Detecting Visually Relevant Sentences for Fine-Grained Classification

Olivia Winn*, **Madhavan Kavanur Kidambi*** **and Smaranda Muresan**†
*Computer Science Department, Columbia University
† Center for Computational Learning Systems, Columbia University
olivia@cs.columbia.edu, mk3700@columbia.edu, smara@columbia.edu

Abstract

Detecting discriminative semantic attributes from text which correlate with image features is one of the main challenges of zero-shot learning for fine-grained image classification. Particularly, using full-length encyclopedic articles as textual descriptions has had limited success, one reason being that such documents contain many non-visual or unrelated sentences. We propose a method to automatically extract visually relevant sentences from Wikipedia documents. Our model, based on a convolutional neural network, is robustly tested through ground truth labeling obtained via Amazon Mechanical Turk, achieving 81.73% F1 measure.

1 Introduction

Current research in multimodal fusion and cross-modal mapping relies primarily on pre-aligned datasets of images and their short captions or tags, where the text is known to contain visually descriptive content directly related to its image (Baroni, 2016). These texts are usually manually collected, and restricted in length to words, phrases, and sentences. Using full-length documents such as Wikipedia articles would potentially allow automated access to already available rich descriptive content and would greatly aid the task of fine-grained classification across numerous domains, many of which have rich image datasets (such as birds (Welinder et al., 2010), flowers (Nilsback and Zisserman, 2008), aircraft (Maji et al., 2013), and dogs (Khosla et al., 2011).)

Unfortunately, most full-length documents contain predominantly non-visual text, making them noisy with respect to visual information and limiting the success of zero-shot learning techniques for fine-grained classification (Elhoseiny et al., 2013; Elhoseiny et al., 2015; Lei Ba et al., 2015). Furthermore, the visual portion of the text often describes objects outside the classifier's interest, such as the color of a bird's eggs when the task is identifying bird species (see Figure 1).

...
Description
The Fish Crow is superficially similar to the American Crow but is smaller and has a more silky plumage by comparison... The eyes are dark brown…Fish crows tend to have more slender bills and feet. There may also be a small sharp hook at the end of the upper bill.. […]
Diet
Food is taken mainly from the ground and even in shallow water where the bird will hover and pluck food items out of the water with its feet. The fish crow is omnivorous. […]
Breeding
The nest is usually built high in a tree and is often accompanied in nearby trees with other nests of the same species forming small, loose colonies. There are usually 4-5 eggs laid. Pale blue-green in colour, they bear blotches of olive-brown.[…]
Conservation
This species appears to be somewhat more resistant to West Nile Virus than the American crow. […]

Figure 1: Example Sentences from Wikipedia article on the Fish Crow.

Thus, the question we address in this paper is as follows: can we *automatically identify visually descriptive sentences relevant to a particular object* from documents that may contain predominantly non-visual text? We refer to this type of sentence as 'visually relevant'. Answering this question would allow us to automatically build aligned datasets of images with rich sentence-level descriptions, removing the necessity of manually creating aligned image-text datasets.

In this work, we focus on bird species, as this is one of the most well-studied and challenging fine-grained classification domains, using Wikipedia articles as our text (Section 2). To build our computational models, we must first define the notion of 'visually relevant' sentences. We use the defini-

Proceedings of the 5th Workshop on Vision and Language, pages 86–91,
Berlin, Germany, August 12 2016.

tion of Visually Descriptive Language (VDL) introduced by Gaizauskas et al. (2015), with some restrictions. Like VDL, we aim to identify 'visually confirmed' rather than 'visually concrete' segments of text as our descriptions correspond to a class (the bird species) rather than a particular image. For example, a sentence describing a bird's feet can be a 'visually relevant' sentence for a bird, though it would not be 'visually concrete' for an image of the bird flying with its feet hidden. Unlike VDL, for the scope of this paper we are interested only in the sentences which are *visually descriptive with respect to the object (i.e., bird species)*. We define such sentences as containing *visually relevant language* (VRL).

