Learning Models for Object Recognition from Natural Language Descriptions

Josiah Wang
Katja Markert
Mark Everingham

School of Computing
University of Leeds
Main Idea

• Learn models using only textual descriptions
  – No training images used

(Source: eNature.com)
Main Idea

• Conventional approaches require many training images
  – Difficult to scale to large number of categories

• Related work in CVPR 2009
  – Farhadi et al. (2009) & Lampert et al. (2009)
  – Describe object categories using named attributes
  – Attributes are defined by hand for categories
  – We learn these from textual descriptions
Textual Descriptions

• Define **appearance** properties of an object category

• Readily available for certain object categories (butterflies, flowers, sign language, judo moves, etc.)

Very large, with forewing long and drawn out. Above, **bright, burnt-orange** with **black veins** and **black margins** sprinkled with **white dots**; forewing **tip broadly black** interrupted by larger **white and orange spots**. Below, paler, duskier **orange**.
Challenges

• Mapping between text and images

• Extracting information from textual descriptions
  – Parsing

• Short descriptions

• Some described properties are not visible in images
Dataset

• Ten butterfly categories
• Training set: Textual descriptions *only* (from eNature)
• Test set: Butterfly images (from Google Images)
Method Outline

**training descriptions**

Above, bright, burnt-orange with black veins and black margins sprinkled with white dots; FW tip broadly black interrupted by larger white and orange spots.

**Natural Language Processing (NLP)**

Representation from text

**Visual Processing**

Representation from image

**Generative Model**

**test image**

Butterfly with bright orange wings and black veins, set against a green background.
Above, bright, burnt-orange with black veins and black margins sprinkled with white dots; FW tip broadly black interrupted by larger white and orange spots.
Above, bright, burnt-orange with black veins and black margins sprinkled with white dots; forewing tip broadly black interrupted by larger white and orange spots. Below, paler, duskier orange.

Natural Language Processing

above fw colour : orange
above fw pattern : [black] veins
above fwm colour : black
above fwm pattern : [white] dots
above hw colour : orange
above hw pattern : [black] veins
above hwm colour : black
above hwm pattern : [white] dots
below fw colour : orange
below fw pattern : [white and orange] spots
below fwm colour : black
below fwm pattern : [white] dots
below hw colour : orange
below hw pattern : [black] veins
below hwm colour : black
below hwm pattern : [white] dots
Above, bright, burnt-orange with black veins and black margins sprinkled with white dots.

above fw colour: orange
above fw pattern: [black] veins
above fwm colour: black
above fwm pattern: white dots
above hw colour: orange
above hw pattern: [black] veins
above hwm colour: black
above hwm pattern: white dots
Above, bright, burnt-orange with black veins and black margins sprinkled with white dots.
above, bright, burnt-orange with black veins and black margins sprinkled with white dots.
above, bright, burnt-orange with black veins and black margins sprinkled with white dots.
above, bright, burnt-orange with black veins and black margins sprinkled with white dots.
Visual Processing

**training descriptions**
Above, bright, burnt-orange with black veins and black margins sprinkled with white dots; FW tip broadly black interrupted by larger white and orange spots.

**Natural Language Processing (NLP)**

**Representation from text**

**Visual Processing**

**Representation from image**

**Generative Model**
Visual Processing

- Test image
- Segmentation
- Colour modelling
- Spot detection
- Dominant (wing) colour
- Coloured spots

Colour modelling:
- Pixel 1 colour
- Pixel 2 colour
- Pixel 3 colour
- Pixel 4 colour
- Pixel 5 colour
...

Coloured spots:
- Spot 1 colour
- Spot 2 colour
- Spot 3 colour
- Spot 4 colour
- Spot 5 colour
...

16
Visual Processing

• Segmentation
  – Interactive ‘star shape’ graph-cut (Veksler 2008)
    • User selects a centre for the butterfly
    • User may specify more foreground/background points
Visual Processing

• Colour modelling
  – Relate a colour name *e.g.* “orange” to $L*a*b*$ values
  – Learn Parzen density model from selected pixels in butterfly images (category labels are **not** used)
Visual Processing

- Spot detection

Candidate Spot Detection → Spot Classifier

DoG detector → SIFT descriptors

Average colour → Logistic regression
Generative Model

**training descriptions**
Above, bright, burnt-orange with black veins and black margins sprinkled with white dots; FW tip broadly black interrupted by larger white and orange spots.

**Natural Language Processing (NLP)**

**Visual Processing**

**Representation from text**

**Representation from image**

**Generative Model**
Generative Model

• Template $\rightarrow$ Spot Colour Name Prior

above fw colour : black
above fw pattern : [orange] bars
above fwm pattern : [white] spots
above hw colour : black
above hwm pattern : [blue] patch
above hwm pattern : [black] spots
Generative Model

- **Template** → Wing Colour Name Prior

<table>
<thead>
<tr>
<th>Feature</th>
<th>Colour/Pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>Above fw colour</td>
<td>black</td>
</tr>
<tr>
<td>Above fw pattern</td>
<td>[orange] bars</td>
</tr>
<tr>
<td>Above fwm pattern</td>
<td>[white] spots</td>
</tr>
<tr>
<td>Above hw colour</td>
<td>black</td>
</tr>
<tr>
<td>Above hwm pattern</td>
<td>[blue] patch</td>
</tr>
<tr>
<td>Above hwm pattern</td>
<td>[black] spots</td>
</tr>
</tbody>
</table>

Dominant Colour
Generative Model

- **Template → Wing Colour Name Prior**

<table>
<thead>
<tr>
<th>Template</th>
<th>Wing Colour Name</th>
<th>Prior</th>
</tr>
</thead>
<tbody>
<tr>
<td>above fw colour</td>
<td>: black</td>
<td></td>
</tr>
<tr>
<td>above fw pattern</td>
<td>: [orange] bars</td>
<td></td>
</tr>
<tr>
<td>above fwm pattern</td>
<td>: [white] spots</td>
<td></td>
</tr>
<tr>
<td>above hw colour</td>
<td>: black</td>
<td></td>
</tr>
<tr>
<td>above hwm pattern</td>
<td>: [blue] patch</td>
<td></td>
</tr>
<tr>
<td>above hwm pattern</td>
<td>: [black] spots</td>
<td></td>
</tr>
</tbody>
</table>

- **Dominant Colour**

- **‘Other’ Colour**
Generative Model

• **Template → Wing Colour Name Prior**

<table>
<thead>
<tr>
<th>Above fw colour</th>
<th>above fw pattern</th>
<th>above fwm pattern</th>
<th>above hw colour</th>
<th>above hwm pattern</th>
<th>above hwm pattern</th>
</tr>
</thead>
</table>

\[ \alpha \text{ probability density} \]
Generative Model

\[ p(I|B_i) = p(S|B_i) \cdot p(W|B_i) \]

Classification: Assign to category which maximises \( p(I|B_i) \)
Humans
As “Upper Bound”

Proposed Method
How well can machine learn from only textual descriptions?

Butterfly Recognition Experiment

Please read the description below and select the butterfly that matches the description by clicking on the corresponding image. There is one (and only one) correct image. You can change your answer by clicking on another image. Please read the description carefully - you have only one chance!

When you have selected an image, please complete the information at the bottom of the page. If you are a native English speaker and/or an expert on butterflies, enter your email address, and click Submit.

Description

Above, bright, burnt-orange with black veins and black margins sprinkled with white dots; FW tip broadly black interrupted by larger white and orange spots.

Template

Classifier

?
Human Performance

Butterfly Recognition Experiment

Please read the description below and select the butterfly that matches the description by clicking on the corresponding image. There is one (and only one) correct image. You can change your answer by clicking on another image. Please read the description carefully - you have only one chance!

When you have selected an image, please complete the information at the bottom of the page - tell us if you are a native English speaker and/or an expert on butterflies, enter your email address, and click Submit.

Description

Wings long and narrow. Jet-black above, banded with lemon-yellow (sometimes pale yellow). Beneath similar; bases of wings have crimson spots.
Human Performance

Native speakers: 72% (201 participants)

Non-native speakers: 51% (52 participants)

Chance performance: 10%
Example Misclassification

_Heliconius charitonius_ (Zebra Longwing)

Wings long and narrow. Jet-black above, banded with lemon-yellow (sometimes pale yellow). Beneath similar; bases of wings have crimson spots. "lemon-yellow" bands? Spots are not mentioned.

Confused with
Results: Proposed Method

Humans vs. Our Method

- Chance: 10.0%
- Non-natives: 51.0%
- Natives: 72.0%
- Our Method: 54.4%

Individual vs. Combined Components

- Dominant colours: 35.3%
- Spot colours: 39.1%
- Both: 54.4%
### Results: Proposed Method

<table>
<thead>
<tr>
<th>Ground truth</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>danaus plexippus</td>
<td>35 0 4 18 41 1</td>
</tr>
<tr>
<td>heliconius charitoni</td>
<td>0 1 1 42 55 1</td>
</tr>
<tr>
<td>heliconius erato</td>
<td>49 10 7 2 33 1</td>
</tr>
<tr>
<td>junonia coenia</td>
<td>1 2 42 20 12 20</td>
</tr>
<tr>
<td>lycaena phlaeas</td>
<td>3 94 2 61 12 20</td>
</tr>
<tr>
<td>nymphalis antiopa</td>
<td>1 1 61 3 12 20</td>
</tr>
<tr>
<td>papilio cresphontes</td>
<td>1 1 94 2 61 12</td>
</tr>
<tr>
<td>pieris rapae</td>
<td>2 2 94 18 12 20</td>
</tr>
<tr>
<td>vanessa atalanta</td>
<td>6 15 18 94 12 20</td>
</tr>
<tr>
<td>vanessa cardui</td>
<td>39 39 77 55 89 100</td>
</tr>
</tbody>
</table>

**Accuracy:** 54.4%
### Results: Proposed Method

<table>
<thead>
<tr>
<th>Ground truth</th>
<th>Predicted</th>
<th>Accuracy: 54.4%</th>
</tr>
</thead>
<tbody>
<tr>
<td>danaus plexippus</td>
<td>danaus plexippus</td>
<td>35</td>
</tr>
<tr>
<td>heliconius charitonius</td>
<td>heliconius charitonius</td>
<td>0</td>
</tr>
<tr>
<td>heliconius erato</td>
<td>heliconius erato</td>
<td>49</td>
</tr>
<tr>
<td>junonia coenia</td>
<td>junonia coenia</td>
<td>2</td>
</tr>
<tr>
<td>lycaena phleas</td>
<td>lycaena phleas</td>
<td>94</td>
</tr>
<tr>
<td>nymphalis antiopa</td>
<td>nymphalis antiopa</td>
<td>1</td>
</tr>
<tr>
<td>papilio cresphontes</td>
<td>papilio cresphontes</td>
<td>85</td>
</tr>
<tr>
<td>pieris rapae</td>
<td>pieris rapae</td>
<td>7</td>
</tr>
<tr>
<td>vanessa atalanta</td>
<td>vanessa atalanta</td>
<td>84</td>
</tr>
<tr>
<td>vanessa cardui</td>
<td>vanessa cardui</td>
<td>6</td>
</tr>
</tbody>
</table>

Accuracy: 54.4%
Proposed method is compared against two standard approaches:

- **Spatial-Colour Histograms**: L*a*b* colour space (8 bins per channel), Nearest neighbour classifier ($\chi^2$ distance)

- **Bag of Words**:
  - Feature Extraction
  - Vector Quantisation
  - Histogram
  - Classifier
  - DoG + SIFT (segmented)
  - $k$-means (10,000 clusters)
  - Continuous valued
  - SVM: $\chi^2$ kernel
Results: Standard Vision Methods

Bag of Words
79.7 ± 5.9%
(1 training image per category)

Spatial Colour Histograms
54.7 ± 3.3%
(5 training images per category)

Our method
54.5 ± 0.9%
(No training images)
Discussion

• We investigated models for linking information in text and images together
• Mapping between textual and image features is a challenging problem
• Initial model achieved modest accuracy with no training images
• State of the art vision methods give good results but depend on the training images used
• Future work:
  – Extract more information from text
  – Combine information from multiple texts
  – **Combine text with images**
Learning Models for Object Recognition from Natural Language Descriptions

Josiah Wang
Katja Markert
Mark Everingham

School of Computing
University of Leeds