Sentences and pictures: not just “more words” and pictures

D.A. Forsyth, UIUC
with
Julia Hockenmaier, Derek Hoiem, Ian Endres, Ali Farhadi,
Nicolas Loeff, Cyrus Rashtchian, Gang Wang
all of UIUC
with echoes from
Kobus Barnard (U. Arizona), Pinar Duygulu (Bilkent U.),
Nando de Freitas (UBC)
Words and pictures: Implicit csp
Words and pictures: Explicit csp

- In its simplest form, missing variable problem
- Pile in with EM
  - given correspondences, conditional probability table is easy (count)
  - given cpt, expected correspondences could be easy
- Caveats
  - might take a lot of data; symmetries, biases in data create issues

“the beautiful sun”

“le soleil beau”

Brown, Della Pietra, Della Pietra & Mercer 93; Melamed 01

“sun sea sky”

See Duygulu et al 02
How to generalize words and pictures?

- More accuracy
- More words
- Predict more structure
## Accuracy

<table>
<thead>
<tr>
<th>Method</th>
<th>P</th>
<th>R</th>
<th>Fl</th>
<th>Ref</th>
</tr>
</thead>
<tbody>
<tr>
<td>Co-occ</td>
<td>0.03</td>
<td>0.23</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>0.23</td>
<td>0.17</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>0.20</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>PicSOM</td>
<td>0.35*</td>
<td>0.35*</td>
<td>0.35*</td>
<td>[73]</td>
</tr>
</tbody>
</table>

Y. Mori et al. 99
Duygulu et al. 02
Jeon et al. 03
Celebi et al. 05
Jeon et al. 04
Lavrenko et al. 03
Yavlinsky et al. 05
Feng et al. 04
Metzler et al. 04
Feng et al. 04
Carneiro et al. 05

Viitaniemi et al. 07
More words

- **Easy case**
  - learn with larger vocabularies
  - tricky bits, but...

- **Hard case**
  - what do we do about out-of-example words?
  - one simple answer doesn’t work (later)
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 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) \quad 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

\[
D \approx GFX
\]

Learn this

Word data (observed)  Image representation (observed)
It was there and we didn’t

It was there and we predicted it  It wasn’t and we did
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>

Loeff Farhadi 08
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 **flipper**s.
A **seal** floats on its back in the **water**, holding a **basketball**.
The **dolphin** on his **back** holds the **orange** **basketball**
General architecture

Feature extraction

Feature Selection

Attribute Predictions

Attribute Classifiers

Category Models

Bird

Has Beak, Has Eye, Has foot, Has Feather

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 it’s 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
Wang et al 09

<table>
<thead>
<tr>
<th>Image</th>
<th>Object class Saliency map</th>
<th>Visual attribute Saliency map</th>
<th>Combined Saliency map</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Ground truth</th>
<th>Round 1</th>
<th>Round 2</th>
<th>Round 3</th>
</tr>
</thead>
</table>

Wang et al 09
Structure

- Correlated words
  - waves go with beaches not cats
- Modifiers
  - pink cadillac
- Attributes
  - has nose
- Modifier-noun pairs
  - green hat
- Relations
  - cat on mat
- Sentences
  - Two women wearing jeans, one with a blue scarf around her head, sit and talk.

Gupta and Davis 08, but there is still a lot here.
Structure

- Correlated words
  - waves go with beaches not cats
- Attributes
  - has nose
- Adjectives
  - green hat
- Relations
  - Gupta and Davis 08,
  - 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 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.
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.
Conclusions

- Real progress in accuracy
- Structure is still hard, but rewarding
- Big problem: predicting sentences