To build our training data, we make a simplifying assumption: a sentence is only considered to contain visually relevant language if it is in the 'Description' section of the article. While other sections may contain visually descriptive language, we assume they describe other objects such as the eggs. This simplifying assumption allows us to approach our problem as a *sentence classification task* (is a sentence VRL or non-VRL), and provides an automatic, though noisy, approach for labeling the training data. We collect a dataset of 1150 Wikipedia articles about birds to train the non-linear, non-consecutive convolution neural network architecture proposed by Lei et al. (2015). The architecture of this particular CNN is well suited to model sentences in our corpus such as *"Adults have upperparts streaked with brown, grey, black and white"* as it captures non-consecutive grams such as *"upperparts brown"*, *"upperparts gray"*, *"streaked white"*, etc.

To test our model in a robust manner, we use crowdsourcing to manually annotate all sentences as either VRL or non-VRL from an unseen set of 200 Wikipedia articles (for a total of 6342 sentences) (Section 2), corresponding to the bird classes in the Caltech-UCSD Birds-200-2011 dataset (Welinder et al., 2010).

Our experiments show that the CNN model trained on the noisy VRL dataset performs very well when tested on a human-labeled VRL dataset: 83.4% Precision, 80.13% Recall, 81.73% F1 measure (Section 4). Our analysis highlights several findings: 1) VRL sentences outside of the description section, or in documents with no Description section, are properly labeled by the model as VRL; 2) non-VRL sentences within the Description section (many documents included descriptions of birdsong in these sections) are correctly labeled by the model as non-VRL (Section 4). The datasets, including the crowdsourcing annotations for the 200 documents are released to the research community (`http://github.com/oh-livia/VRL_Wiki_Dataset`). This dataset will be useful to advance research on fine-grained classification, given that the Caltech-UCSD Birds-200-2011 is one of the most highly used datasets for this task.

	Training	**Development**
VRL	6355	794
non-VRL	27292	3411
Total	33647	4205

Table 1: Statistics of the Training and Dev. Sets

2 Datasets

To train our models we collected a set of 1150 Wikipedia articles of bird species. As a future goal of this work is to correlate the extracted textual information with image data, the training documents were specifically chosen not to correspond to the 200 birds species in the Caltech-UCSD Birds-200-11 dataset, which were set aside as test data. Of these 1150 documents, 690 of them contained sections labeled "Description" or related headings such as "Appearance", which allowed us to build our training and development sets. All sentences in the sections labeled "Description", "Appearance" and "Identification" were considered instances of the VRL class and everything else as instances of the non-VRL class; this labeling scheme we refer to as 'noisy'. Table 1 shows the statistics of the number of training and development instances used to build the computational models. The dataset is highly unbalanced: VRL sentences comprise 19% of both training and development. This skew is typical of many descriptive documents, and as such provides an appropriate model to train on.

To test our models we use the Wikipedia articles of the 200 birds in the Caltech-UCSD Birds-200-11 previously collected by Elhoseiny et al. (2013), consisting of 6342 sentences, which we call $\mathbf{200}_{VRL}$. To see whether our computational models trained on the noisy VRL dataset are able to detect VRL sentences as judged by humans, we conducted a crowdsourcing experiment.

2.1 Crowdsourcing to Annotate Sentences as Visually Relevant

We define a sentence-level annotation task, where each sentence in a document is assigned one of the following labels: **1** — the sentence contains visually relevant language (VRL), i.e. it is visually descriptive with respect to the object under consideration (birds species) (see examples (1) and (2)); and **0** — the sentence does not contain visually relevant language (see examples (3), (4), (5)).

Label **1** (VRL sentence) is assigned when the entire sentence is visually relevant (ex (1)) or when it is partially visually relevant (e.g., in example (2) only the underlined part is visually relevant):

(1) It has a black cap and a chestnut lower belly

(2) Males give increasingly vocal displays and show off the white markings of the wings in flight and of the tail [...]

Label **0** (non-VRL sentence) is assigned when the sentence describes the object of interest (bird species) but it is not visually descriptive (ex (3)), when it is visually descriptive but not relevant to the object (ex (4)), or when it is neither visually descriptive nor associated with the bird species.

(3) Males have 2 distinct types of songs - classified as short and long songs.

(4) The egg coloring is a brown spotted greenish-white.

(5) Finally volcanic eruptions on Torishima continues to be a threat.

In addition to the above labeling, for cases where a Turker chose the label **1** they were asked to provide information about the particular visually relevant text segments by specifying the *bird*, the *body part* and the *description*. While these phrase-level annotations are not used for our current task, they could be used in future work when joint-learning from text and images, especially to align information related to each body part of the bird. In addition, they could be used to build a graph-based representation of image descriptions similar to scene graphs (Schuster et al., 2015).

