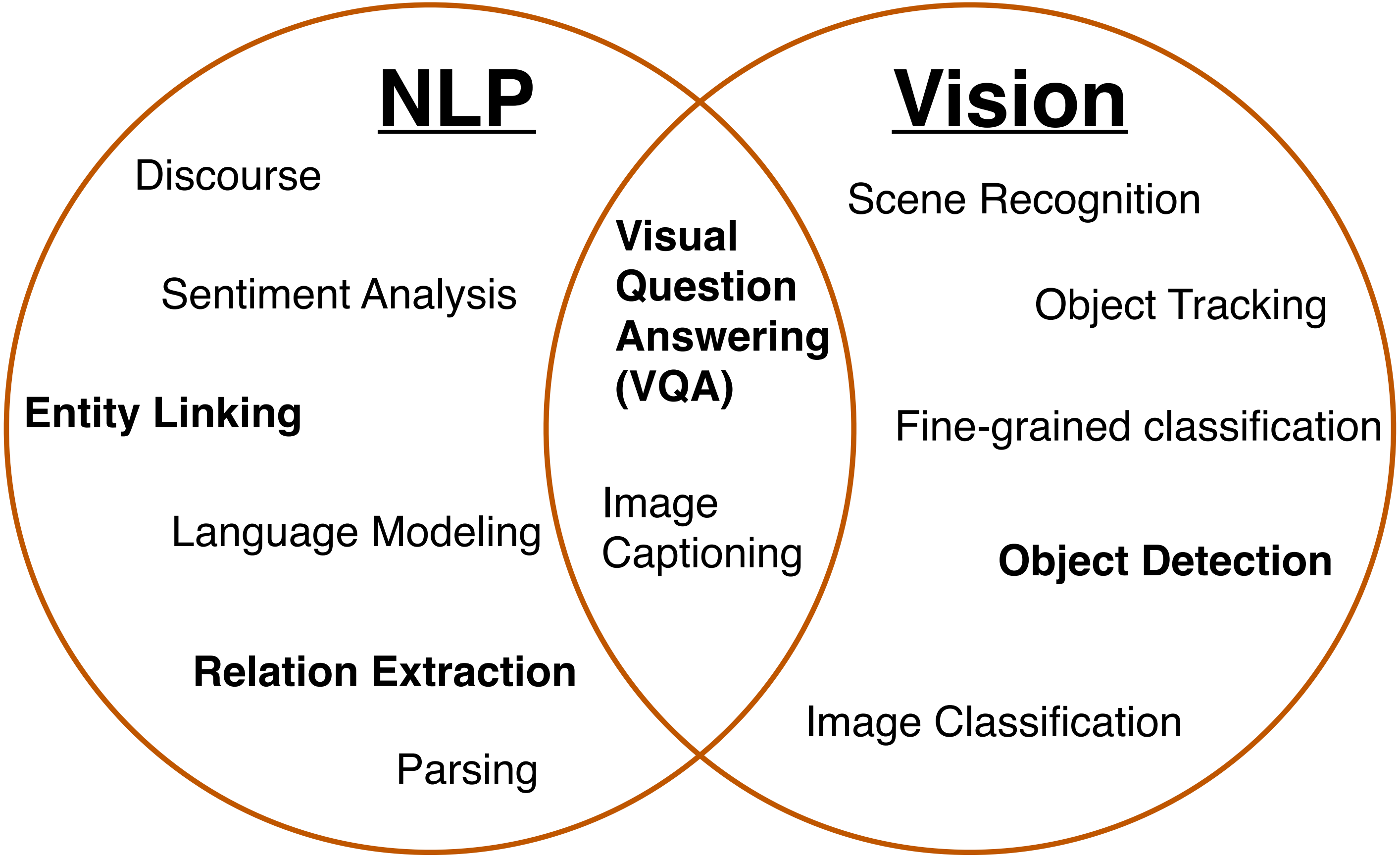
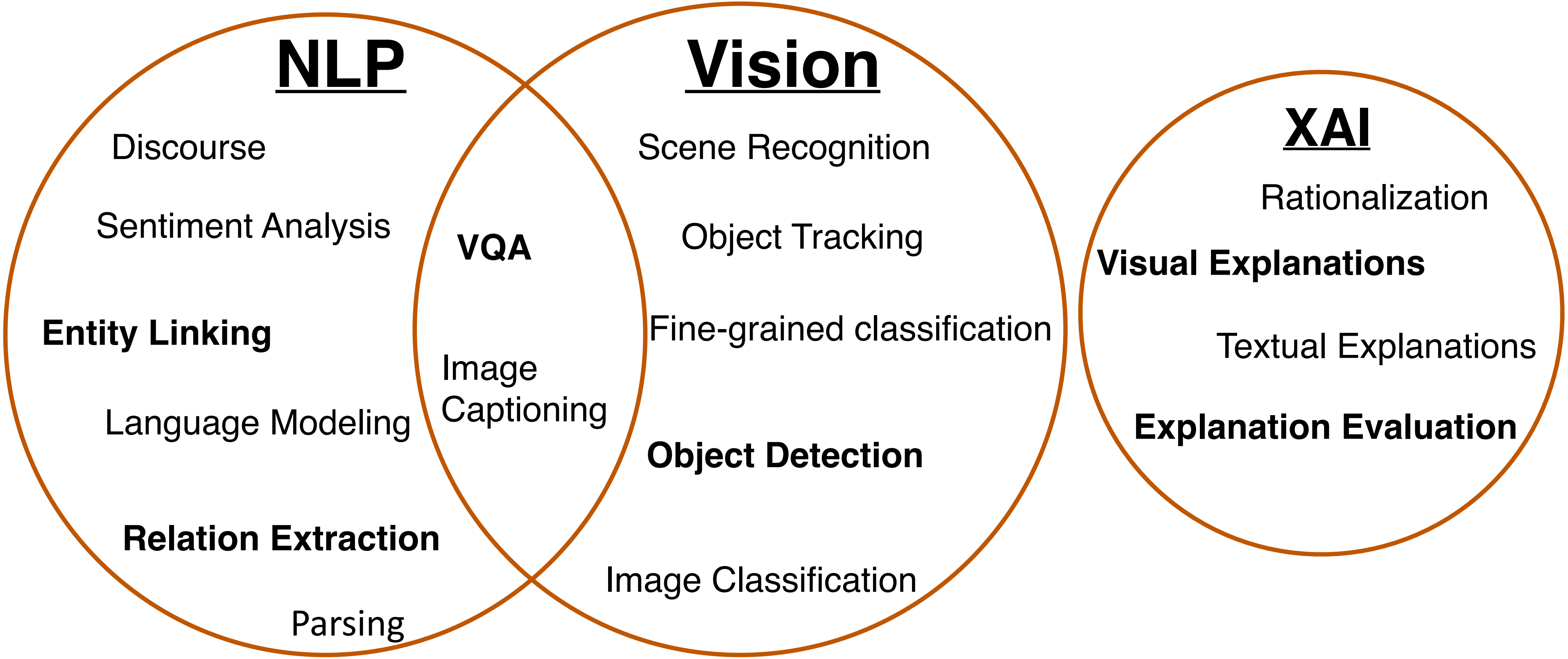
Explainable Improved Ensembling for Natural Language and Vision

Nazneen Rajani
University of Texas at Austin

Ph.D. Defense (12th July, 2018)


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

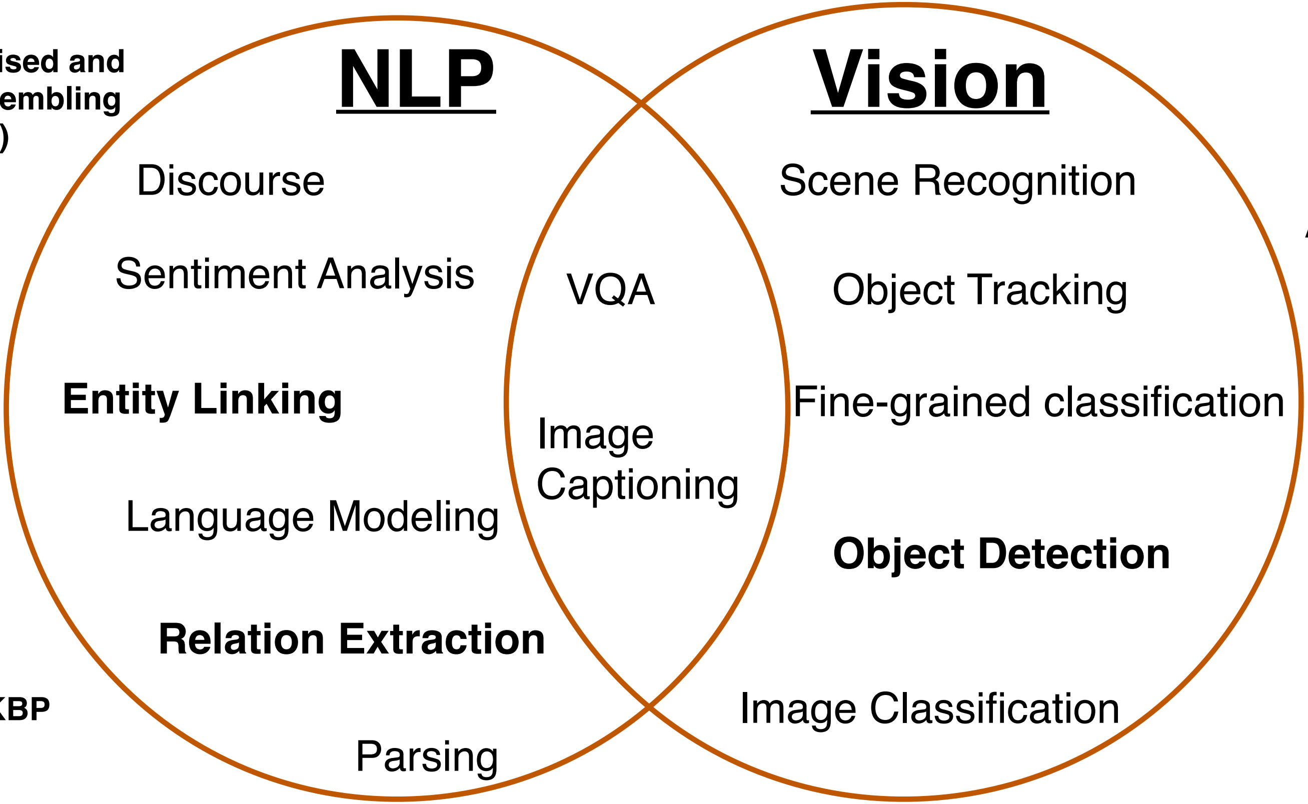

NLP
Discourse
Sentiment Analysis
Entity Linking
Language Modeling
Relation Extraction
Parsing
VQA
Image Captioning
Vision
Scene Recognition
Object Tracking
Fine-grained classification
Object Detection
Image Classification
XAI
Rationalization
Visual Explanations
Textual Explanations
Explanation Evaluation

My Research

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

Before Proposal

Combining supervised and Unsupervised Ensembling (EMNLP’16)



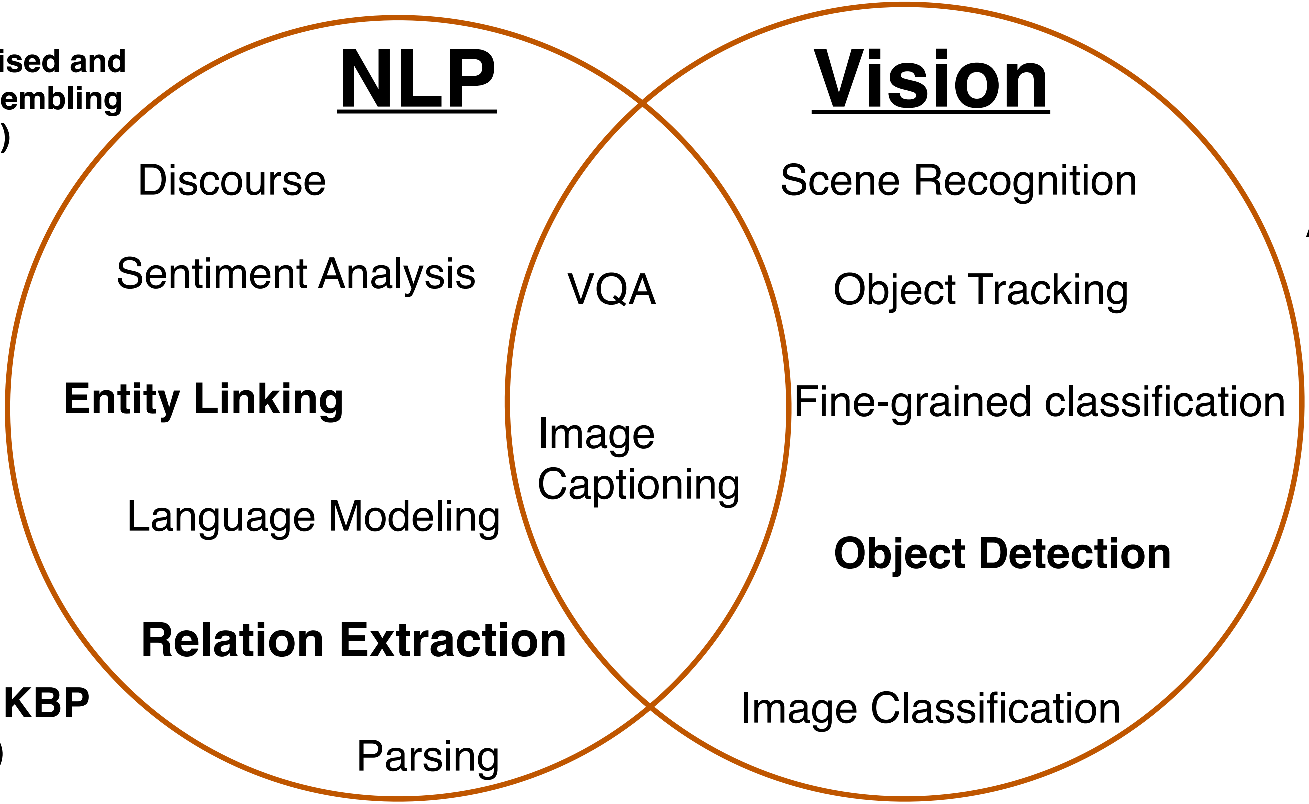
Stacking With Auxiliary Features (IJCAI’17)

Stacking for KBP (ACL’15)

Before Proposal

Combining supervised and Unsupervised Ensembling (EMNLP’16)

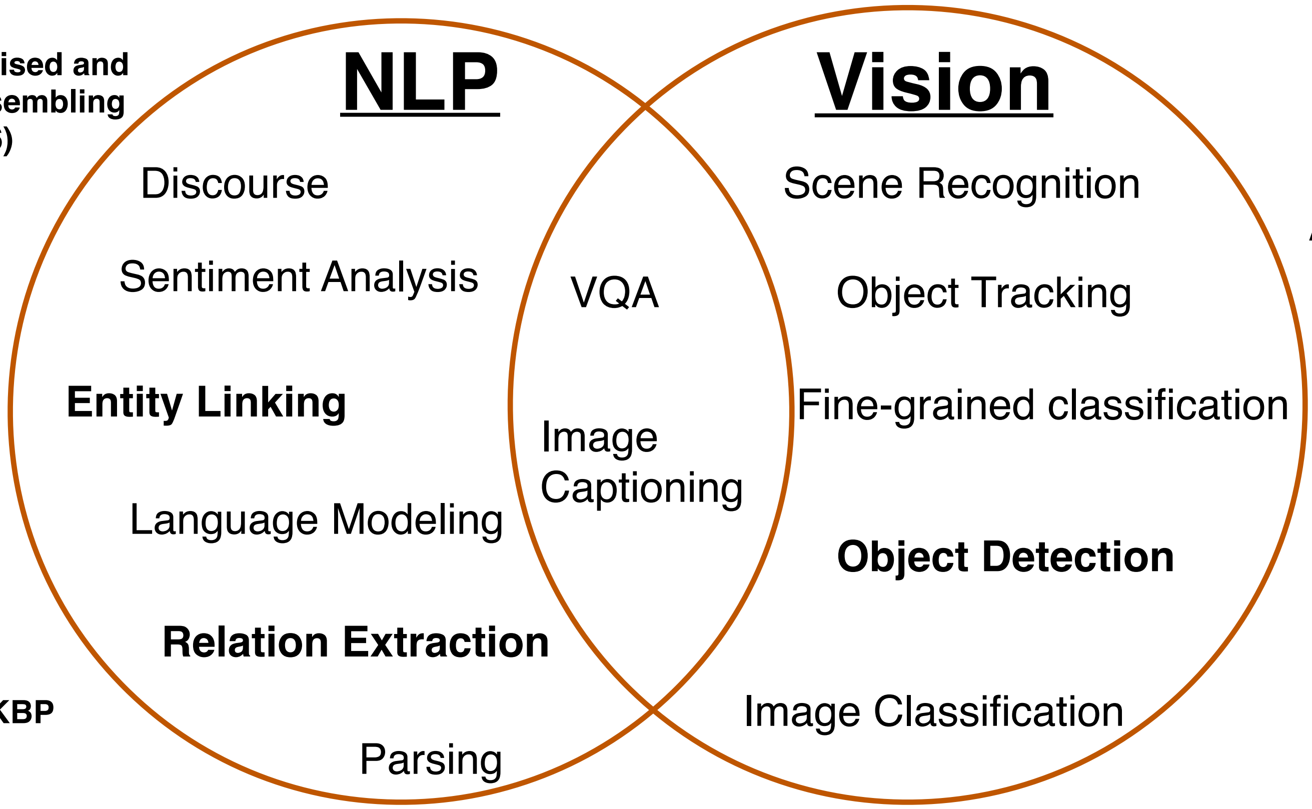
Stacking With Auxiliary Features (IJCAI’17)



Stacking for KBP (ACL’15)

Before Proposal

Combining supervised and Unsupervised Ensembling (EMNLP'16)

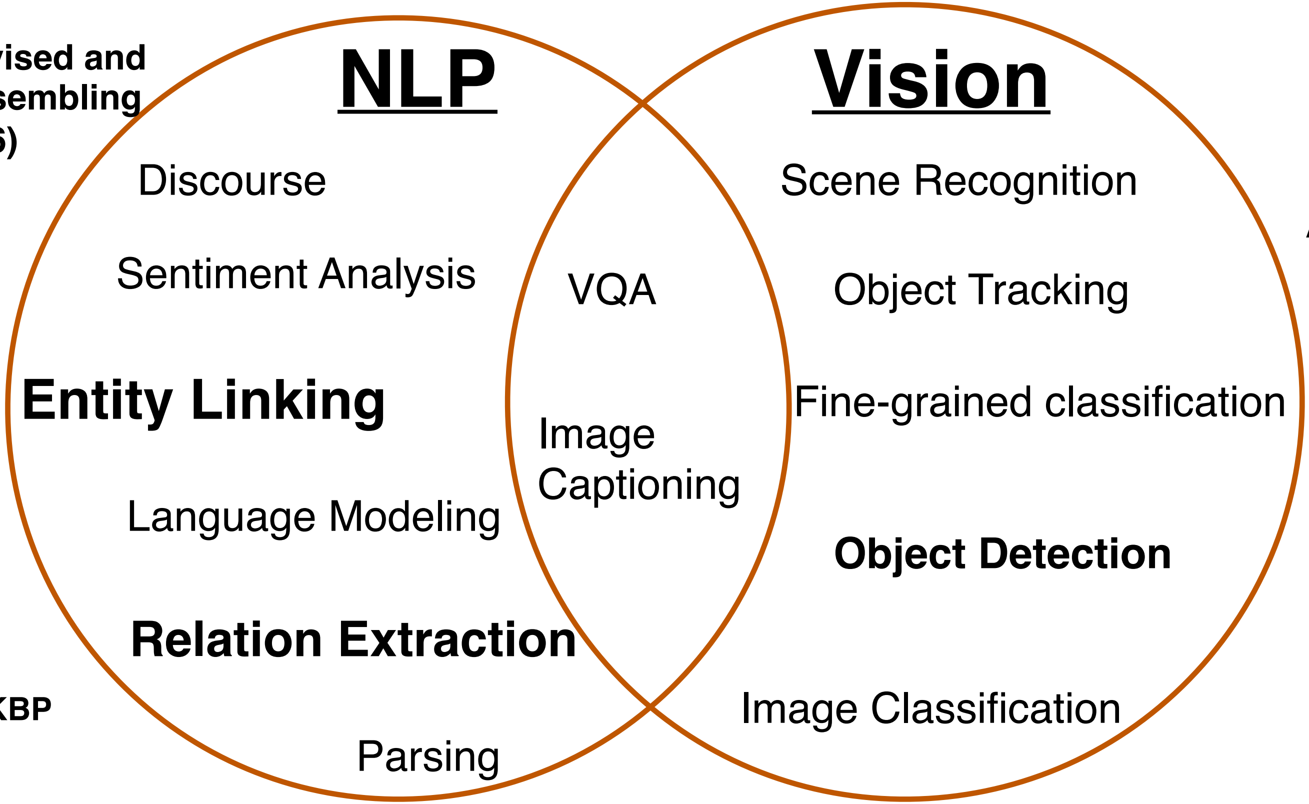


Stacking With Auxiliary Features (IJCAI'17)

Stacking for KBP (ACL'15)

Before Proposal

Combining supervised and Unsupervised Ensembling (EMNLP'16)



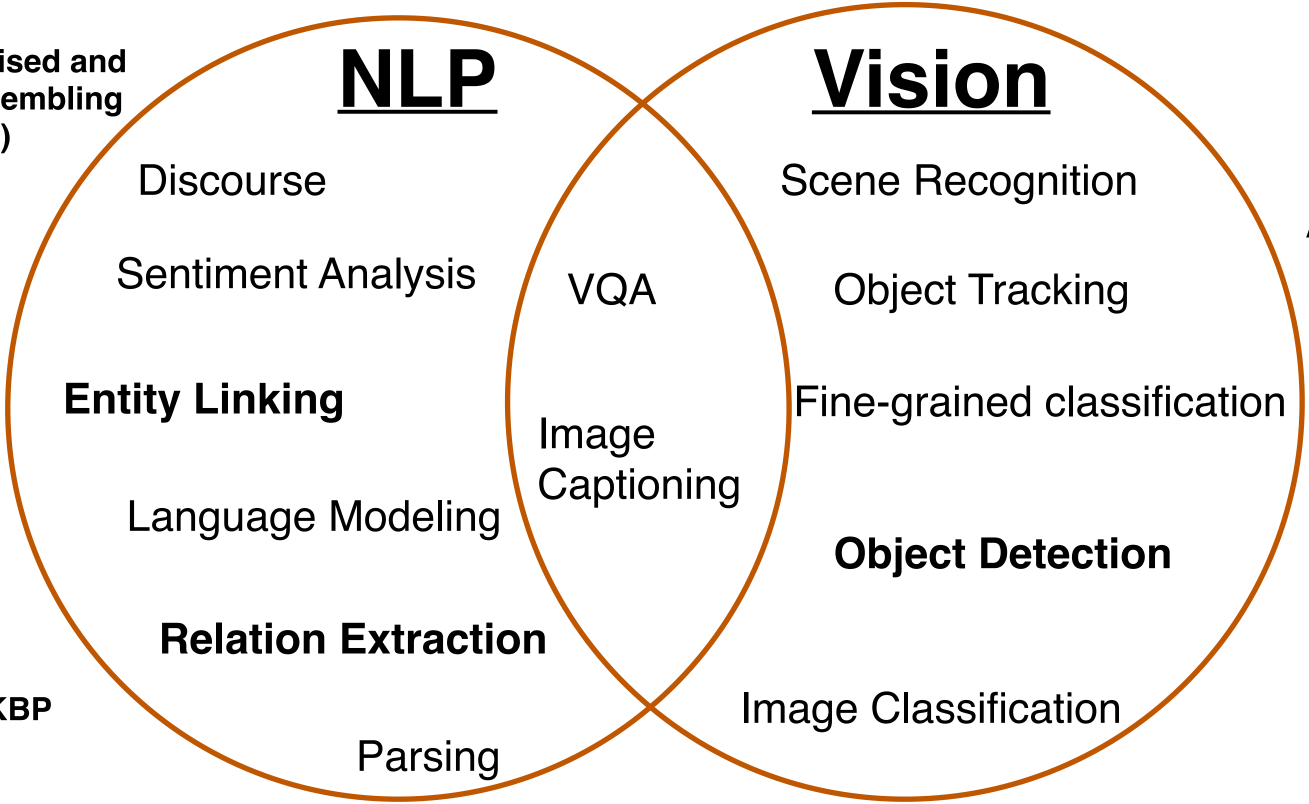
Stacking With Auxiliary Features (IJCAI'17)

Stacking for KBP (ACL'15)

Before Proposal

Combining supervised and Unsupervised Ensembling (EMNLP’16)

Stacking With Auxiliary Features (IJCAI’17)

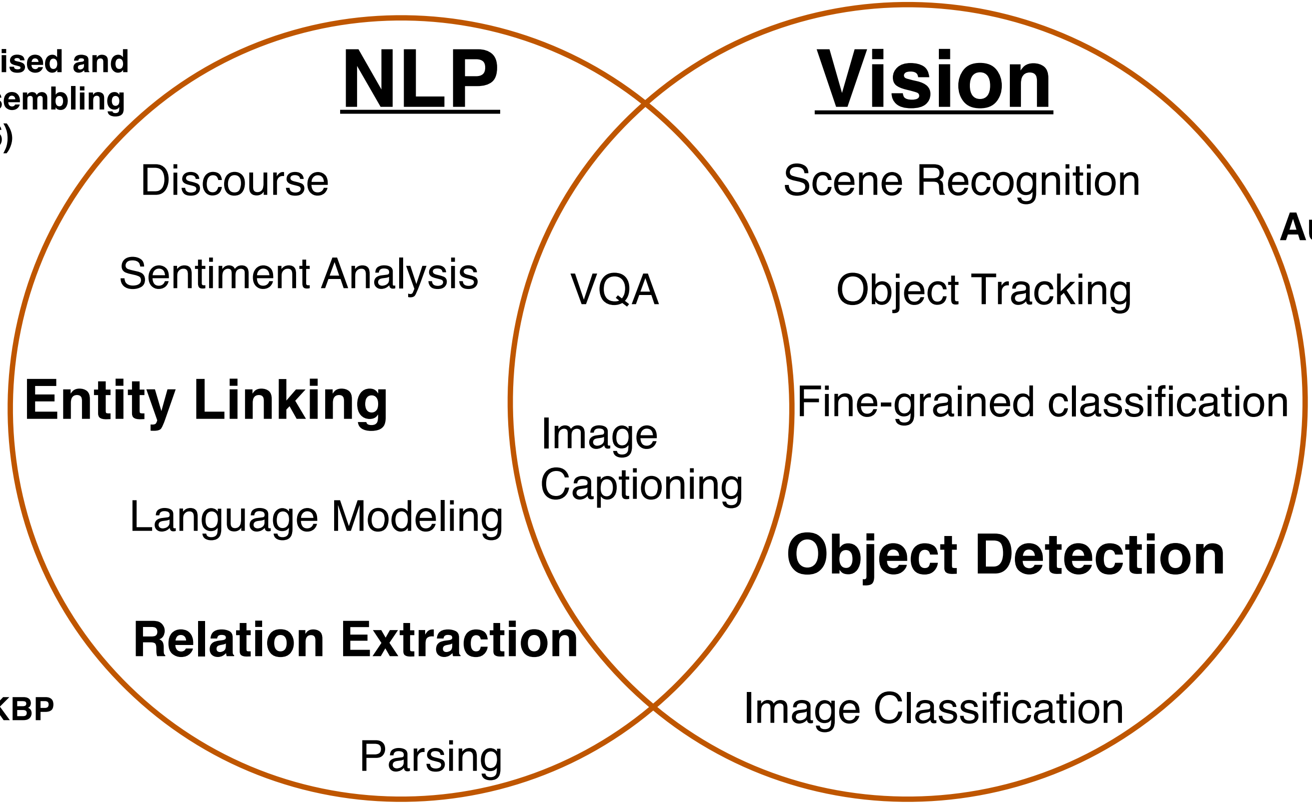


Stacking for KBP (ACL’15)

Before Proposal

Combining supervised and Unsupervised Ensembling (EMNLP'16)

Stacking With Auxiliary Features (IJCAI'17)



Stacking for KBP (ACL'15)

Since Proposal

NLP

Discourse

Sentiment Analysis

Entity Linking

Language Modeling

Relation Extraction

Parsing

VQA

Image Captioning

Vision

Scene Recognition

Object Tracking

Fine-grained classification

Object Detection

Image Classification

Stacking with Auxiliary Features for VQA (NAACL'18)

XAI

Rationalization

Visual Explanations

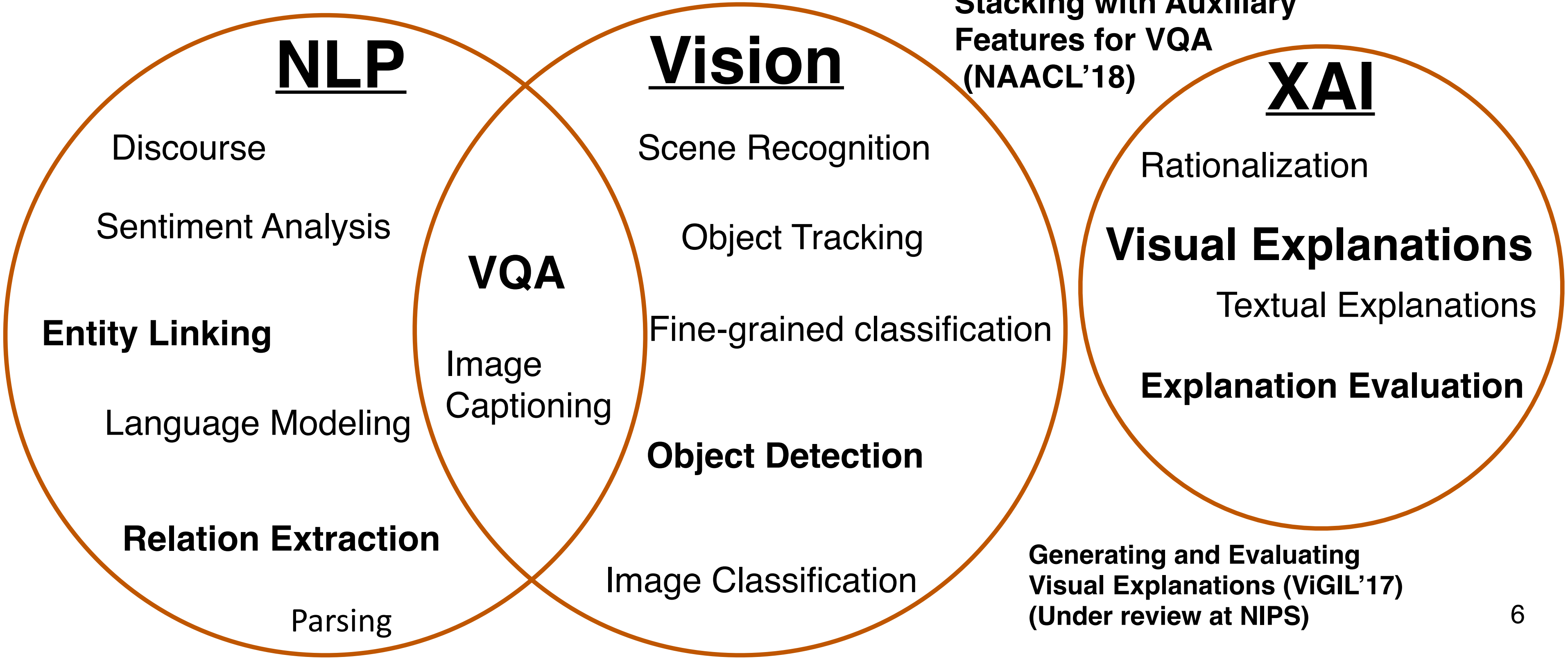
Textual Explanations

Explanation Evaluation

Generating and Evaluating Visual Explanations (ViGIL'17) (Under review at NIPS)

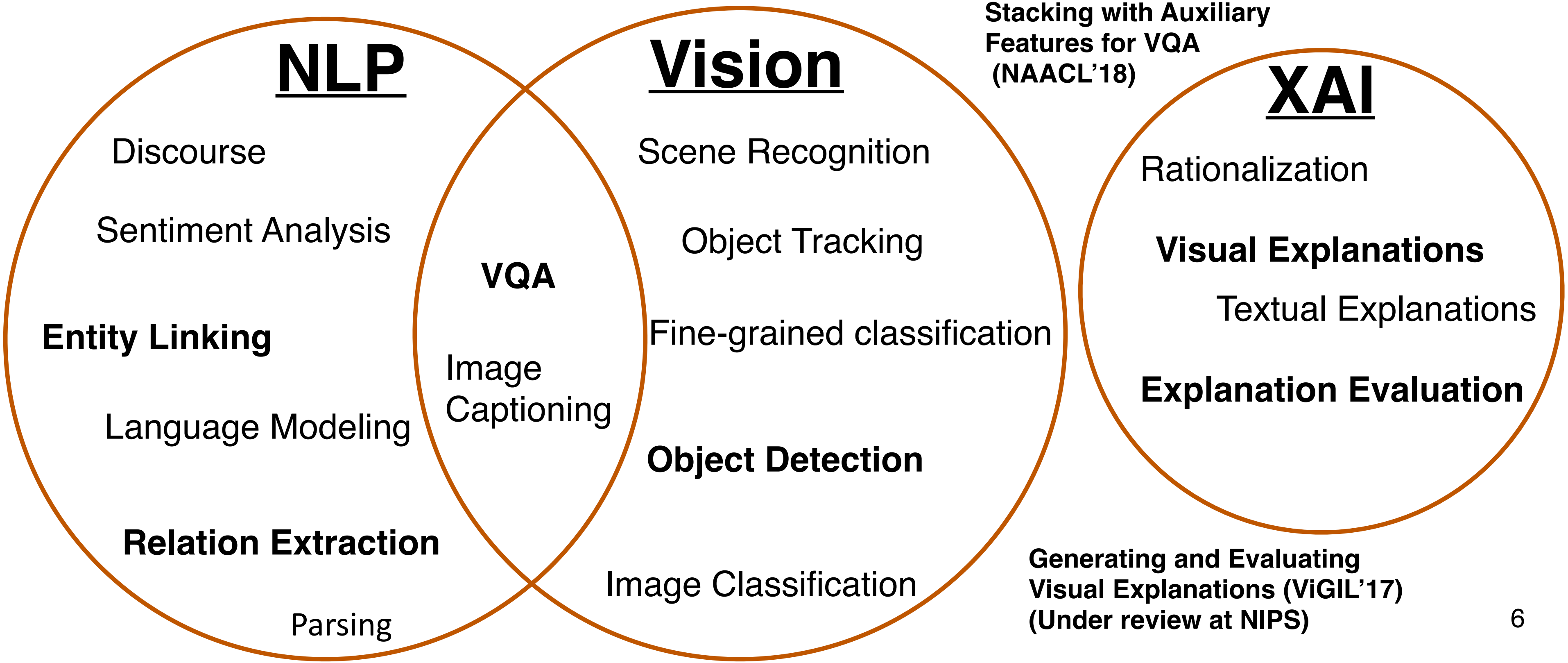
Since Proposal

Stacking with Auxiliary Features for VQA (NAACL'18)



Generating and Evaluating Visual Explanations (ViGIL'17) (Under review at NIPS)

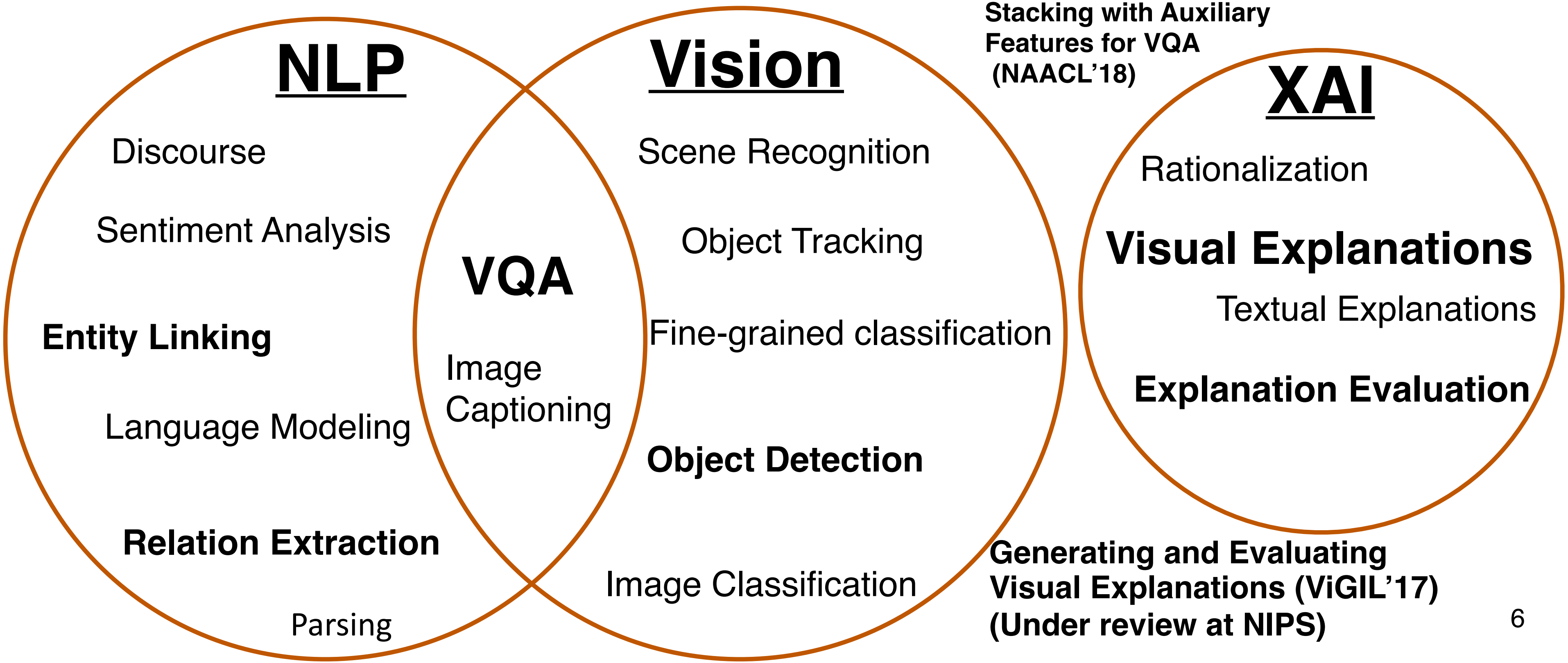
Since Proposal



Stacking with Auxiliary Features for VQA (NAACL'18)

Generating and Evaluating Visual Explanations (ViGIL'17) (Under review at NIPS)

Since Proposal

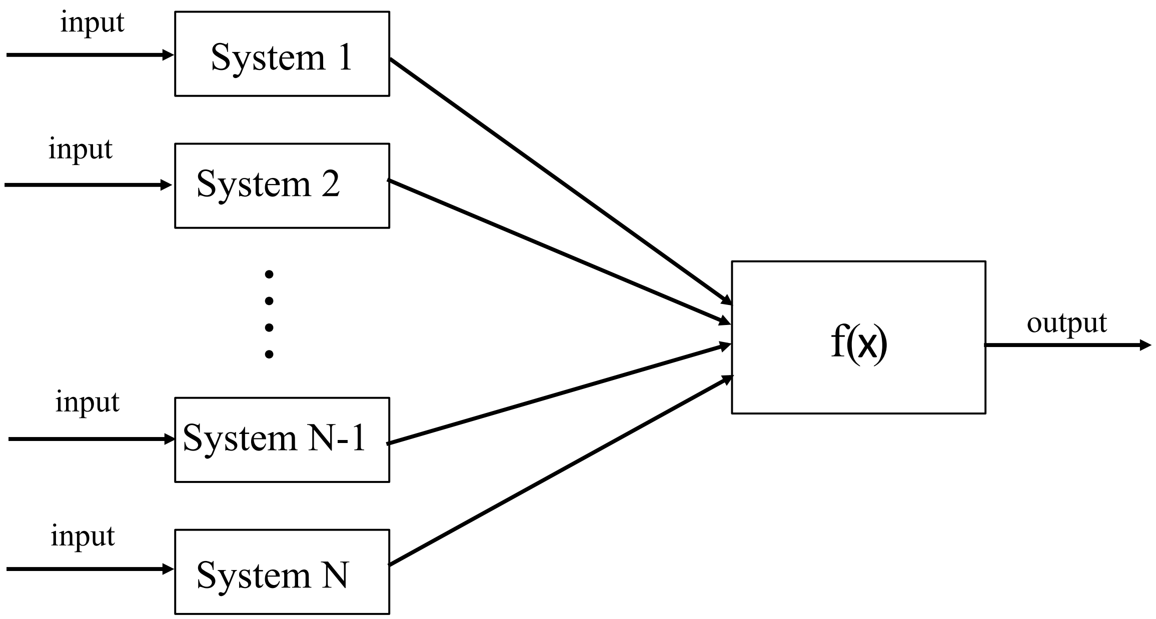


Stacking with Auxiliary Features for VQA (NAACL'18)

Generating and Evaluating Visual Explanations (ViGIL'17) (Under review at NIPS)

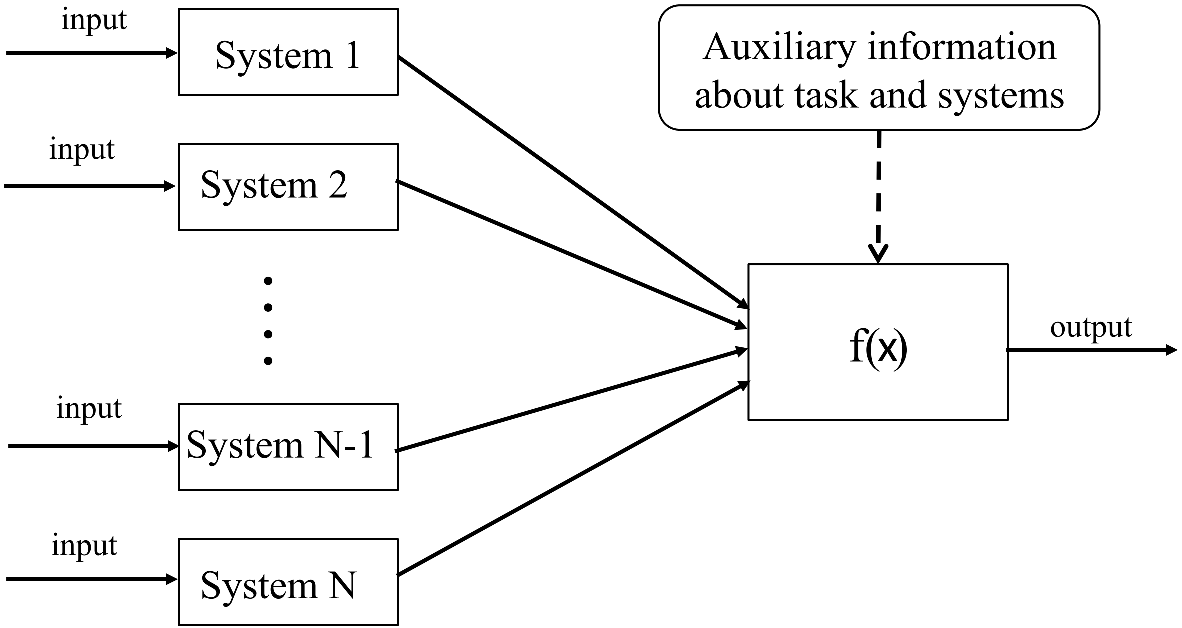
Ensembling

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



Ensembling

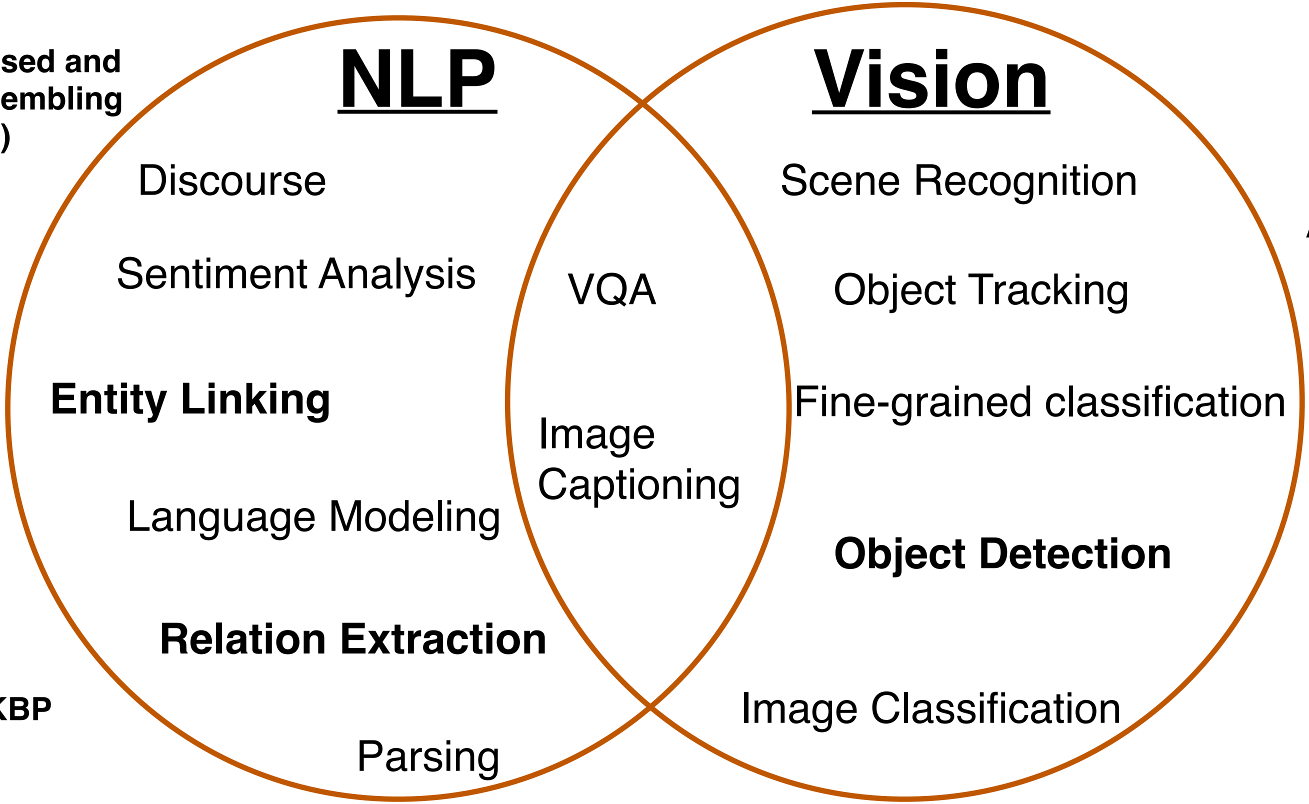
- Make **auxiliary** information accessible to the ensemble.



Before Proposal

Combined supervised and Unsupervised Ensembling (EMNLP'16)

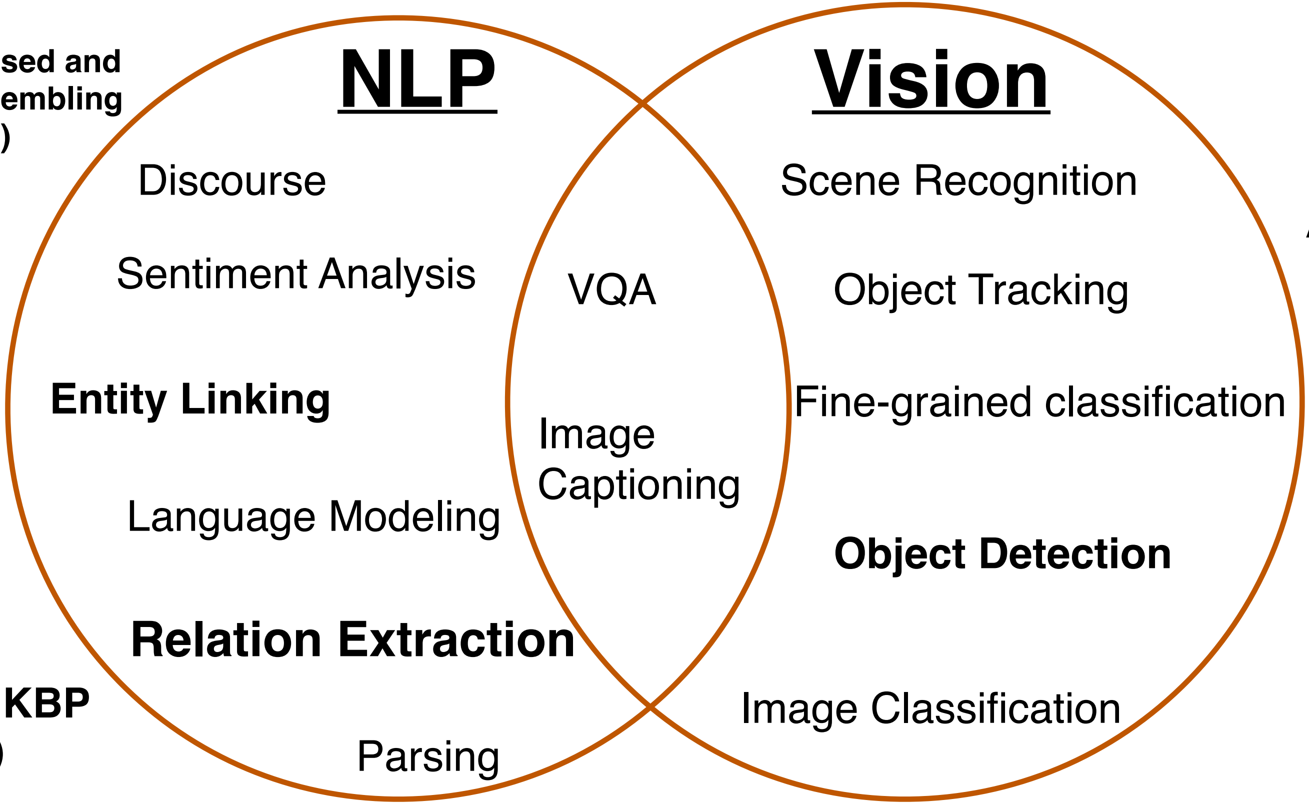
Stacking With Auxiliary Features (IJCAI'17)



Stacking for KBP (ACL'15)

Before Proposal

Combined supervised and Unsupervised Ensembling (EMNLP'16)



Stacking With Auxiliary Features (IJCAI'17)

Stacking for KBP (ACL'15)

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*

Slot-Filling

Unstructured web text → **Slot-Filling**

Microsoft is a technology company, headquartered in Redmond, Washington.

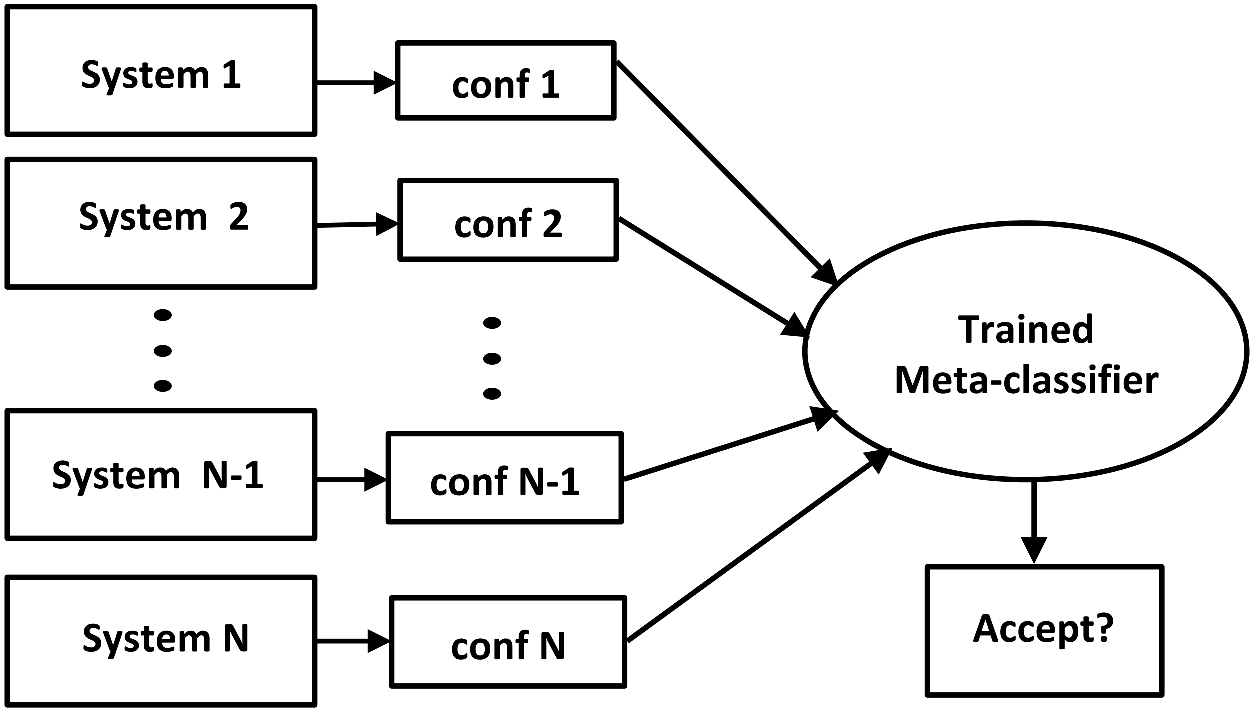
Microsoft was founded by Paul Allen and Bill Gates on April 4, 1975, to develop and sell BASIC interpreters for the Altair 8800.

org: Microsoft

city of headquarters	confidence
Redmond	1.0
founded by	**confidence**
Paul Allen	0.8
Bill Gates	0.95

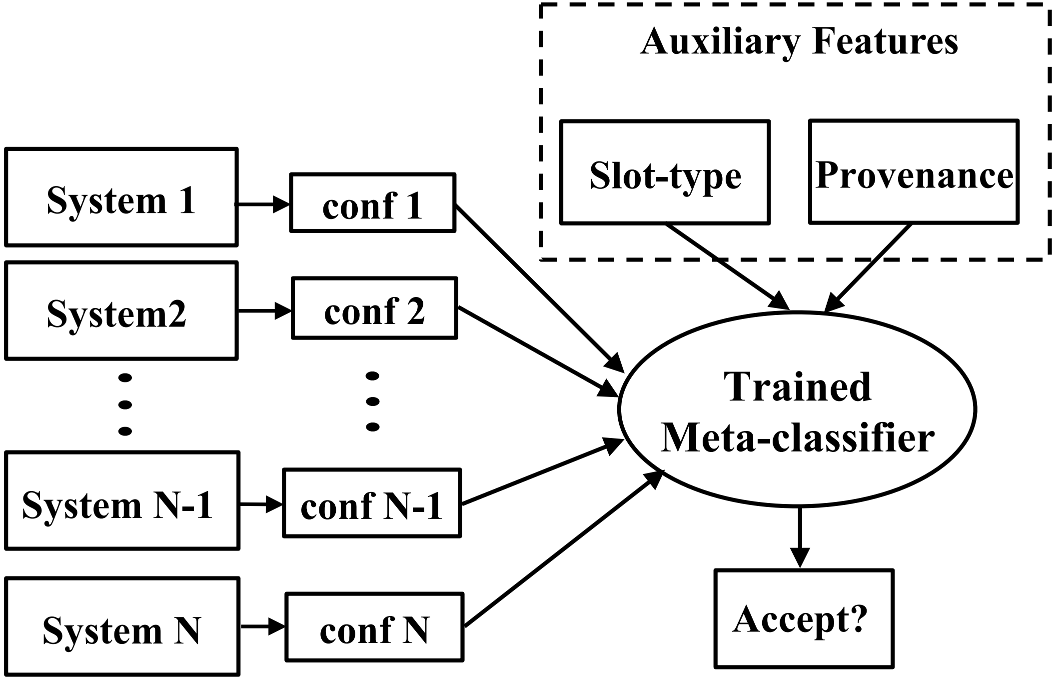
(Wolpert, 1992)

Stacking



(Viswanathan* et al., ACL'15)

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:

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

(Viswanathan* et al., ACL'15)

Offset Provenance

System 2

System 1

Former President Barack Obama

System 3

Offsets	System 1	System 2	System 3
Start offset	7	0	7
End Offset	28	15	22

$$OP_1 = \frac{1}{2}\left(\frac{9}{29} + \frac{16}{22}\right) \quad OP_2 = \frac{1}{2}\left(\frac{9}{29} + \frac{9}{23}\right) \quad OP_3 = \frac{1}{2}\left(\frac{16}{22} + \frac{9}{23}\right)$$

(Viswanathan* et al., ACL'15)

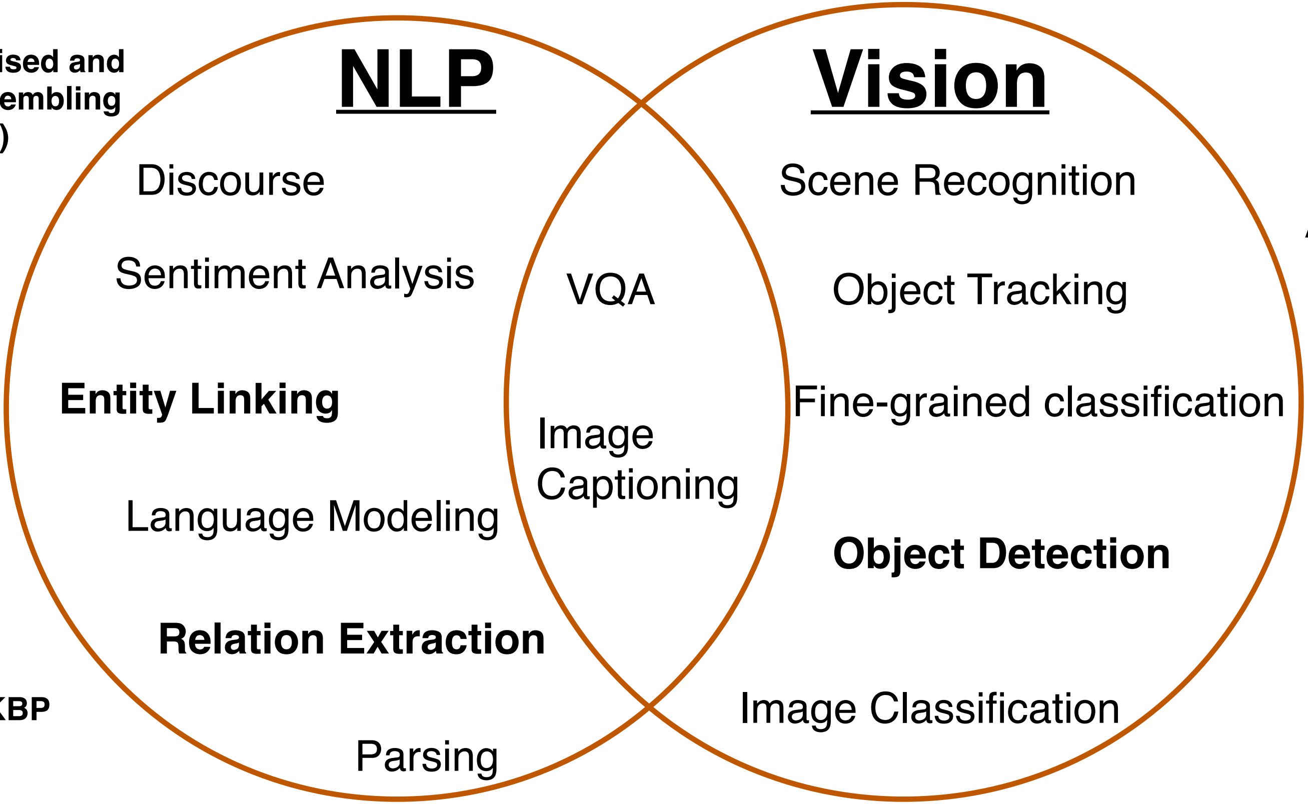
Slot-Filling Results

- 2014 KBP SF task— 10 shared systems

Approach	Precision	Recall	F1
Union	0.176	**0.647**	0.277
Voting	**0.694**	0.256	0.374
Best SF system in 2014 (Stanford)	0.585	0.298	0.395
Stacking	0.606	0.402	0.483
Stacking + Slot-type	0.607	0.406	0.486
Stacking + Provenance + Slot-type	0.541	0.466	**0.501**

Before Proposal

Combining supervised and Unsupervised Ensembling (EMNLP'16)

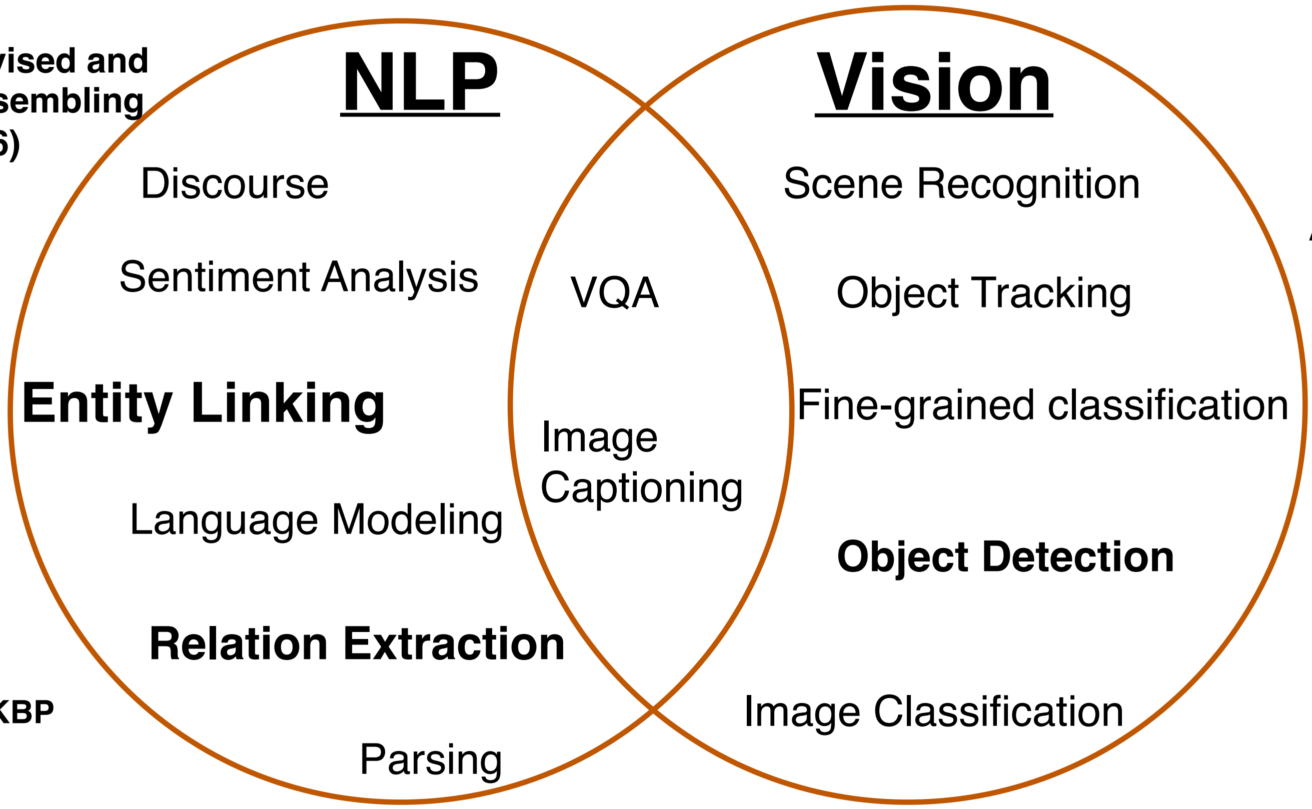


Stacking With Auxiliary Features (IJCAI'17)

Stacking for KBP (ACL'15)

Before Proposal

Combining supervised and Unsupervised Ensembling (EMNLP'16)



Stacking With Auxiliary Features (IJCAI'17)

Stacking for KBP (ACL'15)

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)

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:

William Jefferson "Bill" Clinton is an American politician who served as the 42nd President of the United States from 1993 to 2001. Clinton was Governor of Arkansas from 1979 to 1981 and 1983 to 1992, and Arkansas Attorney General from 1977 to 1979.

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:

William Jefferson "Bill" Clinton is an American politician who served as the 42nd President of the United States from 1993 to 2001. Clinton was Governor of Arkansas from 1979 to 1981 and 1983 to 1992, and Arkansas Attorney General from 1977 to 1979.

Entity Discovery and Linking (EDL)

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

FreeBase entry:

