Words and Pictures: more structure

D.A. Forsyth, UIUC
with
Okan Arikan (UT Austin), Nazli Ikizler (Boston U), Leslie Ikemoto (animate-me), Derek Hoiem (UIUC), Ali Farhadi (UIUC), Ian Endres (UIUC), Ryan White (Euclid media)
Obtain dataset

Build features

Light entertainment (the way we do it)

Mess around with classifiers, probability, etc

Produce representation
Big questions

- What signal representation should we use for recognition?

- What should we say about what we see?
Structure

- **Correlated words**
  - waves go with beaches not cats
- **Attributes**
  - has nose
- **Adjectives**
  - green hat
- **Relations**
  - cat on mat
- **Sentences**
  - A dolphin holds a basketball as it swims on its back
  - A dolphin swimming upside down while holding a basketball
  - A dolphin swims upside down holding a basketball between it’s flippers.
  - A seal floats on it’s back in the water, holding a basketball.
  - The dolphin on his back holds the orange basketball.
Correlated Words

- **Simple method:**
  - rack up some features, build a bunch of linear classifiers one per word
  - works poorly
    - few examples per word
    - many features, only some are stable

Learn this

\[ D \approx MX \]

Word data (observed)  Image representation (observed)
Correlated words

- Idea
  - some features are not helpful
  - a low dimensional subspace is good at predicting most things (Ando + Zhang, )
  - We can find this space by penalizing rank in the matrix of linear classifiers

Learn this

\[ D \approx GFX \]

Word data (observed) \hspace{1cm} Image representation (observed)
It was there and we didn’t

It was there and we predicted it  

It wasn’t and we did
Scenes as object bags

- We could build collections of labelled scene images
  - useful, but..
    - kitchen, bathroom, outdoor, and then?

- We could collect images of similar appearance
  - but...
    - might not really have similar objects in them

- Unsupervised bag discovery
  - Pictures of the same scene tend to contain similar objects
    - i.e. tend to attract the same image annotations
In this space, images are “close” if they “look similar” AND if they predict “similar” words.
Correlated word predictors are quite good

<table>
<thead>
<tr>
<th>Method</th>
<th>P</th>
<th>R</th>
<th>F1</th>
<th>Ref</th>
</tr>
</thead>
<tbody>
<tr>
<td>Co-occ</td>
<td>0.03</td>
<td>0.02</td>
<td>0.02</td>
<td>[53]</td>
</tr>
<tr>
<td>Trans</td>
<td>0.06</td>
<td>0.04</td>
<td>0.05</td>
<td>[27]</td>
</tr>
<tr>
<td>CMRM</td>
<td>0.10</td>
<td>0.09</td>
<td>0.10</td>
<td>[37]</td>
</tr>
<tr>
<td>TSIS</td>
<td>0.10</td>
<td>0.09</td>
<td>0.10</td>
<td>[19]</td>
</tr>
<tr>
<td>MaxEnt</td>
<td>0.09</td>
<td>0.12</td>
<td>0.10</td>
<td>[39]</td>
</tr>
<tr>
<td>CRM</td>
<td>0.16</td>
<td>0.19</td>
<td>0.17</td>
<td>[44]</td>
</tr>
<tr>
<td>CT-3×3</td>
<td>0.18</td>
<td>0.21</td>
<td>0.19</td>
<td>[82]</td>
</tr>
<tr>
<td>CRM-rect</td>
<td>0.22</td>
<td>0.23</td>
<td>0.23</td>
<td>[31]</td>
</tr>
<tr>
<td>InfNet</td>
<td>0.17</td>
<td>0.24</td>
<td>0.23</td>
<td>[50]</td>
</tr>
<tr>
<td>MBRM</td>
<td>0.24</td>
<td>0.25</td>
<td>0.25</td>
<td>[31]</td>
</tr>
<tr>
<td>MixHier</td>
<td>0.23</td>
<td>0.29</td>
<td>0.26</td>
<td>[17]</td>
</tr>
<tr>
<td>(section 2.2)</td>
<td>0.27</td>
<td>0.27</td>
<td>0.27</td>
<td></td>
</tr>
<tr>
<td>(section 2.2, kernel)</td>
<td>0.29</td>
<td>0.29</td>
<td>0.29</td>
<td></td>
</tr>
<tr>
<td>PicSOM</td>
<td>0.35*</td>
<td>0.35*</td>
<td>0.35*</td>
<td>[73]</td>
</tr>
</tbody>
</table>
Structure

