Language and Vision: Learning Knowledge about the World

Yejin Choi
Computer Science & Engineering

UNIVERSITY of WASHINGTON
Goal: Intelligent Communication
Intelligent Communication

Reading between the lines

Understanding what is said + what is not said
Blueberry Muffins

Ingredients
1 cup milk
1 egg
1/3 cup vegetable oil
2 cups all-purpose flour
2 teaspoons baking powder
1/2 cup white sugar
1/2 cup fresh blueberries

Procedure
1. Preheat oven to 400 degrees F. Line a 12-cup muffin tin with paper liners.
2. In a large bowl, stir together milk, egg, and oil. Add flour, baking powder, sugar, and blueberries; gently mix the batter with only a few strokes. Spoon batter into cups.

http://allrecipes.com/Recipe/Blueberry-Muffins-I/
Intelligent Communication

Reading between the lines

Understanding
  what is said
  +
  what is *not* said

Language is contextual:
  - social / emotional context
  - visual / physical context
HAL (A space odyssey, 1968)

- David Stork (HAL’s Legacy, 1998)

“Imagine, for example, a computer that could look at an arbitrary scene anything from a sunset over a fishing village to Grand Central Station at rush hour and produce a verbal description. This is a problem of overwhelming difficulty, relying as it does on finding solutions to both vision and language and then integrating them. I suspect that scene analysis will be one of the last cognitive tasks to be performed well by computers”
Welcome to Amazon.com Books!

One million titles, consistently low prices.

(If you explore just one thing, make it our personal notification service. We think it's very cool!)

**Spotlight! -- August 16th**

These are the books we love, offered at Amazon.com low prices. The spotlight moves EVERY day so please come often.

**One Million Titles**

Search Amazon.com's million title catalog by author, subject, title, keyword, and more... Or take a look at the books we recommend in over 20 categories... Check out our customer reviews and the award winners from the Hugo and Nebula to the Pulitzer and Nobel... and bestsellers are 30% off the publishers list...

**Eves & Editors, a personal notification service**

Like to know when that book you want comes out in paperback or when your favorite author has a new book? Sign up for our personal notification service and we'll let you know!
Web Today: Increasingly Visual
-- social media, news media, online shopping

- Facebook.com has over 250 billion images uploaded as of Jun 2013
- 1.15 billion users uploading 350 million images a day on average
“This picture shows one person, one grass, one chair, and one potted plant. The person is near the green grass, and in the chair. The green grass is by the chair, and near the potted plant.”

Conditional random fields (CRF) model to combine visual detection with language priors
A cow was staring at me in the grass in the countryside.
A cow in the grass was staring at me in the countryside.

Tree Structure --- Probabilistic Context Free Grammars (PCFG)

Object (NP)

```
NP
| DT a
| NP cow
```

Action (VP)

```
VP
| VBD was
| VBG staring
| IN at
| NP PRP me
```

Stuff (PP)

```
PP
| IN in
| NP
| | DT the
| | NP NN grass
```

Scene (PP)

```
PP
| IN in
| NP
| | DT the
| | NP NN countryside
```
Blue flowers have no scent. Small white flowers have no idea what they are.

Blue flowers are running rampant in my garden.

My cat laying in my duffel bag.
Deja Image-caption corpus (NAACL 2015):
  - Of 750 million pairs of image-caption pairs from Flickr
  - Retain only those captions that are repeated verbatim by more than one user
  - Yielding 4 million images with 180K unique captions

The sun sets for another day (12)
Sun is going to bed (21)
After the sun has set (9)
The sky looks like it is on fire (58)
Rippled sky (44)
Related Work

• Donahue et al., 2015, Vinyals et al, 2015, Fang et al., 2015, Karpahty et al, 2015, Xu et al, 2015, Delvin et al., 2015, …

• MS CoCo Dataset
  – 120,000 images, 5 captions per image
  – 80 objects

  – sports (10 categories):
    • tennis racket (3561 images), baseball bat, baseball gloves, snowboard, skateboard, surf board,…

  – street (5 categories)
    • traffic light (4330 images), fire hydrant (1797 images), stop sign (1803 images), parking meters (742 images), bench (5805 images)

  – person (6 categories)
    • tie (3955 images), umbrella (4142 images)

Data problem? Or Modeling problem?
Moving Forward ...