The annotation task was done at the sentence level and each sentence was annotated by three Turkers on Amazon Mechanical Turk. Besides the two labels **1** and **0**, the Turkers could also select "I don't know" and provide an explanation for why they could not determine whether or not the sentence contains VRL. We used highly skilled Turkers ($\geq$ 500 completed HITS and $\geq$ 95% approval rate) and we paid 5 cents per HIT (each HIT contained only one sentence). The inter-annotator agreement was very high, with a Fleiss $\mathcal{K}$ score of 0.8273. Only 8.64% of the sentences did not have a unanimous vote. Less than 2% of the sentences had at least one Turker vote 'I don't know'; of these, less than 0.05% garnered one vote each of **1**, **0** and 'I don't know'.

To build the test set for the computational models we use majority voting (at least two annotators selected the label). For the few cases where we did not have majority voting (0.05% of data) we selected the **0** label, as only one Turker voted **1** while the other two said **0** and 'I don't know'. This test set, which we call $\mathbf{200}_{HumVRL}$, contains 1248 sentences of class **1** (VRL) and 5094 sentences of class **0** (non-VRL).

3 Detecting Visually Relevant Sentences

As mentioned earlier, our task can be framed as a binary sentence classification problem, where each sentence is labeled either as VRL or non-VRL. Deep learning methods, and in particular convolutional neural networks (CNNs), have become some of the top performing methods on various NLP tasks that can be modeled as sentence classification (e.g, sentiment analysis, question type classification) (Kim, 2014; Kalchbrenner et al., 2014; Lei et al., 2015).

We use the non-linear, non-consecutive convolution neural network architecture proposed by Lei et al. (2015), which we refer to as $\mathbf{CNN}_{\boldsymbol{Lei}}$. This CNN uses tensor products to combine non-consecutive n-grams of each sentence to create an embedding per sentence. The non-consecutive aspect of the n-gram allows it to capture co-occurrence of words spread across sentences: *"yellow crown, rump and flank patch"* will generate representations of the relevant noun-adjective pairs *"yellow crown"*, *"yellow rump"*, and *"yellow flank patch"*. The tensor product is used as a "generalized approach" to linear concatenation of the n-grams, as concatenation is "insufficient to directly capture relevant information in the n-gram" (Lei et al., 2015, p 1). We use the training and development set described in Table 1 that comes from the 690 documents with 'Description' headings.

Hyperparameters and Word Vectors. The word vectors are pre-trained on the entire set of 1150 Wikipedia articles about birds using the word2vec model of Mikolov et al. (2013) with a window context of 20 words and vectors of 150 dimensions. Notice that we do not use the documents in the test set $\mathbf{200}_{VRL}$ for training the word vectors. We chose to use domain specific text to pre-train the word vectors in order to make sure we are capturing domain specific semantics such as proper word senses. Words such as "crown", when trained on a different corpus, would typically have an embedding very close to words such as "royalty", "tiara", etc; in the domain of bird descriptions, "crown" maps most closely to "feathers" and "head". The hyperparameters for the CNN model are: L2 regularization weight is 0.0001, n-gram order is 3 and hidden feature dimension is 50.

