Explainable Improved Ensembling for Natural Language and Vision

Nazneen Rajani
University of Texas at Austin
Ph.D. Defense (12th July, 2018)
NLP
- Discourse
- Sentiment Analysis
- Entity Linking
- Language Modeling
- Relation Extraction
- Parsing

Vision
- Visual Question Answering (VQA)
- Image Captioning
- Scene Recognition
- Object Tracking
- Fine-grained classification
- Object Detection
- Image Classification
NLP
- Discourse
- Sentiment Analysis
- Entity Linking
- Language Modeling
- Relation Extraction
- Parsing

Vision
- Scene Recognition
- Object Tracking
- Fine-grained classification
- Object Detection
- Image Classification

VQA
- Image Captioning

XAI
- Rationalization
- Visual Explanations
- Textual Explanations
- Explanation Evaluation
My Research

• Develop **improved ensemble models** for language and vision applications.
• Develop methods to **generate and evaluate explanations** for ensemble models.
Before Proposal

NLP
- Discourse
- Sentiment Analysis
- Entity Linking
- Language Modeling
- Relation Extraction
- Parsing

Vision
- Scene Recognition
- Object Tracking
- Fine-grained classification
- Object Detection
- Image Classification
- Image Captioning
- VQA

Combining supervised and Unsupervised Ensembling (EMNLP’16)
Stacking for KBP (ACL’15)
Stacking With Auxiliary Features (IJCAI’17)
Before Proposal

NLP
- Discourse
- Sentiment Analysis
- Entity Linking
- Language Modeling
- Relation Extraction
- Parsing

Vision
- Scene Recognition
- Object Tracking
- Fine-grained classification
- Object Detection
- Image Classification
- Image Captioning
- VQA

Combining supervised and Unsupervised Ensembling (EMNLP’16)
Stacking for KBP (ACL’15)
Stacking With Auxiliary Features (IJCAI’17)
Before Proposal

**NLP**
- Discourse
- Sentiment Analysis
- Entity Linking
- Language Modeling
- Relation Extraction
- Parsing

**Vision**
- Scene Recognition
- Object Tracking
- Fine-grained classification
- Object Detection
- Image Classification
- VQA
- Image Captioning

Combining supervised and Unsupervised Ensembling (EMNLP’16)

Stacking for KBP (ACL’15)

Stacking With Auxiliary Features (IJCAI’17)
Before Proposal

**NLP**
- Discourse
- Sentiment Analysis
- Language Modeling
- Entity Linking
- Relation Extraction
- Parsing

**Vision**
- Scene Recognition
- Object Tracking
- Fine-grained classification
- Object Detection
- Image Classification
- Image Captioning
- VQA

**Combining supervised and Unsupervised Ensembling (EMNLP’16)**

**Stacking for KBP (ACL’15)**

**Stacking With Auxiliary Features (IJCAI’17)**
Before Proposal

NLP
- Discourse
- Sentiment Analysis
- Entity Linking
- Language Modeling
- Relation Extraction
- Parsing

Vision
- Scene Recognition
- Object Tracking
- Fine-grained classification
- Object Detection
- Image Classification
- VQA
- Image Captioning

Combining supervised and Unsupervised Ensembling (EMNLP’16)
Stacking for KBP (ACL’15)
Stacking With Auxiliary Features (IJCAI’17)
Before Proposal

NLP
- Discourse
- Sentiment Analysis
- Language Modeling
- Relation Extraction
- Parsing

Vision
- Scene Recognition
- Object Tracking
- Fine-grained classification
- Image Classification
- Image Captioning

Entity Linking

Stacking for KBP (ACL’15)
Combining supervised and Unsupervised Ensembling (EMNLP’16)
Stacking With Auxiliary Features (IJCAI’17)
Since Proposal

NLP
- Discourse
- Sentiment Analysis
- Entity Linking
- Language Modeling
- Relation Extraction
- Parsing

Vision
- Scene Recognition
- Object Tracking
- Fine-grained classification
- Object Detection
- Image Classification
- Image Captioning

XAI
- Rationalization
- Visual Explanations
- Textual Explanations
- Explanation Evaluation
- Generating and Evaluating Visual Explanations (ViGIL’17) (Under review at NIPS)
- Stacking with Auxiliary Features for VQA (NAACL’18)
Since Proposal

NLP
- Discourse
- Sentiment Analysis
- Entity Linking
- Language Modeling
- Relation Extraction
- Parsing

Vision
- Scene Recognition
- Object Tracking
- Fine-grained classification
- Object Detection
- Image Classification

XAI
- Rationalization
- Visual Explanations
  - Textual Explanations
  - Explanation Evaluation
- Stacking with Auxiliary Features for VQA (NAACL’18)

VQA
- Image Captioning

Generating and Evaluating Visual Explanations (ViGIL’17)
(Under review at NIPS)
Since Proposal