- **Correlated words**
  - waves go with beaches not cats

- **Attributes**
  - has nose

- **Adjectives**
  - green hat

- **Relations**
  - cat on mat

- **Sentences**
  - A dolphin holds a basketball as it swims on its back
  - A **dolphin** holds a **basketball** as it swims on its **back**.
  - A **dolphin** swimming upside down while holding a **basketball**.
  - A **dolphin** swims upside down holding a **basketball** between its **flippers**.
  - A **seal** floats on its back in **water**, holding a **basketball**.
  - The **dolphin** on his **back** holds the **orange** **basketball**.
Object recognition = k class classification

• current data sets ok,
  • improve coverage

• research agenda:
  • more features, better classifiers:
  • perhaps category hierarchies for statistical leverage (tying)
Are these monkeys?
Object recognition = describing what objects are like

- current datasets
  - are largely of the wrong form
    - and no declarative data about objects
- research agenda
  - learning by reading
  - sensible responses to objects of unknown category
  - within class variance has semantics
  - architectures, representations, semantics
General architecture

Feature extraction → Feature Selection

Attribute Predictions

Category Models

Bird

Has Beak, Has Eye, Has foot, Has Feather

Farhadi et al 09; cf Lampert et al 09
Farhadi et al 09; cf Lampert et al 09
How is an object different from typical?

- Pragmatics suggests this is how adjectives are chosen
  - If we are sure it’s a cat, and we know that
    - an attribute is different from normal
    - the detector is usually reliable
  - we should report the missing/extra attribute
Missing attributes

Aeroplane
No “wing”

Car
No “window”

Boat
No “sail”

Aeroplane
No “jet engine”

Motorbike
No “side mirror”

Car
No “door”

Bicycle
No “wheel”

Sheep
No “wool”

Train
No “window”

Sofa
No “wood”

Bird
No “tail”

Bird
No “leg”

Bus
No “door”
Extra attributes

- Bird: "Leaf"
- Bus: "face"
- Motorbike: "cloth"
- Dining Table: "skin"
- People: "Furn. back"
- Aeroplane: "beak"
- People: "label"
- Sofa: "wheel"
- Bike: "Horn"
- Monitor: window"
Structure

- **Correlated words**
  - waves go with beaches not cats

- **Attributes**
  - has nose

- **Adjectives**
  - green hat

- **Relations**
  - cat on mat

- **Sentences**
  - A dolphin holds a basketball as it swims on its back.
  - A dolphin swimming upside down while holding a basketball.
  - A dolphin swims upside down holding a basketball between it’s flippers.
  - A seal floats on its back in the water, holding a basketball.
  - The dolphin on his back holds the orange basketball.
“Pink” from Google

Yanai Barnard 05
“Pink” after 10 EM iterations

Yanai Barnard 05
Partially supervised recognition

- **Training**
  - we know what is in the image, but not where

- **But if we have adjectives**
  - we can improve location estimates
  - and so recognition

Wang et al 09
Structure

- **Correlated words**
  - waves go with beaches not cats

- **Attributes**
  - has nose

- **Adjectives**
  - green hat

- **Relations**
  - cat on mat, but there is still a lot here

- **Sentences**
  - A dolphin holds a basketball as it swims on its back

A dolphin holds a basketball as it swims on its back.
A dolphin swimming upside down while holding a basketball.
A dolphin swims upside down holding a basketball between its flippers.
A seal floats on its back in the water, holding a basketball.
The dolphin on his back holds the orange basketball.
Relations as an MRF

Sun IN Sky
Sky ABOVE Grass
Sun ABOVE Grass

Gupta and Davis 08
Duygulu et al 02

Gupta and Davis 08
Relations distort participants
Relations distort participants
Relations distort participants
Structure

- Correlated words
  - waves go with beaches not cats
- Attributes
  - has nose
- Adjectives
  - green hat
- Relations
  - cat on mat
- Sentences
  - A dolphin holds a basketball as it swims on its back
    - A dolphin holds a basketball as it swims on its back.
    - A dolphin swimming upside down while holding a basketball.
    - A dolphin swims upside down holding a basketball between it's flippers.
    - A seal floats on it's back in the water, holding a basketball.
    - The dolphin on his back holds the orange basketball.
Two girls take a break to sit and talk.

Two women are sitting, and one of them is holding something.

Two women chatting while sitting outside.

Two women sitting on a bench talking.

Two women wearing jeans, one with a blue scarf around her head, sit and talk.
A crowd of young adults in a dark room.

A girl in a brown shirt and a blue jean skirt is dancing with a young man dressed in a blue shirt wearing a black backpack.

A group of people standing in a dark building.

A large group of people dancing in a bar.

Dancing at club and two guys bucking up.