4 Experimental Setup and Results

Test Datasets. We first evaluate the $\mathbf{CNN}_{\boldsymbol{Lei}}$ model on the $\mathbf{200}_{HumVRL}$ dataset described in Section 2, which contains the 6342 sentences labeled by Turkers (class distribution: 1248 sentences in class **1** and 5094 sentences in class **0**). Since our computational model was trained on the noisy visually relevant sentences (where the labels were determined by the 'Description' section of the documents), we wanted to evaluate how the model performed on a similarly constructed test set. Thus, instead of considering the human labels for the 6342 sentences, a sentence was assigned to class **1** if it belonged to the Description, Appearance or Identification sections and to class **0** otherwise. We call this dataset $\mathbf{200}_{NoisyVRL}$ (class distribution: 1258 sentences in class **1** and 5084 sentences in class **0**). Note that while it seems as if only 10 sentences changed, many of the sentences in the 'Description' sections were labeled by humans as class 0, and many sentences outside these sections labeled as class 1. However, one possible issue with the $\mathbf{200}_{NoisyVRL}$ dataset is that some documents do not contain any description-type sections and thus all sentences are labeled **0**, which might affect measuring the performance of the model. Thus, we considered additional test sets containing only the documents that had sections labeled with 'Description', 'Appearance' or 'Identification' (142 documents out of the original 200 documents). Using these documents, we constructed a dataset $\mathbf{142}_{NoisyVRL}$, where class **1** contained sentences that were part of the three description-type sections, and class **0** contained all other sentences (class distribution: 1156 class **1** and 3836 class **0**). In addition, we also used the Turkers' labels (majority voting) for the corresponding sentences in these 142 documents. We call this dataset $\mathbf{142}_{HumVRL}$ (class distribution: 992 class **1** and 4000 class **0**). Since the CNN model was trained on the noisy labeling, a reasonable assumption is that the classification results would be better on the $\mathbf{200}_{NoisyVRL}$ and $\mathbf{142}_{NoisyVRL}$ datasets than on the $\mathbf{200}_{HumVRL}$ and $\mathbf{142}_{HumVRL}$ datasets.

Baseline. As baseline, we used the same neural bag-of-words model (**nBoW**) as Lei et al. (2015). We use the same training and development sets as for the CNN model (Table 1), along with the same word embeddings.

Results and Discussion. Table 2 shows the results of the $\mathbf{CNN}_{\boldsymbol{Lei}}$ model and the **nBoW** model on the four datasets. The CNN model performs slightly better than the baseline on all datasets in terms of F1 measure, with a much better Recall but worse Precision. Given that the end goal is to use the extracted visually relevant sentences together with images for fine-grained classification, and that the amount of visually relevant sentences in a document is small with respect to the document length, having high Recall is important.

One of the most interesting findings of this study is that both of the computational models perform much better on the human-labeled visually relevant datasets ($\mathbf{200}_{HumVRL}$, $\mathbf{142}_{HumVRL}$) than on the noisy visually relevant datasets ($\mathbf{200}_{NoisyVRL}$, $\mathbf{142}_{NoisyVRL}$). In particular, the recall increases significantly (e.g., from 63.24% on $\mathbf{142}_{NoisyVRL}$ to 80.15% on $\mathbf{142}_{HumVRL}$ using the $\mathbf{CNN}_{\boldsymbol{Lei}}$ model).

An error analysis highlights that the computational models are more 'conservative' with the classification of VRL than the noisy labeling. As mentioned earlier, the Description sections of the Wikipedia articles often (though not always) contain details pertaining to the birds' song. However, despite being trained on such a labeling, the computational models do not classify most sentences related primarily to the description of birds' song as VRL. This result was most likely aided by the fact that some of the training documents contain

Models	$\mathbf{200}_{HumVRL}$			$\mathbf{200}_{NoisyVRL}$			$\mathbf{142}_{HumVRL}$			$\mathbf{142}_{NoisyVRL}$		
	P(%)	R(%)	F(%)	P(%)	R(%)	F(%)	P(%)	R(%)	F(%)	P(%)	R(%)	F(%)
$\mathbf{CNN}_{\boldsymbol{Lei}}$	83.40	80.13	81.73	66.06	62.96	64.47	82.94	80.15	81.52	82.04	63.24	71.42
nBoW	88.61	73.56	80.39	67.31	55.33	60.73	88.48	73.32	80.19	84.11	55.88	67.15

Table 2: Classification results on the four datasets

song descriptions outside of the description-type sections, so the words pertaining to sound were not correlated as strongly with the VRL class. It is also possible that the abundance of appearance descriptions in each description section would encourage the visual words to have a much stronger effect on the 'visualness' of a sentence. One such example is the sentence *"The song is a series of musical notes which sound like: wheeta wheeta whee-tee-oh, for which a common pneumonic is 'The red, the red T-shirt'"*. Even the repetition of the word *'red'* is not enough to make the classifier label the sentence as VRL.

Another type of example that explains these results are sentences that describe the weight of the birds, such as *"Recorded weights range from 0.69 to 2 kg,[...]"* These sentences were part of the Description section, but were not marked as VRL by either the Turkers or the computational models.