NLP
- Discourse
- Sentiment Analysis
- Entity Linking
- Language Modeling
- Relation Extraction
- Parsing

Vision
- Scene Recognition
- Object Tracking
- Fine-grained classification
- Object Detection
- Image Classification
- VQA
- Image Captioning

XAI
- Rationalization
- Visual Explanations
- Textual Explanations
- Explanation Evaluation

Stacking with Auxiliary Features for VQA (NAACL’18)
Generating and Evaluating Visual Explanations (ViGIL’17) (Under review at NIPS)
Since Proposal

NLP
- Discourse
- Sentiment Analysis
- Entity Linking
- Language Modeling
- Relation Extraction
- Parsing

Vision
- Scene Recognition
- Object Tracking
- Fine-grained classification
- Object Detection
- Image Classification

VQA
- Image Captioning

XAI
- Rationalization
- Visual Explanations
- Textual Explanations
- Explanation Evaluation

Stacking with Auxiliary Features for VQA (NAACL’18)
Generating and Evaluating Visual Explanations (ViGIL’17) (Under review at NIPS)
Ensembling

- Used by the $1M winning team for the Netflix competition.
Ensembling

• Make **auxiliary** information accessible to the ensemble.
Before Proposal

NLP
- Discourse
- Sentiment Analysis
- Entity Linking
- Language Modeling
- Relation Extraction
- Parsing

Vision
- Scene Recognition
- Object Tracking
- Fine-grained classification
- Object Detection
- Image Classification
- VQA
- Image Captioning

Stacking for KBP (ACL’15)
Combined supervised and Unsupervised Ensembling (EMNLP’16)
Stacking With Auxiliary Features (IJCAI’17)
Before Proposal

**NLP**
- Discourse
- Sentiment Analysis
- Entity Linking
- Language Modeling
- Relation Extraction
- Parsing

**Vision**
- Scene Recognition
- Object Tracking
- Fine-grained classification
- Object Detection
- Image Classification
- Image Captioning
- VQA

**Methods**
- Combined supervised and Unsupervised Ensembling (EMNLP’16)
- Stacking for KBP (ACL’15)
- Stacking With Auxiliary Features (IJCAI’17)
Relation Extraction

• Knowledge Base Population (KBP) sub-task of discovering entity facts and adding to a KB.
• Relation extraction using fixed ontology is slot-filling.
• Along with extracted entities, systems provide:
  - confidence score
  - provenance — docid:startoffset-endoffset
Microsoft is a technology company, headquartered in Redmond, Washington.

Microsoft was founded by Paul Allen and Bill Gates on April 4, 1975, to develop and sell BASIC interpreters for the Altair 8800.
Stacking

(Wolpert, 1992)
Stacking with Auxiliary Features for KBP

(Viswanathan* et al., ACL’15)
Provenance Feature

• Document Provenance:
  - $DP_i = n/N$ for a system $i$ where $n$ is number of systems that extracted from the same document and $N$ is total number of systems.

• Offset Provenance using Jaccard similarity:

$$OP_n = \frac{1}{|N|} \sum_{i \in N, i \neq n} \frac{|\text{substring}(i) \cap \text{substring}(n)|}{|\text{substring}(i) \cup \text{substring}(n)|}$$

(Viswanathan* et al., ACL’15)
# Offset Provenance

(Viswanathan* et al., ACL’15)

## Offsets

<table>
<thead>
<tr>
<th>Offsets</th>
<th>System 1</th>
<th>System 2</th>
<th>System 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start offset</td>
<td>7</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>End Offset</td>
<td>28</td>
<td>15</td>
<td>22</td>
</tr>
</tbody>
</table>

\[
OP_1 = \frac{1}{2} \left( \frac{9}{29} + \frac{16}{22} \right) \\
OP_2 = \frac{1}{2} \left( \frac{9}{29} + \frac{9}{23} \right) \\
OP_3 = \frac{1}{2} \left( \frac{16}{22} + \frac{9}{23} \right)
\]
Slot-Filling Results

- 2014 KBP SF task — 10 shared systems

<table>
<thead>
<tr>
<th>Approach</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Union</td>
<td>0.176</td>
<td>0.647</td>
<td>0.277</td>
</tr>
<tr>
<td>Voting</td>
<td>0.694</td>
<td>0.256</td>
<td>0.374</td>
</tr>
<tr>
<td>Best SF system in 2014 (Stanford)</td>
<td>0.585</td>
<td>0.298</td>
<td>0.395</td>
</tr>
<tr>
<td>Stacking</td>
<td>0.606</td>
<td>0.402</td>
<td>0.483</td>
</tr>
<tr>
<td>Stacking + Slot-type</td>
<td>0.607</td>
<td>0.406</td>
<td>0.486</td>
</tr>
<tr>
<td>Stacking + Provenance + Slot-type</td>
<td>0.541</td>
<td>0.466</td>
<td>0.501</td>
</tr>
</tbody>
</table>

(Viswanathan* et al., ACL’15)
Before Proposal

**NLP**
- Discourse
- Sentiment Analysis
- Entity Linking
- Language Modeling
- Relation Extraction
- Parsing

**Vision**
- Scene Recognition
- Object Tracking
- Fine-grained classification
- Object Detection
- Image Classification
- VQA
- Image Captioning

**Combining supervised and Unsupervised Ensembling** (EMNLP’16)

**Stacking for KBP** (ACL’15)

**Stacking With Auxiliary Features** (IJCAI’17)
Before Proposal

NLP
- Discourse
- Sentiment Analysis
- Language Modeling
- Relation Extraction
- Parsing

Vision
- Scene Recognition
- Object Tracking
- Fine-grained classification
- Object Detection
- Image Classification
- Image Captioning
- VQA

Entity Linking
- Combining supervised and Unsupervised Ensembling (EMNLP’16)
- Stacking for KBP (ACL’15)
- Stacking With Auxiliary Features (IJCAI’17)
Entity Linking

• KBP sub-task involving two NLP problems:
  - Named Entity Recognition (NER)
  - Disambiguation
• Link mentions to English KB (FreeBase).
• If no KB entry found, cluster into a NIL ID.
Entity Discovery and Linking (EDL)

**FreeBase entry:**
Hillary Diane Rodham Clinton is a US Secretary of State, U.S. Senator, and First Lady of the United States. From 2009 to 2013, she was the 67th Secretary of State, serving under President Barack Obama. She previously represented New York in the U.S. Senate.

**FreeBase entry:**

**Source Corpus Document:**
*Hillary Clinton* Not Talking About ’92 *Clinton-Gore* Confederate Campaign Button..
Entity Discovery and Linking (EDL)

Source Corpus Document: Hillary Clinton
Not Talking About ’92 Clinton-Gore
Confederate Campaign Button..

FreeBase entry:
Hillary Diane Rodham Clinton is a US Secretary of State, U.S. Senator, and First Lady of the United States. From 2009 to 2013, she was the 67th Secretary of State, serving under President Barack Obama. She previously represented New York in the U.S. Senate.

FreeBase entry:
Hillary Diane Rodham Clinton is a US Secretary of State, U.S. Senator, and First Lady of the United States. From 2009 to 2013, she was the 67th Secretary of State, serving under President Barack Obama. She previously represented New York in the U.S. Senate.

Combining supervised & unsupervised ensembles

Sup System 1 \rightarrow conf 1

Sup System 2 \rightarrow conf 2

\vdots

Sup System N \rightarrow conf N

Unsup System 1

Unsup System 2

\vdots

Unsup System M

Auxiliary Features

Trained linear SVM

Calibrated conf

Constrained Optimization (Weng et al, 2013)

Accept?

(Rajani and Mooney, EMNLP’16)
Constrained Optimization

- Approach to aggregate raw confidence values.
- Re-weight the confidence score of an instance:
  - number of systems that produce it.
  - performance of those systems.
- Uniform weights for all systems.
- Our work extends to entity linking.

(Wang et al., 2013)
### Results

- **2015 SF** — \#sup systems = 10, \#unsup systems = 13

<table>
<thead>
<tr>
<th>Approach</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constrained optimization</td>
<td>0.1712</td>
<td>0.3998</td>
<td>0.2397</td>
</tr>
<tr>
<td>Oracle voting (&gt;=3)</td>
<td>0.4384</td>
<td>0.2720</td>
<td>0.3357</td>
</tr>
<tr>
<td>Top ranked system (Angeli et al., 2015)</td>
<td>0.3989</td>
<td>0.3058</td>
<td>0.3462</td>
</tr>
<tr>
<td>Stacking + slot-type + provenance</td>
<td>0.4656</td>
<td>0.3312</td>
<td>0.3871</td>
</tr>
<tr>
<td><strong>Stacking for combining sup + unsup (constrained optimization)</strong></td>
<td><strong>0.4676</strong></td>
<td><strong>0.4314</strong></td>
<td><strong>0.4489</strong></td>
</tr>
</tbody>
</table>

(Rajani and Mooney, EMNLP’16)
## Results

- 2015 EDL — #sup systems=6, #unsup systems=4

<table>
<thead>
<tr>
<th>Approach</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constrained optimization</td>
<td>0.176</td>
<td>0.445</td>
<td>0.252</td>
</tr>
<tr>
<td>Oracle voting (&gt;=4)</td>
<td>0.514</td>
<td>0.601</td>
<td>0.554</td>
</tr>
<tr>
<td>Top ranked system (Sil et al., 2015)</td>
<td>0.693</td>
<td>0.547</td>
<td>0.611</td>
</tr>
<tr>
<td>Stacking + entity-type +provenance</td>
<td>0.813</td>
<td>0.515</td>
<td>0.630</td>
</tr>
<tr>
<td>Stacking for combining sup + unsup (constrained optimization)</td>
<td>0.686</td>
<td>0.624</td>
<td>0.653</td>
</tr>
</tbody>
</table>

(Rajani and Mooney, EMNLP’16)
Before Proposal

**NLP**
- Discourse
- Sentiment Analysis
- Entity Linking
- Language Modeling
- Relation Extraction
- Parsing

**Vision**
- Scene Recognition
- Object Tracking
- Fine-grained classification
- Object Detection
- Image Classification
- VQA

**Combining supervised and Unsupervised Ensembling (EMNLP’16)**

**Stacking with Auxiliary Features (IJCAI’17)**

**Stacking for KBP (ACL’15)**
Before Proposal

NLP
- Discourse
- Sentiment Analysis
- Language Modeling
- Relation Extraction
- Parsing

Vision
- Scene Recognition
- Object Tracking
- Fine-grained classification
- Image Classification
- Image Captioning
- VQA

Entity Linking

Stacking for KBP (ACL’15)
Combining supervised and Unsupervised Ensembling (EMNLP’16)
Stacking With Auxiliary Features (IJCAI’17)
Object Detection

• Well known vision problem for object recognition.
• Annually conducted by ImageNET on very large datasets.
• Object detection:
  - detect all instances of object categories in images (total 200).
  - localize using axis-aligned Bounding Boxes (BB).
ImageNet Object Detection
Stacking with Auxiliary Features (SWAF)

- Stacking using two types of auxiliary features:

  - System 1
  - System 2
  - ... System N-1
  - System N

  conf 1
  conf 2
  ... conf N-1
  conf N

  Trained Meta-classifier

  Accept?

  Auxiliary Features
  - Instance Features
  - Provenance Features

(Rajani and Mooney, IJCAI’17)
Instance Features

- Enables stacker to discriminate between input instance types.
- Some systems are better at certain input types.
- Slot-filling — slot type (per:age, org:headquarters).
- Entity Linking — entity type (PER/ORG/GPE).
- Object detection — object category and SIFT feature descriptors.

(Rajani and Mooney, IJCAI’17)
Provenance Features

- Enables the stacker to discriminate between systems.
- Output is reliable if systems agree on source.
- Slot-filling & Entity Linking — substring overlap.
- Object detection — measure BB overlap.

\[ BBO(n) = \frac{1}{|N|} \times \sum_{i \in N, i \neq n} \frac{|\text{Area}(i) \cap \text{Area}(n)|}{|\text{Area}(i) \cup \text{Area}(n)|} \]
Object Detection Provenance Features

(Rajani and Mooney, IJCAI’17)
## Slot Filling Results

- **2016 SF — 8 shared systems**

<table>
<thead>
<tr>
<th>Approach</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oracle voting (&gt;=4)</td>
<td>0.191</td>
<td>0.379</td>
<td>0.206</td>
</tr>
<tr>
<td>Top ranked system (Zhang et al., 2016)</td>
<td>0.265</td>
<td>0.302</td>
<td>0.260</td>
</tr>
<tr>
<td>Stacking</td>
<td><strong>0.311</strong></td>
<td>0.253</td>
<td>0.279</td>
</tr>
<tr>
<td>Stacking + instance features</td>
<td>0.257</td>
<td>0.346</td>
<td>0.295</td>
</tr>
<tr>
<td>Stacking + provenance features</td>
<td>0.252</td>
<td>0.377</td>
<td>0.302</td>
</tr>
<tr>
<td>SWAF</td>
<td>0.258</td>
<td><strong>0.439</strong></td>
<td><strong>0.324</strong></td>
</tr>
</tbody>
</table>

(Rajani and Mooney, IJCAI’17)
## Entity Linking Results

- 2016 EDL — 6 shared systems

<table>
<thead>
<tr>
<th>Approach</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oracle voting (&gt;=4)</td>
<td>0.588</td>
<td>0.412</td>
<td>0.485</td>
</tr>
<tr>
<td>Top ranked system (Sil et al., 2016)</td>
<td>0.717</td>
<td>0.517</td>
<td>0.601</td>
</tr>
<tr>
<td>Stacking</td>
<td>0.723</td>
<td>0.537</td>
<td>0.616</td>
</tr>
<tr>
<td>Stacking + instance features</td>
<td>0.752</td>
<td>0.542</td>
<td>0.630</td>
</tr>
<tr>
<td>Stacking + provenance features</td>
<td><strong>0.767</strong></td>
<td>0.544</td>
<td>0.637</td>
</tr>
<tr>
<td>SWAF</td>
<td>0.739</td>
<td><strong>0.600</strong></td>
<td><strong>0.662</strong></td>
</tr>
</tbody>
</table>
Object Detection Results

• 2015 ImageNet object detection—3 shared systems

<table>
<thead>
<tr>
<th>Approach</th>
<th>Mean AP</th>
<th>Median AP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oracle voting (&gt;=1)</td>
<td>0.366</td>
<td>0.368</td>
</tr>
<tr>
<td>Best standalone system (VGG + selective search)</td>
<td>0.434</td>
<td>0.430</td>
</tr>
<tr>
<td>Stacking</td>
<td>0.451</td>
<td>0.441</td>
</tr>
<tr>
<td>Stacking + instance features</td>
<td>0.461</td>
<td>0.45</td>
</tr>
<tr>
<td>Stacking + provenance features</td>
<td>0.502</td>
<td>0.494</td>
</tr>
<tr>
<td>SWAF</td>
<td>0.506</td>
<td>0.497</td>
</tr>
</tbody>
</table>
Since Proposal

**NLP**
- Discourse
- Sentiment Analysis
- Entity Linking
- Language Modeling
- Relation Extraction
- Parsing

**Vision**
- Scene Recognition
- Object Tracking
- Fine-grained classification
- Object Detection
- Image Classification
- Image Captioning

**XAI**
- Rationalization
- Visual Explanations
- Textual Explanations
- Explanation Evaluation
- Stacking with Auxiliary Features for VQA (NAACL’18)
- Generating and Evaluating Visual Explanations (ViGIL’17) (Under review at NIPS)
Since Proposal

NLP
- Discourse
- Sentiment Analysis
- Entity Linking
- Language Modeling
- Relation Extraction
- Parsing

Vision
- Scene Recognition
- Object Tracking
- Fine-grained classification
- Object Detection
- Image Classification

VQA
- Image Captioning

XAI
- Rationalization
- Visual Explanations
- Textual Explanations
- Explanation Evaluation

Stacking with Auxiliary Features for VQA (NAACL’18)
Generating and Evaluating Visual Explanations (ViGIL’17) (Under review at NIPS)
Visual Question Answering (VQA)

- VQA involves both language and vision understanding.
- Data in the form of image and a set of questions.
- Requires inferring from the image.
- Multiple datasets:
  - DAQUAR (Malinowski and Fritz, 2014)
  - VQA (Antol et al., 2015)
  - CLEVR (Johnson et al., 2017)
  - NLVR (Suhr et al., 2017)
Visual Question Answering (VQA)

**Visual-Question** and **Only-Question**

<table>
<thead>
<tr>
<th>What is in the child's mouth?</th>
<th>her thumb</th>
<th>it's thump</th>
<th>thumb</th>
<th>candy</th>
<th>cookie</th>
<th>lollipop</th>
</tr>
</thead>
<tbody>
<tr>
<td>What is the child harnessed to?</td>
<td>her thumb</td>
<td>high chair</td>
<td>seat</td>
<td>bike</td>
<td>child seat</td>
<td>seat</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>What is the animal in the water?</th>
<th>dog</th>
<th>dog</th>
<th>duck</th>
<th>duck</th>
<th>guppy</th>
</tr>
</thead>
<tbody>
<tr>
<td>How many people are present?</td>
<td>15</td>
<td>15</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>
Component VQA systems

- Three deep learning models:
  1. LSTM (Antol et al., 2015)
  2. Hierarchical Co-Attention (HieCoAtt) (Lu et al., 2016)
  3. Multimodal Compact Bilinear Pooling (MCB) (Fukui et al., 2016)
SWAF for VQA

- Three types of auxiliary features that can be inferred from image-question pair

1. Question & Answer types
   - Question prefixes — “What is the color of the vase?”
   - Answer types — yes/no, number and other

2. Question Features
   - BOW representation of words in the question

3. Image Features
   - VGGNet’s fc7 layer

(Rajani and Mooney, NAACL’18)
Visual Explanation as Auxiliary Features

• DNNs attend to relevant regions of image while doing VQA (Goyal et al., 2016).

• The parts of images that the models focus on can be viewed as a **visual explanation**.

• We use *heat-maps* to visualize explanations in images.

• Enable the stacker to learn to rely on systems that “look” at the right region of the image while predicting the answer.
Visual Explanation

What is the man doing?  
Surfing

What is she holding?  
Baseball bat

What is that?  
Elephant

What is that?  
Zebra
Generating Visual Explanation

- **GradCAM** (Selvaraju et al., 2017) is used to generate heat-map explanations.
Generating Visual Explanation Features

- Measure agreement between systems’ heat-maps using *rank order correlation*.

**Q:** What sport is this?

**A:** Tennis

**A:** Baseball
Generating Visual Explanation Features

Q: What is the kid doing?  A: Skateboarding

LSTM  MCB  HieCoAtt
Generating Visual Explanation Features

Q: Are there mushrooms in the grass by the zebra?  
A: Yes

LSTM  
MCB  
HieCoAtt

(Rajani and Mooney, NAACL’18, XAI’17)
## VQA Results

(Rajani and Mooney, NAACL’18)

<table>
<thead>
<tr>
<th>Approach</th>
<th>All</th>
<th>Yes/No</th>
<th>Number</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>DPPNet (Noh et al., 2016)</td>
<td>57.36</td>
<td>80.28</td>
<td>36.92</td>
<td>42.24</td>
</tr>
<tr>
<td>NMNs (Andreas et al., 2016)</td>
<td>58.70</td>
<td>81.20</td>
<td>37.70</td>
<td>44.00</td>
</tr>
<tr>
<td>MCB (Best component system) (Fukui et al., 2016)</td>
<td>62.56</td>
<td>80.68</td>
<td>35.59</td>
<td>52.93</td>
</tr>
<tr>
<td>MCB (Ensemble) (Fukui et al., 2016)</td>
<td>66.50</td>
<td><strong>83.20</strong></td>
<td>39.50</td>
<td>58.00</td>
</tr>
<tr>
<td>Voting (MCB + HieCoAtt + LSTM)</td>
<td>60.31</td>
<td>80.22</td>
<td>34.92</td>
<td>48.83</td>
</tr>
<tr>
<td>Stacking</td>
<td>63.12</td>
<td>81.61</td>
<td>36.07</td>
<td>53.77</td>
</tr>
<tr>
<td>+ Q/A type features</td>
<td>65.25</td>
<td>82.01</td>
<td>36.50</td>
<td>57.15</td>
</tr>
<tr>
<td>+ Question features</td>
<td>65.50</td>
<td>82.26</td>
<td>38.21</td>
<td>57.35</td>
</tr>
<tr>
<td>+ Image features</td>
<td>65.54</td>
<td>82.28</td>
<td>38.63</td>
<td>57.32</td>
</tr>
<tr>
<td>+ Explanation (SWAF)</td>
<td><strong>67.26</strong></td>
<td>82.62</td>
<td><strong>39.50</strong></td>
<td><strong>58.34</strong></td>
</tr>
</tbody>
</table>
Feature Ablation Analysis

(Rajani and Mooney, NAACL’18)
Takeaways

• Proposed four categories of auxiliary features:
  - Three can be inferred from the image-question pair.
  - Explanation generated from component systems.
• SOTA even with just 3 component systems.
• Explanation can be used to improve accuracy, not just gain human trust.

(Rajani and Mooney, NAACL’18)
Since Proposal

**NLP**
- Discourse
- Sentiment Analysis
- Entity Linking
- Language Modeling
- Relation Extraction
- Parsing

**Vision**
- Scene Recognition
- Object Tracking
- Fine-grained classification
- Object Detection
- Image Classification
- Image Captioning

**XAI**
- Rationalization
- Visual Explanations
- Textual Explanations
- Explanation Evaluation

- Stacking with Auxiliary Features for VQA (NAACL’18)
- Generating and Evaluating Visual Explanations (ViGIL’17) (Under review at NIPS)
Since Proposal

**NLP**
- Discourse
- Sentiment Analysis
- Entity Linking
- Language Modeling
- Relation Extraction
- Parsing

**Vision**
- Scene Recognition
- Object Tracking
- Fine-grained classification
- Object Detection
- Image Classification

**XAI**
- Rationalization
- Visual Explanations
- Textual Explanations
- Explanation Evaluation

**VQA**
- Image Captioning

*Stacking with Auxiliary Features for VQA (NAACL’18)*

*Generating and Evaluating Visual Explanations (ViGIL’17)*

(Under review at NIPS)
Explainable AI (XAI)

- Generate explanations for an ensemble.
- Evaluate explanations.
Visual Explanation for Ensembles

• Current VQA systems are complex DNNs that are *opaque* and can make odd mistakes, decreasing trustworthiness.
• Visual explanations can help make their reasoning more *transparent*.
• Ensembling VQA systems produces better results but further complicates explaining their results.
• Visual explanations for ensemble models also improves explanation quality over those of the individual component models.
Visual Explanations for Ensemble Models

• Explain a complex VQA ensemble by ensembling the visual explanations of its component systems.

• Ensembling visual explanation methods:
  1. Weighted Average (WA)
  2. Penalized Weighted Average (PWA)
Weighted Average (WA) Approach

- **Average** the explanatory heat-maps of systems that *agree* with the ensemble, weighted by their performance on validation data.
- E - explanation map of ensemble
- \(A^k\) - explanation map of \(k^{th}\) component model
- \(w_k\) - weight of the component model
- t - thresholding parameter

\[
E_{i,j} = \begin{cases} 
\frac{1}{|K|} \sum_{k \in K} w_k A^k_{i,j}, & \text{if } A^k_{i,j} \geq t \\
0, & \text{otherwise}
\end{cases}
\]

subject to \(\sum_{k \in K} w_k = 1\)

(Rajani and Mooney, ViGIL’17 & BC)
WA Example

Q: What color is the umbrella? A: Yellow

\[
\frac{1}{3} \begin{bmatrix}
W_1 \\
+ W_2 \\
+ W_3 \\
\end{bmatrix}
\]

LSTM  + HieCoAtt  + MCB  = Ensemble

(Rajani and Mooney, ViGIL’17 & BC)
Penalized Weighted Average (PWA) Approach

- Complimentary to WA.
- Subtract the explanatory heat-maps of systems that disagree with the ensemble.
- $I^m$ - explanation map of $m^{th}$ model that disagrees.

$$E_{i,j} = \begin{cases} 
\frac{1}{|K|} \sum_{k \in K} \sum_{m \in M} w_k A_{i,j}^k - w_m I_{i,j}^m, & \text{if } p \geq t \\
0, & \text{otherwise}
\end{cases}$$

subject to $\sum_{k \in K} w_k + \sum_{m \in M} w_m = 1$

(Rajani and Mooney, ViGIL’17 & BC)
**PWA Example**

Q: The car in front of the train is what color?  
A: Red

HieCoAtt, MCB answer: red and LSTM answer: white

(Rajani and Mooney, ViGIL’17 & BC)
PWA Example

Q: What direction are the giraffe looking? A: Right

LSTM, HieCoAtt answer: right and MCB answer: left

(Rajani and Mooney, ViGIL’17 & BC)
Crowd-sourced Hyper-parameter Tuning

- We used crowd-sourcing to determine the value of the threshold parameter $t$.
- The idea is to optimize the explanation map generation based on the evaluation metric.
- The human subjects were shown thresholded maps in steps of 0.05 in $[0.1,0.25]$ and asked to choose the one that highlighted the most appropriate regions.
- For LSTM, MCB, WA and PWA: $t = 0.2$.
- For HieCoAtt: $t = 0.15$.

(Rajani and Mooney, ViGIL’17 & BC)
Crowd-sourced Hyper-parameter Tuning

(Rajani and Mooney, ViGIL'17 & BC)
Evaluating Visual Explanations

- Crowd-sourcing has been used for evaluation but metrics vary widely.
- Some metrics rely on human-generated explanations as gold standard.
- However, research shows that machines and humans do not have the same “view” of visual explanations (Das et al., 2017).
- We propose two novel metrics:
  - Comparison metric
  - Uncovering metric

(Rajani and Mooney, ViGIL’17 & BC)
Comparison Metric

- Human judges were shown two visual explanations (one was the ensemble and the other was an individual system) and asked: “Which picture highlights the part of the image that best supports the answer to the question?”
- Our ensemble explanation was judged better on an average 61% of the time compared to any individual system’s explanation.
Comparison Metric

Which picture highlights the part of the image that best supports the answer to the question?

- You are given a reference image along with a question and answer about the image.
- Below are two pictures that highlight parts of the reference image that support the answer to the question. Shades of red, orange, and yellow mean high intensity while shades of green and blue mean lower intensity in those regions of the image.
- Use your best judgment to pick one of the two pictures that highlights (accounting for intensity as well) the best-supporting evidence to the answer.
- A picture that highlights more of the wrong regions of an image is worse than a picture that highlights less of the wrong regions.
- Only if you cannot decide between the two images, please choose Option C: cannot decide.
- If you think the answer to the question is incorrect, please choose Option D: incorrect answer.
- For the following example, the left image is clearly a better option than the right image because, although both highlight regions of the image that support the answer with high intensity, the right one also highlights other irrelevant regions of the image with high intensity.

Question: Which foot is not on the ground? Answer: right.

<table>
<thead>
<tr>
<th>Reference Image</th>
<th>Question: Is there a giraffe eating? Answer: yes</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Reference Image" /></td>
<td><img src="image2" alt="Reference Image" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Option A</th>
<th>Option B</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="optionA" alt="Option A" /></td>
<td><img src="optionB" alt="Option B" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Option C</th>
<th>Option D</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="optionC" alt="Option C" /></td>
<td><img src="optionD" alt="Option D" /></td>
</tr>
</tbody>
</table>

Submit
## Comparison Results

(Rajani and Mooney, ViGIL’17 & BC)

<table>
<thead>
<tr>
<th>Approach</th>
<th>Ensemble</th>
<th>Single System</th>
<th>Cannot decide</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ensemble (WA)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LSTM</td>
<td>58</td>
<td>36</td>
<td>3</td>
</tr>
<tr>
<td>HieCoAtt</td>
<td>62</td>
<td>27</td>
<td>6</td>
</tr>
<tr>
<td>MCB</td>
<td>52</td>
<td>41</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Ensemble (PWA)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LSTM</td>
<td>64</td>
<td>28</td>
<td>3</td>
</tr>
<tr>
<td>HieCoAtt</td>
<td>69</td>
<td>26</td>
<td>1</td>
</tr>
<tr>
<td>MCB</td>
<td>61</td>
<td>35</td>
<td>1</td>
</tr>
</tbody>
</table>
Uncovering Metric

- Human judges were shown partially uncovered images that only show the part of the image highlighted in the explanation.
- Uncover 1/3, 2/3, or all of the “hottest” part of the explanation map for an image.
- Measure for what percentage of the test cases a human judge decided they were able to answer the question from the partial image, and then picked the correct answer.