- Image captioning is an emblematic task, not the end goal
- Seeing beyond the literal content

- Why did this happen?
- How do they feel?
- Reasoning about the situation
- Need knowledge about the world
Learning Knowledge about the World

I: Size
II: Entailment
III: Cooking
IV: Event
Learning Knowledge about the World
Take I: Size

Bagherinezhad et al. @ AAAI 2016
Are Elephants Bigger than Butterflies?
Knowledge on Size Useful for

• Vision:
  – Prune out implausible detections
Knowledge on Size Useful for

• Vision:
  – Prune out implausible detections

• Language:
  – The trophy would not fit in the brown suitcase because it was too **big**. What was too **big**?
    Answer 0: the trophy
    Answer 1: the suitcase
Related Work

• Narisawa et al. 2013 -- Is a 204 cm Man Tall or Small?
• Tandon et al. 2014 -- WebChild
• Takamura et al. 2015

➢ Text only
Elephants Bigger than People?

• Reporting bias: do not state the obvious
• Use both language and images!
• Elephants bigger than butterflies?

→ Need multi-hop inference
Construction of size graph
• Not all object pairs co-occur in many images.
  • e.g. “airplane” and “watermelon”
• It is not scalable to see images for all object pairs.
• An edge (A,B) only if A and B co-occur in many images.
• 2 edge connected (2 disjoint edge paths between every pair)
Language – absolute estimation

- "car is * x * m"
- "person is * m tall"

Vision – relative estimation

- From Flickr images that are tagged with both objects
- LEVAN [CVPR14], a webly supervised object detector.
- Run a depth estimator to infer the object distances

\[
\frac{\text{size}(O_i)}{\text{size}(O_j)} = \frac{\text{area}(\text{box}_1)}{\text{area}(\text{box}_2)} \times \frac{\text{depth}(\text{box}_1)^2}{\text{depth}(\text{box}_2)^2}
\]
Collective Inference

- Resolving potential inconsistencies across different language and vision estimates
- Assumption: size follows log-normal distribution
- Size is always positive, thus log-normal instead of normal
- Also motivated by a psychology study (Konkle and Olivia 2011)

![Diagram of log-normal distribution curves with labels for μ = 0, σ = 1/4, μ = 0, σ = 1/2, μ = 0, σ = 1]
Collective Inference

- By optimizing LL over the entire graph (MLE)

\[
\sum_{(i,j) \in E} \sum_{r=1}^{n_{ij}} \log f(g_i - g_j = y_{ij}^{(r)} | g_i \sim N(\mu_i, \sigma_i^2), g_j \sim N(\mu_j, \sigma_j^2))
\]

\[\] + \sum_{i \in V} \sum_{r=1}^{n_i} \log f(g_i = y_i^{(r)} | g_i \sim N(\mu_i, \sigma_i^2))

- Coordinate ascent (not convex)
Final output: log-normal dist of sizes
Evaluation

Dataset: annotated labels for 41 physical objects with 486 comparisons.
Baselines

• Text Baseline (inspired by Davidov et al. ACL 2010): Search for some fixed templates and get the mean for each object.
  • e.g. “object is * x * m” and “object’s width is * m”

• Vision Baseline: To answer query (A < B) find a reliable path between A and B in the complete graph and multiply ratios.
Which of objects A or B is bigger?