We also analyzed some of the false positives of the $\mathbf{CNN}_{\boldsymbol{Lei}}$ model on the $\mathbf{142}_{HumVRL}$ and $\mathbf{200}_{HumVRL}$ datasets. One type of error comes from sentences that are visually descriptive, but not visually relevant, such as sentences that describe other objects like eggs. For example, the sentence *"The egg shells are of various shades of light or bluish grey with irregular, dark brown spots or greyish-brown splotches"* was labeled as VRL by the model but not by the Turkers. More interesting are the false positives that contain comparison words such as *"clapping or clicking has been observed more often in females than in males"*, and words having to do with appearance that do not specifically describe how the bird looks such as *"this bird is more often seen than heard"*.

5 Related Work

There are two lines of work most closely related to ours. First, Gaizauskas et al. (2015) propose a definition and typology of Visually Descriptive Language (VDL). They show that humans are able to reliably annotate text segments as containing 'visually descriptive' language or not, providing evidence that standalone text can be classified by the visualness of its contents. In our work, motivated by the end task of fine-grained classification, we restrict the definition to 'visually relevant'. As Gaizauskas et al. (2015) do, we show that humans can reliably annotate text as visually relevant or not. Unlike Gaizauskas et al. (2015), we propose a method to automatically detect visually relevant sentences from full-text documents. Second, Dodge et al. (2012) propose a method to separate visual text from non-visual text in image captions. However, their method focuses just on noun-phrases, while our approach finds visually relevant sentences in full-length documents.

While our end result is a set of visually relevant text descriptions, our approach is complementary to the rich body of work on generating text descriptions from images (see (Bernardi et al., 2016) for a survey), since our method *extracts such descriptions from existing text*.

6 Conclusion

Our work shows that it is possible to take domain-specific full-length documents—such as Wikipedia articles for birds species—and classify their sentences by visual relevancy using a CNN model trained on a noisy dataset. As many documents generally have a small proportion of visually relevant sentences, this approach automatically generates high quality visually relevant textual descriptions for images to be used by zero-shot learning approaches for fine-grained image classification tasks (e.g., (Wang et al., 2009)). While our study has focused on bird species, we believe that this method is generally applicable for other domains used in fine-grained classification research such as flowers and dogs (all have associated Wikipedia articles and Description/Appearance sections). In future work, we plan to use the outcomes of this work for joint learning from text and images.

Acknowledgments

This research was funded by the NSF (award IIS-409257). We thank the anonymous reviewers for helpful feedback.

References

M. Baroni. 2016. Grounding distributional semantics in the visual world. *Language and Linguistics Compass*, 10(1):3–13.

Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research*, 55.

Jesse Dodge, Amit Goyal, Xufeng Han, Alyssa Mensch, Margaret Mitchell, Karl Stratos, Kota Yamaguchi, Yejin Choi, Hal Daumé III, Alexander C Berg, et al. 2012. Detecting visual text. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 762–772. Association for Computational Linguistics.

Mohamed Elhoseiny, Babak Saleh, and Ahmed Elgammal. 2013. Write a classifier: Zero-shot learning using purely textual descriptions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2584–2591.

Mohamed Elhoseiny, Ahmed Elgammal, and Babak Saleh. 2015. Tell and predict: Kernel classifier prediction for unseen visual classes from unstructured text descriptions. *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR)*.

Robert Gaizauskas, Josiah Wang, and Arnau Ramisa. 2015. Defining visually descriptive language. In *Proceedings of the 2015 Workshop on Vision and Language (VL15): Vision and Language Integration Meets Cognitive Systems*.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pages 655–665.

Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. 2011. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, June.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751.

Jimmy Lei Ba, Kevin Swersky, Sanja Fidler, et al. 2015. Predicting deep zero-shot convolutional neural networks using textual descriptions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4247–4255.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2015. Molding cnns for text: Non-linear, non-consecutive convolutions. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. 2013. Fine-grained visual classification of aircraft. Technical report.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in Neural infgormation Processing Systems*, pages 3111–3119.

M-E. Nilsback and A. Zisserman. 2008. Automated flower classification over a large number of classes. In *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing*, Dec.

Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D Manning. 2015. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the Fourth Workshop on Vision and Language*, pages 70–80.

Josiah Wang, Katja Markert, and Mark Everingham. 2009. Learning models for object recognition from natural language descriptions. In *British Machine Vision Conference (BMVC)*, volume 1, page 2.

P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. 2010. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology.