Learning Models for Object Recognition from Natural Language Descriptions

Josiah Wang
scs6jwks@comp.leeds.ac.uk
Katja Markert
markert@comp.leeds.ac.uk
Mark Everingham
me@comp.leeds.ac.uk

School of Computing
University of Leeds
Leeds, UK

Progress in machine learning approaches to object category recognition has been significant in recent years, yet scaling current methods to many object classes is limited by the onerous task of manually collecting and labelling large training sets. In this paper we propose methods for learning models for visual object recognition from textual natural language descriptions. For fine-grain categories such as animal or plant species, such descriptions are readily available in the form of online nature guides (Figure 1). By extracting salient visual properties from such descriptions, our approach enables learning of visual recognition with *no* training images.

We investigate the task of learning to recognise fine-grain object categories from natural language descriptions alone, using species of butterflies as an example. We learn models to recognise ten categories (species) of butterflies solely from textual descriptions obtained from the eNature online nature guide; *no* training images are used. The method comprises three components: (i) *natural language processing (NLP)* to build models from textual descriptions; (ii) *visual processing* to extract visual attributes from test images; and (iii) a *generative model* learnt from text connecting textual terms with visual attributes.

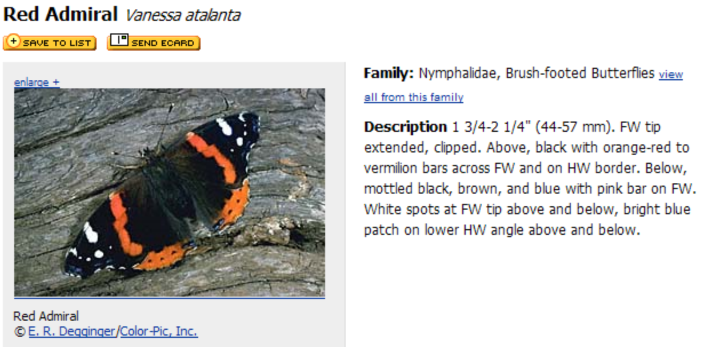
Natural language processing. Our NLP approach extracts models from the textual descriptions obtained from eNature (Figure 1). We treat the problem as an *information extraction* task, where unstructured information in text is converted into structured data in the form of a *template*. The input text is first divided into tokens, then a Part-of-Speech (PoS) tagger computes PoS tags for each token. The tags are modified by a list of rules to adapt to the specific style of the eNature descriptions. Chunking is then performed to extract noun phrases (NP) and adjective phrases (AP), for example “wing has blue spots” or “wings are black”. Finally a template is filled by matching the resulting ‘chunks’ against a list of colours, patterns and location terms.

Visual processing. Visual attributes of an image are extracted for matching against the models learnt from text. Our method bases recognition on two simple visual attributes determined salient from the textual descriptions: (i) dominant (wing) colour; (ii) coloured spots. To remove background clutter, images are first segmented using a graph-cut method with a ‘star shape’ prior. Colour models are learnt which relate pixel values to named colours *e.g.* ‘orange’. For each colour name, example pixels from unlabelled butterfly images are used to learn a Parzen model in $L^*a^*b^*$ colour space.

Spots are extracted by finding candidates using the Difference-of-Gaussians (DoG) interest point operator. Each candidate is described by SIFT descriptors extracted around the interest point, and a linear classifier labels candidates as either ‘spot’ or ‘non-spot’. The colour of each spot is estimated as the Gaussian-weighted spatial average of pixels within the spot region in $L^*a^*b^*$ colour space.

Generative model. The task of predicting the category of butterfly depicted in an input image is cast as one of Bayesian inference using a generative model for each of the ten butterfly categories. An input image I is classified by assigning it the category B_i which maximises the likelihood $p(I|B_i)$:

$$p(I|B_i) = \underbrace{\prod_j \sum_k p(\mathbf{z}^s_j|c^s_k)P(c^s_k|B_i)}_{\text{spot colours}} \; \underbrace{\prod_j \sum_k p(\mathbf{z}^w_j|c^w_k)P(c^w_k|B_i)}_{\text{dominant (wing) colour}}$$



Red Admiral *Vanessa atalanta*

Family: Nymphalidae, Brush-footed Butterflies view all from this family

Description 1 3/4-2 1/4" (44-57 mm). FW tip extended, clipped. Above, black with orange-red to vermilion bars across FW and on HW border. Below, mottled black, brown, and blue with pink bar on FW. White spots at FW tip above and below, bright blue patch on lower HW angle above and below.

Red Admiral
© E. R. Degginger/Color-Pic, Inc.

Figure 1: Example visual description from eNature, for the Red Admiral butterfly *Vanessa atalanta*. The question we investigate in this paper is whether a computer can learn to recognise this species of butterfly from the textual description alone, and indeed can humans?

- $\mathbf{z}^s_j$ is the observed $L^*a^*b^*$ colour of spot j
- $\mathbf{z}^w$ are the observed $L^*a^*b^*$ colours of non-spot (wing) pixels
- $p(\mathbf{z}|c)$ is the probability of observing pixel value $\mathbf{z}$ for colour name c
- $P(c^s_k|B_i)$ and $P(c^w_k|B_i)$ are learnt priors for category B_i, over spot and dominant (wing) colour names respectively.

The category-specific spot colour priors $P(c^s_k|B_i)$ are constructed from the learnt template by assigning equal probability to each colour in the template. For example, if the template contains ‘white spots’ and ‘black spots’, the probability of each is assigned 0.5. The dominant colour name priors $P(c^w_k|B_i)$ are defined as a mixture of two components:

$$P(c^w_k|B_i) = \alpha P(c^w_k|\Theta^d_i) + (1-\alpha)P(c^w_k|\Theta^o_i)$$

where $P(c^w_k|\Theta^d_i)$ and $P(c^w_k|\Theta^o_i)$ denote the prior over colour names for the dominant colour and ‘other’ (pattern) colours respectively. These are set uniformly for all corresponding colour names appearing in the template. For example, if the template contains a dominant colour of ‘orange’ and other colours ‘black’ and ‘white’, then $P(c^w_k|\Theta^d_i)$ is 1 for orange and zero for all other colours, and $P(c^w_k|\Theta^o_i)$ is 0.5 for black and white. The parameter α controls how much of the image is expected to be explained by the dominant colour. Rather than setting this to an arbitrary value we define a Beta hyper-prior over its value, and marginalise.

Results. The paper reports results of two sets of experiments, measuring performance of humans (as an ‘upper-bound’) and the proposed method. A dataset of 832 images of ten butterfly categories (species) with associated descriptions is used. The task of learning to recognise butterflies from text alone proves challenging for humans, with native and non-native English speakers achieving accuracy of 72% and 51% respectively. Our proposed method achieves 54% accuracy, substantially better than chance (10%), and slightly outperforming the non-native English speakers!

Conclusions. Our work proposes new approaches for exploiting NLP methods to learn object recognition without example images, and has potential for expanding current object recognition to fine-grain categories where it is difficult to find many training images. Further discussion can be found in the paper.