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chance</td>
<td>50%</td>
</tr>
<tr>
<td>Text Only...</td>
<td>63%</td>
</tr>
<tr>
<td>Vision Only...</td>
<td>72%</td>
</tr>
<tr>
<td>Our Model</td>
<td>84%</td>
</tr>
<tr>
<td>Our Model (image only)</td>
<td>78%</td>
</tr>
<tr>
<td>Our Model (text only)</td>
<td>75%</td>
</tr>
</tbody>
</table>
To Conclude

- Learning size of objects
- Integrating language and vision
  - to overcome the reporting bias
- Future work: learning physical knowledge
Learning Knowledge about the World
Take II: Entailment

Izadinia et al. @ ICCV 2015
A horse is eating.
Is that horse standing or sitting?
Inspiration: Visual Dictionary

Walk

Trot

Gallop
Segment-Phrase Table:
Webly supervised over 50000 instances

<table>
<thead>
<tr>
<th>Chimpanzee running</th>
<th>Horse standing</th>
<th>Dog running</th>
<th>Person sitting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bear jumping</td>
<td>Cat eating</td>
<td>Bear lying</td>
<td>Horse eating</td>
</tr>
<tr>
<td>Dog sitting</td>
<td>Sheep lying</td>
<td>Cow fighting</td>
<td>Cat jumping</td>
</tr>
<tr>
<td>Chimpanzee sleeping</td>
<td>Person jumping</td>
<td>Bear standing up</td>
<td>Bird sitting</td>
</tr>
<tr>
<td>Bird sitting</td>
<td>Bear standing up</td>
<td>Sheep eating</td>
<td>Cow sleeping</td>
</tr>
<tr>
<td>Sheep eating</td>
<td>Chimpanzee running</td>
<td>Chimpanzee running</td>
<td>Bear stretching</td>
</tr>
</tbody>
</table>
A horse is eating. Is that horse standing or sitting?
a horse eating => a horse standing

- Reporting bias: do not state the obvious
- Another case where language + vision can help!
Entailment $X \Rightarrow Y$

- $T(Y)$
- $T(X)$
Entailment $X \Rightarrow Y$

$T(\text{horse standing})$

$T(\text{horse eating})$
Entailment $X \Rightarrow Y$

$entail(X \models Y) := Sim_{R2I}^>(X, Y) - Sim_{R2I}^<(Y, X)$

$Sim_{R2I}^>(X, Y) = \text{average asymmetric region-to-image similarity measure (Kim and Grauman 2010)} \text{ using top K segmentation masks}$
Global Inference

- Transitivity of entailment relations

\[
\max \sum_{x \neq y} \text{entail}_{xy} W_{xy} - \lambda |W| \quad s.t. \quad W_{xy} \in \{0, 1\}
\]

\[
\forall x, y, z \in \mathcal{V}, W_{xy} + W_{yz} - W_{xz} \leq 1
\]
Visual Semantic Tasks

1. Visual Entailment
Visual Semantic Tasks

1. Visual Entailment
2. Visual Paraphrasing
Visual Semantic Tasks

1. Visual Entailment
2. Visual Paraphrasing
3. Semantic Similarity
To Conclude

- **Segmant-Phrase Table**
  - Translation dictionary between images and text
- Can learn visual entailment and paraphrases
Learning Knowledge about the World
Take III: Cooking with Action Diagrams

Kiddon et al. @ EMNLP 2015
Interpreting Natural Language Instructions as Action Diagrams

Smart devices and personal robots executing commands in natural language instructions not just one line command, but a sequence of commands

Step 1: interpret instructions as action diagrams
Instructional Recipes

Blueberry Muffins

Ingredients
1 cup milk
1 egg
1/3 cup vegetable oil
2 cups all-purpose flour
2 teaspoons baking powder
1/2 cup white sugar
1/2 cup fresh blueberries

Procedure
1. Preheat oven to 400 degrees F. Line a 12-cup muffin tin with paper liners.
2. In a large bowl, stir together milk, egg, and oil. Add flour, baking powder, sugar, and blueberries; gently mix the batter with only a few strokes. Spoon batter into cups.
3. **Bake for 20 minutes.** Serve hot.

http://allrecipes.com/Recipe/Blueberry-Muffins-I/
From Kitchen to Biology Labs

DNA Precipitation

Materials
3M NaOAc pH 5.2
EtOH 95%
Glycogen (optional)

Procedure
1. Add 0.1 volumes of 3M Sodium Acetate solution to 1 volume of DNA sample.
2. Add 1ul Glycogen to the DNA sample.
3. Add 2 volumes of 95% EtOH to the DNA Sample.
4. Store the solution overnight at -20°C or for 30 minutes at -80°C.
5. Centrifuge the solution at maximum speed for least 15 minutes.
6. Decant and discard the supernatant.
7. (Optional) Add 1 ml of 70% EtOH to the pellet and let sit for 5 minutes.
8. (Optional) Centrifuge the sample at maximum speed for 5 minutes.
9. (Optional) Decant and Discard the supernatant.
10. Air-dry the pellet for 10-15 minutes at room temperature until all liquid is gone.
11. Resuspend in desired volume of water or buffer

http://openwetware.org/wiki/DNA_Precipitation
Blueberry Muffins

Ingredients
1 cup milk
1 egg
1/3 cup vegetable oil
2 cups all-purpose flour
2 teaspoons baking powder
1/2 cup white sugar
1/2 cup fresh blueberries

Procedure
1. Preheat oven to 400 degrees F (205 degrees C). Line a 12-cup muffin tin with paper liners.
2. In a large bowl, stir together milk, egg, and oil. Add flour, baking powder, sugar, and blueberries; gently mix the batter with only a few strokes. Spoon batter into cups.
Finding best action graph

Stir together milk, egg, and oil.

Add flour, baking powder, sugar, and blueberries;

Gently mix the batter with only a few strokes.

Spoon batter into cups.

Bake for 20 minutes.
Stir together milk, egg, and oil.

Spoon batter into cups.

Add flour, baking powder, sugar, and blueberries; gently mix the batter with only a few strokes.

Bake for 20 minutes.
Semantic challenges

• Traditional parsers have trouble with imperatives
  – **Grease** with butter. **Grease = noun?**

• Elided arguments are common.
  – Bake for 30 minutes. **Bake what? Bake where?**

• Referring expressions use physical properties
  – Whisk eggs. Add flour. Fold sugar into **the wet mixture.**
Blueberry Muffins

Ingredients
1 cup milk
1 egg
1/3 cup vegetable oil
2 cups all-purpose flour
2 teaspoons baking powder
1/2 cup white sugar
1/2 cup fresh blueberries

Procedure
1. Preheat oven to 400 degrees F (205 degrees C). Line a 12-cup muffin tin with paper liners.
2. In a large bowl, stir together milk, egg, and oil. Add flour, baking powder, sugar, and blueberries; gently mix the batter with only a few strokes. Spoon batter into cups.
“In a large bowl, stir together milk, egg, and oil.”
“In a large bowl, stir together milk, egg, and oil. Add flour to the wet mixture.”
"In a large bowl, stir together milk, egg, and oil. Add flour."
Related Work

• Maeta et al. 2015,
• Mori et al. 2014
• Tasse and Smith 2008
Unsupervised Learning (Kiddon et al. 2015)

- **Chicken and Egg**
  - Parsing (unstructured text → action graph) requires knowledge
  - Knowledge requires parsing

- **Model:**
  - Probabilistic Model

- **Learning:**
  - Expectation-Maximization
Probability model $P(C,R)$ (Kiddon et al. 2015)

- **Input**: A set of connections $C$ and a recipe $R$ segmented (Sec. 6) into its actions $\{e_1 = (v_1, a_1), \ldots, e_n = (v_n, a_n)\}$
- The joint probability of $C$ and $R$ is $P(C, R) = P(C)P(R|C)$, each defined below:

1. **Connections Prior (Sec. 3.1)**: $P(C) = \prod_i P(d_i|d_1, \ldots, d_{i-1})$
   Define $d_i$ as the list of connections with destination index $i$. Let $c_p = (o, i, j, k, t^{syn}, t^{sem}) \in d_i$. Then,
   - $P(d_i|d_1, \ldots, d_{i-1}) = P(vs(d_i)) \prod_{c_p \in d_i} P(1(o \rightarrow s^k_{ij})|vs(d_i), d_1, \ldots, d_{i-1}, c_1, \ldots, c_{p-1})$
     (a) $P(vs(d_i))$: multinomial verb signature model (Sec. 3.1.1)
     (b) $P(1(o \rightarrow s^k_{ij})|vs(d_i), d_1, \ldots, d_{i-1}, c_1, \ldots, c_{p-1})$: multinomial connection origin model, conditioned on the verb signature of $d_i$ and all previous connections (Sec. 3.1.2)

2. **Recipe Model (Sec. 3.2)**: $P(R|C) = \prod_i P(e_i|C, e_1, \ldots, e_{i-1})$
   For brevity, define $h_i = (e_1, \ldots, e_{i-1})$.
   - $P(e_i|C, h_i) = P(v_i|C, h_i)P(a_{ij}|C, h_i)$ (Sec. 3.2)
     Define argument $a_{ij}$ by its types and spans, $a_{ij} = (t^{syn}_{ij}, t^{sem}_{ij}, S_{ij})$.
     (a) $P(v_i|C, h_i) = P(v_i|g_i)$: multinomial verb distribution conditioned on verb signature (Sec. 3.2)
     (b) $P(a_{ij}|C, h_i) = P(t^{syn}_{ij}, t^{sem}_{ij}|C, h_i) \prod_{s^k_{ij} \in S_{ij}} P(s^k_{ij}|t^{syn}_{ij}, t^{sem}_{ij}, C, h_i)$
       i. $P(t^{syn}_{ij}, t^{sem}_{ij}|C, h_i)$: determinstic argument types model given connections (Sec. 3.2.1)
       ii. $P(s^k_{ij}|t^{syn}_{ij}, t^{sem}_{ij}, C, h_i)$: string span model computed by case (Sec. 3.2.2):
          A. $t^{sem}_{ij} = \text{food}$ and $\text{origin}(s^k_{ij}) \neq \emptyset$: IBM Model 1 generating composites (Part-composite model)
          B. $t^{sem}_{ij} = \text{food}$ and $\text{origin}(s^k_{ij}) = \emptyset$: naïve Bayes model generating raw food references (Raw food model)
          C. $t^{sem}_{ij} = \text{location}$: model for generating location referring expressions (Location model)

Figure 2: Summary of the joint probabilistic model $P(C, R)$ over connection set $C$ and recipe $R$. 
Probabilistic model

• Assume we are given a preprocessed recipe text \( R \) that has been segmented into actions
• Probabilistic model over action graphs to determine most likely connections \( C \) for the recipe

\[
P(C, R) = P(C)P(R|C)
\]

prior over connections

probability of recipe text given connections
Recipe distribution: $P(R|C)$

- $R$ is a sequence of actions $e_1, \ldots, e_n$

$$P(R|C) = \prod_i P(e_i|C, e_1, \ldots, e_{i-1})$$

- Actions decompose into the probability of the verbs, arguments, and spans

Diagram:

- combine
- batter
- blueberries
- bowl
Local search

- Initialize with sequential connections
- Score local search operators and greedily apply

Add new connection
Swap connections
Model learning

• Unsupervised hard EM method
• First, initialize models. Then:

Recurse:

• **E-step**: Update $C \leftarrow \text{argmax}_c P(C,R)$ for each $R$ in dataset using local search

• **M-step**: Update parameters of $P(C,R)$ using action graphs generated in E-step
Knowledge in the Model

- **Part-composite model:** how likely it is to generate a composite word given the incoming ingredients/raw materials
  
  - \( P(\text{“dressing”} \mid \text{“oil” “vinegar”}) > P(\text{“batter”} \mid \text{“oil” “vinegar”}) \)

- **Raw materials model:** how likely a word is to be an initial reference
  
  - \( P(\text{“batter”} \mid \text{initial reference}) < P(\text{“flour”} \mid \text{initial reference}) \)

- **Location model:** how likely a location is given the action verb
Learned cooking knowledge

Learned good composite words for different ingredients

<table>
<thead>
<tr>
<th>Ingredient</th>
<th>Composite Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>eggs</td>
<td>egg, yolk, mixture, noodles, whites</td>
</tr>
<tr>
<td>beef</td>
<td>beef, mixture, grease, meat, excess</td>
</tr>
<tr>
<td>flour</td>
<td>flour, mixture, dough, batter, top, crust</td>
</tr>
</tbody>
</table>

Learned selectional preferences for verb

- **add** is 58% likely to have two arguments that are not both raw materials
- **bake** is 95% likely to have one non-raw material argument
Evaluation

- Cooking recipe domain, 2456 recipes, 20 dish types
- 100 manually-annotated gold-standard recipes
To Conclude

• Unsupervised parsing of instructional recipes to action diagrams
• Possible due to repeated patterns in naturally existing data
• Knowledge is a recurring theme.
What’s Next: Composing a New Recipe

Compose new recipes given a recipe title (or what’s in the fridge)!
- With or without explicit meaning representation
- New challenge: generating a cohesive discourse
- zero-shot learning for recipes

Grounding instructions with multimodal perception
Learning Knowledge about the World
Take IV: Prototypical Events

Bosselut et al. @ ACL 2016
What makes a wedding a wedding?
Learned Events:

Dance
Kiss
Cut the cake
Vows
Exchange rings

Temporal Knowledge:

1 - Reading our vows.
2 - Exchanging our rings.
3 - Cake cutting.
4 - The cake was so solid.
5 - Dancing excitement.

Prototypical Captions:

-Dancing excitement.
-First dance.
-Ballroom dancing.
-Our first ever kiss.
-You may kiss the bride.
-Sealed with a kiss.
-Cake cutting.
-Our vows.
Circular Dependency

Better knowledge about stereotypical event structure

Better understanding of a new photo album
Data Compilation

- 12 common life scenarios
Learning Prototypical Events

• k-means clustering (on language only)

• Multimodal cluster representation
  – Weighted unigram features of content words
  – Visual Features from VGGNet

• Name each cluster with the most common word
### Sample Events and Prototypical Captions

<table>
<thead>
<tr>
<th>Wedding</th>
<th>Camping</th>
<th>Funeral</th>
</tr>
</thead>
<tbody>
<tr>
<td>aisle</td>
<td>tent</td>
<td>Graveside service</td>
</tr>
<tr>
<td>Walking down the aisle</td>
<td>Inside out tent</td>
<td>The service</td>
</tr>
<tr>
<td>Bride walking down the aisle</td>
<td>Setting up the tent</td>
<td>Paying Respects</td>
</tr>
<tr>
<td>Bride walking down the aisle</td>
<td></td>
<td>Respect</td>
</tr>
<tr>
<td>vow</td>
<td>fire</td>
<td></td>
</tr>
<tr>
<td>Exchanging vows</td>
<td>Building the fire</td>
<td></td>
</tr>
<tr>
<td>Reading the vows</td>
<td>Around the fire</td>
<td></td>
</tr>
<tr>
<td>Reciting vows to each other</td>
<td>Getting the fire going</td>
<td></td>
</tr>
<tr>
<td>dance</td>
<td>sunset</td>
<td>Sunset on the first night</td>
</tr>
<tr>
<td>First dance</td>
<td>Sunset from camp</td>
<td></td>
</tr>
<tr>
<td>Everybody dancing</td>
<td>Watching the sunset</td>
<td></td>
</tr>
<tr>
<td>Dancing the night away</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Learn Temporal Knowledge

- **Local transition probabilities** – Probability that a photo assigned event A being followed by a photo assigned to event B.

\[
P_L(e_k \rightarrow e_l) = \frac{C(e_k \rightarrow e_l)}{\sum_{m=1}^{N} C(e_k \rightarrow e_m)}
\]

- \(P_L(\text{kiss} \rightarrow \text{dance}) = 0.04\)
- \(P_L(\text{dance} \rightarrow \text{toast}) = 0.17\)
Learn Temporal Knowledge

- Global pairwise ordering probabilities – Probability that a photo assigned event A precedes a photo assigned event B anywhere in the album

\[
P_G(e_k \Rightarrow e_l) = \frac{C(e_k \Rightarrow e_l)}{C(e_k \Rightarrow e_l) + C(e_l \Rightarrow e_k)}
\]

\[P_G(\text{vows} \Rightarrow \text{dance}) = .79\]

\[P_G(\text{vows} \Rightarrow \text{toast}) = .84\]
Circular Dependency

Better knowledge about stereotypical event structure

Better understanding of a new photo album
Individual Photo Album Analysis

• Input: An album of photos
• Output: An album partitioned by the scenario’s compositional events

ready aisle vows kiss dancing toasts
Inference

• Constrained Optimization to decode assignment and ordering of events

\[ F = \phi_{\text{event}} + \phi_{\text{seg}} + \phi_{\text{temporal}} \]
\[
\phi_{\text{event}} = \sum_{i=1}^{M} \sum_{k=1}^{N} (\gamma_{\text{ce}} A_{i,k}^c + \gamma_{\text{ve}} A_{i,k}^v) X_{i,k}
\]

Textual Affinity  Visual Affinity

\[
\phi_{\text{temporal}} = \gamma_{lp} \sum_{i=0}^{M} \sum_{k=1}^{N} L_{k,l} Z_{i,i+1,k,l} + \gamma_{gp} \sum_{i=1}^{M} \sum_{j=i}^{M} \sum_{k=1}^{N} G_{k,l} Z_{i,j,k,l}
\]

Local Transition Probabilities  Global ordering probabilities

\[
\phi_{\text{seg}} = \sum_{i=1}^{M-1} \sum_{k=1}^{N} (\gamma_{\text{cs}} b_{i}^c + \gamma_{\text{vs}} b_{i}^v) Z_{i,i+1,k,k}
\]

Textual Similarity  Visual Similarity

\[F = \phi_{\text{event}} + \phi_{\text{seg}} + \phi_{\text{temporal}}\]
Experiments

• Temporal Ordering
• Album Segmentation
• Learned Knowledge
  – Summarization
  – Captioning
Temporal Ordering

- Compile pairwise event training set ordering statistics between all events
- In every album of the test set, pick two photos
- Based on the events assigned to those photos, predict which photo was taken before the other
Temporal Ordering

Pairwise Event Ordering Accuracy

- **k-means**
- **No Temporal**
- **Full Model**
Temporal Ordering

Pairwise Event Ordering Accuracy

- k-means
- No Temporal
- Full Model
Learned Knowledge: Summarization

• Pick a set of $b$ photos from an album as a summary

• Choose photos from $b$ different events
• Choose photo with highest affinity for event
• Replace caption with a prototypical caption
Wedding Summaries

- Watching the ceremony
- Getting married
- Greeting guests
- Laughing
- Mother of the bride
- Eating cake
- Listening to speeches
- Watching the ceremony
- Bridal party
- Listening to the toasts
- Father daughter dance
- Cutting the cake
- The Library
- Last dance
- The Arch
- Watching the ceremony
- ...heading for the reception
- The kiss
- Reciting vows to each other
- First dance
- Listening to the toasts
Baby Birth Summaries

Figure 4: Wedding Example Summaries; Each row represents a summary generated by our full model.

Figure 5: Baby Birth Example Summaries; Each row represents a summary generated by our full model.

Figure 6: Marathon Example Summaries; Each row represents a summary generated by our full model.
To Conclude

• Multimodal script learning from photo albums
• Prototypical event structure of 12 common scenarios
• Future work: integration of videos, and scaling up the knowledge
In this talk

- Toward Intelligent Communication

- Learning knowledge about the world
  - Physical Knowledge (size)
  - Visual Entailment
  - Recipe Parsing with Cooking Knowledge
  - Prototypical Event Knowledge

- From naturally existing data
  - No manually curated data for training
Acknowledgements

My PhD
Chloe Kiddon, Antoine Bosselut, Song Feng, Polina Kuznetsova, Jianfu Chen, Vicente Ordonez, Karl Stratos, Siming Li, Jesse Dodge, Hamid Izadinia, Fereshteh Sadeghi

Other PhD
Girish Kulkarni, Sagnik Dhar, Visruth Premraj

MS

Professor
Margaret Mitchell, Santosh Divvala, Sujith Ravi, Ravi Kumar, Amit Goyal

Industry
Thanks!
Thanks!

- Eunsol Choi
- Hannah Rashkin
- Maarten Sap
- Max Forbes
- Chloe Kiddon
- Li Zilles
- Antoine Bosselut