  - Our ensemble explanation allowed judges to correctly answer more questions at least 64% of the time when shown such partially covered images compared to any individual system’s explanation.

(Rajani and Mooney, ViGIL’17 & BC)
Uncovering Evaluation


(Rajani and Mooney, ViGIL’17 & BC)
Uncovering Evaluation


(Rajani and Mooney, ViGIL’17 & BC)
Uncovering Results

<table>
<thead>
<tr>
<th>System</th>
<th>One-third</th>
<th>Two-thirds</th>
<th>Entire map</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ensemble (PWA)</td>
<td>29</td>
<td>35</td>
<td>69</td>
</tr>
<tr>
<td>Ensemble (WA)</td>
<td>17</td>
<td>28</td>
<td>64</td>
</tr>
<tr>
<td>LSTM</td>
<td>10</td>
<td>22</td>
<td>42</td>
</tr>
<tr>
<td>HieCoAtt</td>
<td>9</td>
<td>19</td>
<td>38</td>
</tr>
<tr>
<td>MCB</td>
<td>11</td>
<td>20</td>
<td>46</td>
</tr>
</tbody>
</table>
Normalized Uncovering

- Uncovering fractions of the explanation does not normalize for the number of pixels revealed, so different systems may uncover different fractions of the overall image.
- Uncover 1/4, 1/2, or 3/4 of the entire image instead.
- Randomly choose zero-weight pixels as needed, resulting in “snowy” images.

(Rajani and Mooney, ViGIL’17 & BC)
Normalized Uncovering Evaluation

Q: What color is the bear?  
## Normalized Uncovering Results

(Rajani and Mooney, ViGIL’17 & BC)

<table>
<thead>
<tr>
<th>System</th>
<th>One-fourth</th>
<th>One-half</th>
<th>Three-fourths</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ensemble (PWA)</td>
<td>23</td>
<td>38</td>
<td>76</td>
</tr>
<tr>
<td>Ensemble (WA)</td>
<td>21</td>
<td>34</td>
<td>71</td>
</tr>
<tr>
<td>LSTM</td>
<td>10</td>
<td>24</td>
<td>65</td>
</tr>
<tr>
<td>HieCoAtt</td>
<td>10</td>
<td>23</td>
<td>57</td>
</tr>
<tr>
<td>MCB</td>
<td>12</td>
<td>25</td>
<td>64</td>
</tr>
</tbody>
</table>
Takeaways

- Explanations provide useful insights into a model’s decision making process.
- We proposed the first approaches to generate visual explanations for ensembles of VQA models.
- Evaluating explanations is difficult especially when you can’t compare them to human-generated GT.
- Our ensemble explanations outperform individuals model’s explanations on both our proposed evaluation metrics:
  - Comparison - 61%
  - Uncovering - 64%
Future Directions
SWAF

• Extend SWAF on VQA to include *textual explanation* features.

• SWAF for actually combining structured o/p instead of casting the structured o/p problem to a binary decision one.

• Extend SWAF to other classification and generation problems in NLP and vision.
  - Question Answering
  - Activity Recognition
XAI

• Generate *textual explanations* that are faithful to the model.

• Ensemble textual explanations to serve as explanation for the ensemble.
  - Challenging but can adopt ideas from MT.

• Use textual explanation as auxiliary features.
  - Measure similarity using MT metrics.

• Combine textual and visual explanations.

• Better evaluation metrics.
Combining Visual and Textual Explanations

• Find natural-language concepts found in sub-regions of the image that contributed to the system’s answer using *network dissection* (Bau *et al.*, 2017).
• Combine these concepts into a coherent explanatory sentence.
• Produce a joint visual and textual explanation where NL concepts in the sentence point to corresponding regions in the image.
Joint Visual and Textual Explanations

This is a closet since it has a *shirt*, a *handbag* and *shoes*.
Conclusion
Conclusion

• General problem of combining outputs from diverse systems.
• SWAF produced significant improvements on NLP and Vision tasks.
• Explanation for improving performance of VQA.
• Ensemble system’s visual explanation is significantly better than single system’s on two novel evaluation metrics.
• Future directions:
  - SWAF
  - XAI
Acknowledgements
Thank You!


Backup Slides
WA Example (forced version)

Q: The car in front of the train is what color? A: Red

\[
\begin{bmatrix}
\frac{1}{3} & \text{forced LSTM} & + \\
\frac{1}{3} & \text{HieCoAtt} & + \\
\frac{1}{3} & \text{MCB} & =
\end{bmatrix}
\]

(Rajani and Mooney, ViGIL’17 & BC)