseline for 1,222 concepts, among which 783 improved by more than 1%. No change in accuracy is measured for 1190, and 221 concepts exhibited a small accuracy decrease. I found that my method provides the most improvement for visual concepts with ambiguous basic-level names – two or more names used with similar frequency. The 221 concepts that exhibit an accuracy decrease are characterised by ambiguous basic-level names and fewer than average training examples.

My system produces natural names for hundreds of concepts by taking visual context into account. This highlights visual context as a feature for inclusion in visual naming systems and verifies the arguments laid out in the psychology literature on a scale that is difficult to achieve with human subjects. My approach is targeted towards caption generation where it can be used to impose human like naming conventions. This will improve the overall naturalness and fluency of captions generated by existing systems.

5. SUMMARY AND OUTLOOK

In my current and past research I have been working towards generating captions which optimise a specific style attribute – either object naming style or positive/negative sentiment. A set of discreet attributes, such as these, is a clear way of defining the style for the generated caption. It is also a very useful method when there is a style attribute that you know you want to optimise for.

5.1 Imitating Captioning Style

Another method for specifying the style of a caption is to provide examples of the style of writing you would like to produce, and then have the system imitate. For example, given a sample of captions written by an individual author the aim would be to generate captions in that authors style. Such a system would need to learn how to represent style compactly and then be able to generate captions from that representation. The system would also need to make use of more general information about caption construction. To train this system I would start by trying to build a style representation that is similar for captions generated by the same author, but different for captions generated by separate authors. Existing datasets can be used for this provided authorship information can be obtained.

5.2 Generating a Diverse Set of Captions

Current state of the art caption generation systems are designed to produce a single caption for an image. In reality there may be many different ways to describe the same image, which are all of a similar quality. A model which is able to enumerate a diverse selection of these sentences could be very useful. It would effectively allow the user to choose the style that they wish to use for a particular caption. In addition, using multiple different ways to describe a scene could be a more effective way to communicate than a single caption. The main challenge would be maintaining visual accuracy while increasing caption diversity.

6. REFERENCES

[1] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. *ACMMM*, 2013.

[2] S. E. Chaigneau, L. W. Barsalou, and M. Zamani. Situational information contributes to object categorization and inference. *Acta psychologica*, 2009.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009.

[4] J. Deng, J. Krause, a. C. Berg, and L. Fei-Fei. Hedging your bets: Optimizing accuracy-specificity trade-offs in large scale visual recognition. *CVPR*, 2012.

[5] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. *CoRR*, 2014.

[6] A. Esuli, F. Sebastiani, and V. G. Moruzzi. SENTIWORDNET : A Publicly Available Lexical Resource for Opinion Mining. *LREC*, 2006.

[7] L. Gatti, M. Guerini, O. Stock, and C. Strapparava. Sentiment Variations in Text for Persuasion Technology. *Persuasive Technology*, 2014.

[8] M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: data, models and evaluation metrics. *JAIR*, 2013.

[9] D. Joshi, R. Datta, E. Fedorovskaya, Q.-T. Luong, J. Wang, J. Li, and J. Luo. Aesthetics and emotions in images. *Signal Processing Magazine, IEEE*, 2011.

[10] A. Karpathy, A. Joulin, and F. F. F. Li. Deep fragment embeddings for bidirectional image sentence mapping. *NIPS*, 2014.

[11] R. Kiros, R. Salakhutdinov, and R. Zemel. Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models. *arXiv*, 2014.

[12] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. *NIPS*, 2012.

[13] G. Kulkarni, V. Premraj, and S. Dhar. Baby talk: Understanding and generating simple image descriptions. *CVPR*, 2011.

[14] P. Kuznetsova, V. Ordonez, and A. Berg. Collective generation of natural image descriptions. *ACL*, 2012.

[15] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille. Explain Images with Multimodal Recurrent Neural Networks. *NIPS*, 2015.

[16] A. Mathews, L. Xie, and X. He. Choosing basic-level concept names using visual and language context. *WACV*, 2015.

[17] V. Ordonez, J. Deng, and Y. Choi. From large scale image categorization to entry-level categories. *ICCV*, 2013.

[18] V. Ordonez, G. Kulkarni, and T. Berg. Im2text: Describing images using 1 million captioned photographs. *NIPS*, 2011.

[19] G. Özbal, D. Pighin, and C. Strapparava. BRAINSUP: Brainstorming Support for Creative Sentence Generation. *ACL*, 2013.

[20] M. Rohrbach, W. Qiu, I. Titov, and S. Thater. Translating Video Content to Natural Language Descriptions. *ICCV*, 2013.

[21] E. Rosch. *Principles of categorization*. 1999.

[22] E. Rosch, C. Mervis, and W. Gray. Basic objects in natural categories. *Cognitive Psychology*, 1976.

[23] F. D. Rosis and F. Grasso. Affective natural language generation. *Affective interactions*, 2000.

[24] R. Socher, A. Perelygin, and J. Wu. Recursive deep models for semantic compositionality over a sentiment treebank. *EMNLP*, 2013.

[25] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas. Sentiment strength detection in short informal text. *JASIST*, 2010.

[26] O. Vinyals and A. Toshev. Show and Tell: A Neural Image Caption Generator. *arXiv*, 2014.

[27] Y. Yang, C. Teo, H. D. III, and Y. Aloimonos. Corpus-guided sentence generation of natural images. *EMNLP*, 2011.