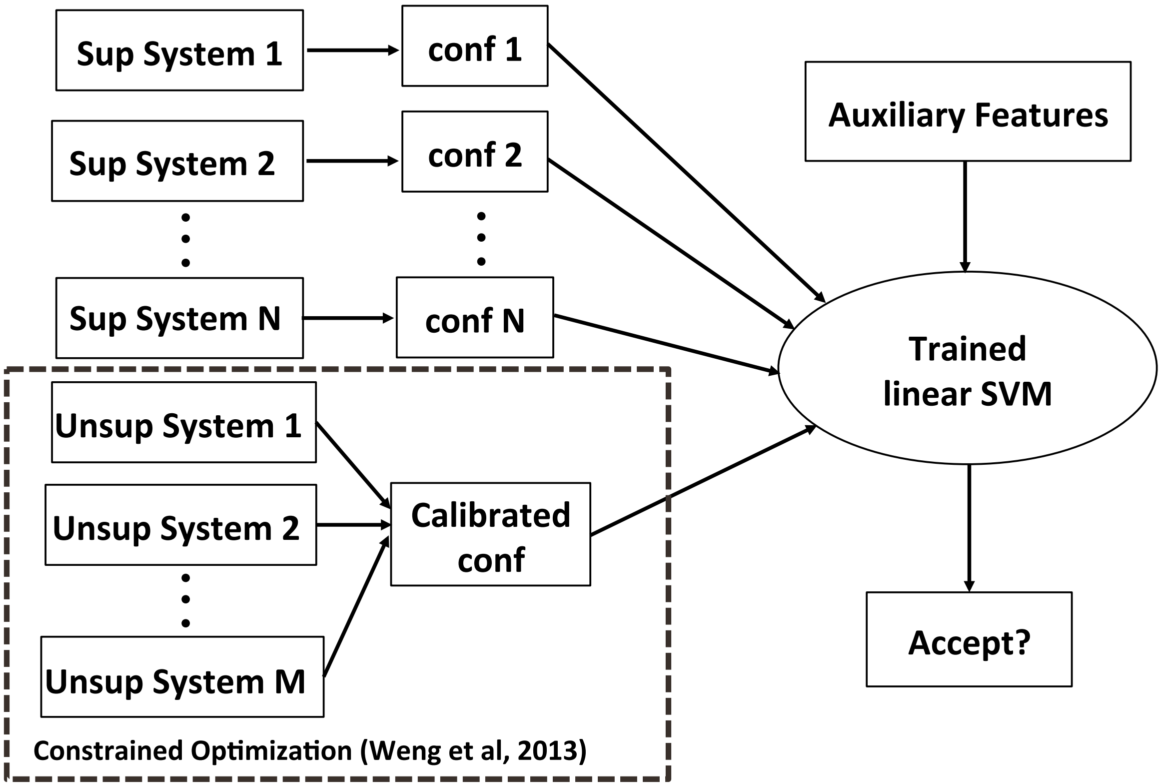
Hillary Diane Rodham Clinton is a US Secretary of State, U.S. Senator, and First Lady of the United States. From 2009 to 2013, she was the 67th Secretary of State, serving under President Barack Obama. She previously represented New York in the U.S. Senate.

FreeBase entry:

William Jefferson "Bill" Clinton is an American politician who served as the 42nd President of the United States from 1993 to 2001. Clinton was Governor of Arkansas from 1979 to 1981 and 1983 to 1992, and Arkansas Attorney General from 1977 to 1979.

(Rajani and Mooney, EMNLP'16)

Combining supervised & unsupervised ensembles



(Wang et al., 2013)

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.

Results

(Rajani and Mooney, EMNLP'16)

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

Approach	Precision	Recall	F1
Constrained optimization	0.1712	0.3998	0.2397
Oracle voting (>=3)	0.4384	0.2720	0.3357
Top ranked system (Angeli et al., 2015)	0.3989	0.3058	0.3462
Stacking + slot-type + provenance	0.4656	0.3312	0.3871
Stacking for combining sup + unsup (constrained optimization)	0.4676	**0.4314**	**0.4489**

Results

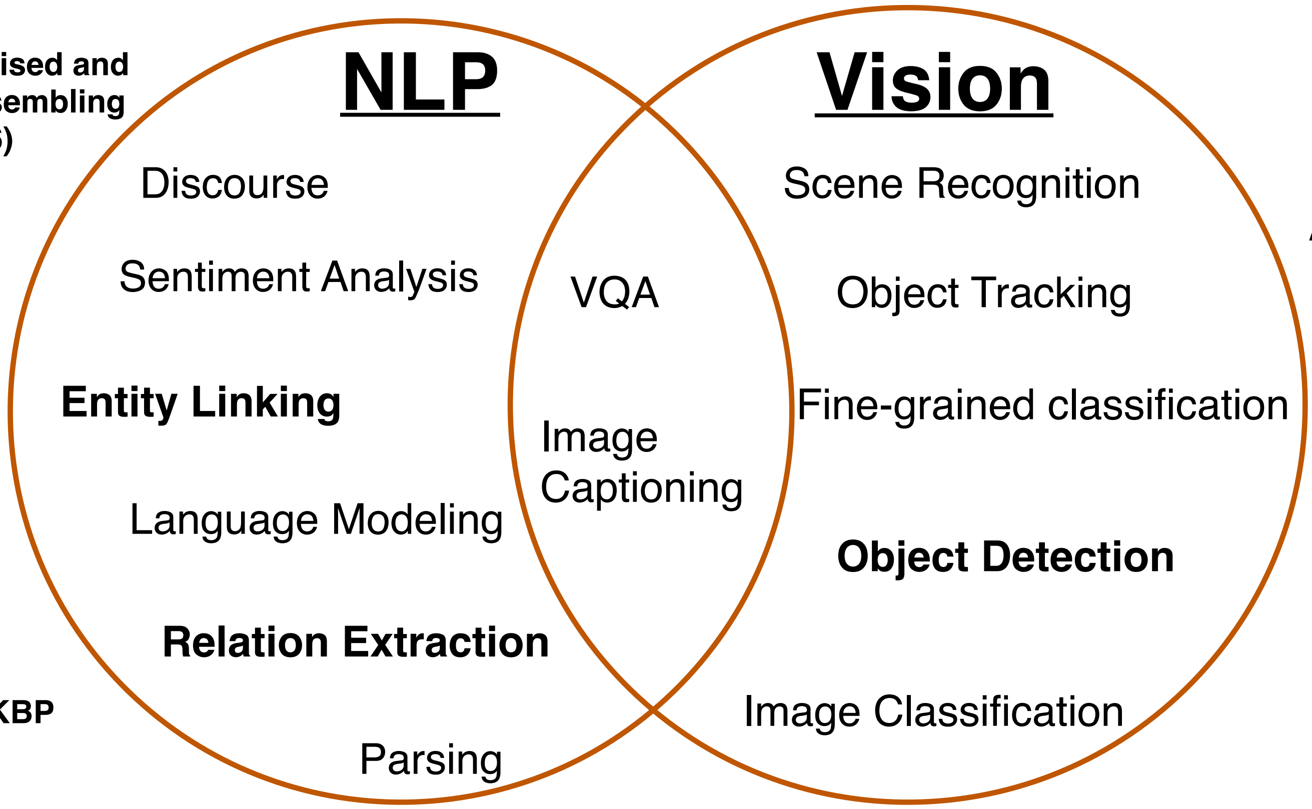
(Rajani and Mooney, EMNLP’16)

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

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

Before Proposal

Combining supervised and Unsupervised Ensembling (EMNLP'16)



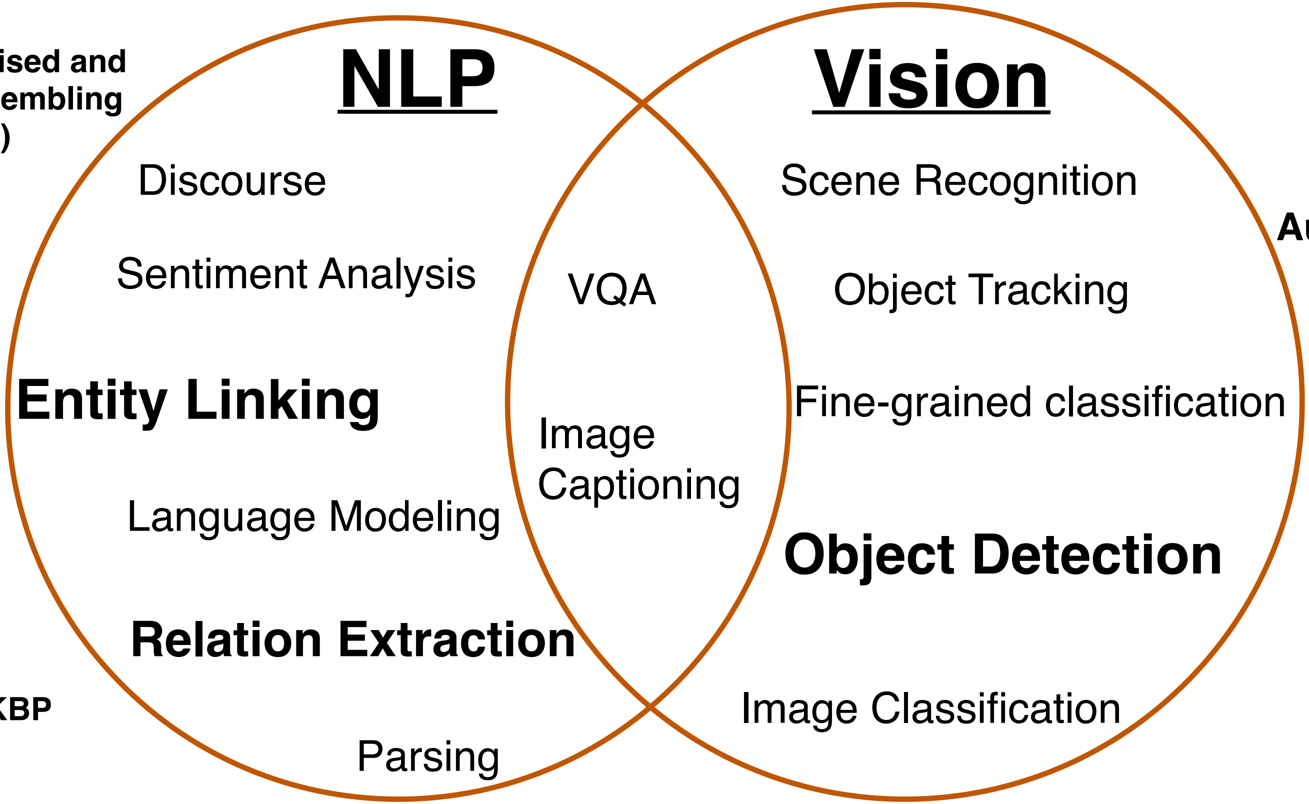
Stacking With Auxiliary Features (IJCAI'17)

Stacking for KBP (ACL'15)

Before Proposal

Combining supervised and Unsupervised Ensembling (EMNLP’16)

Stacking With Auxiliary Features (IJCAI’17)

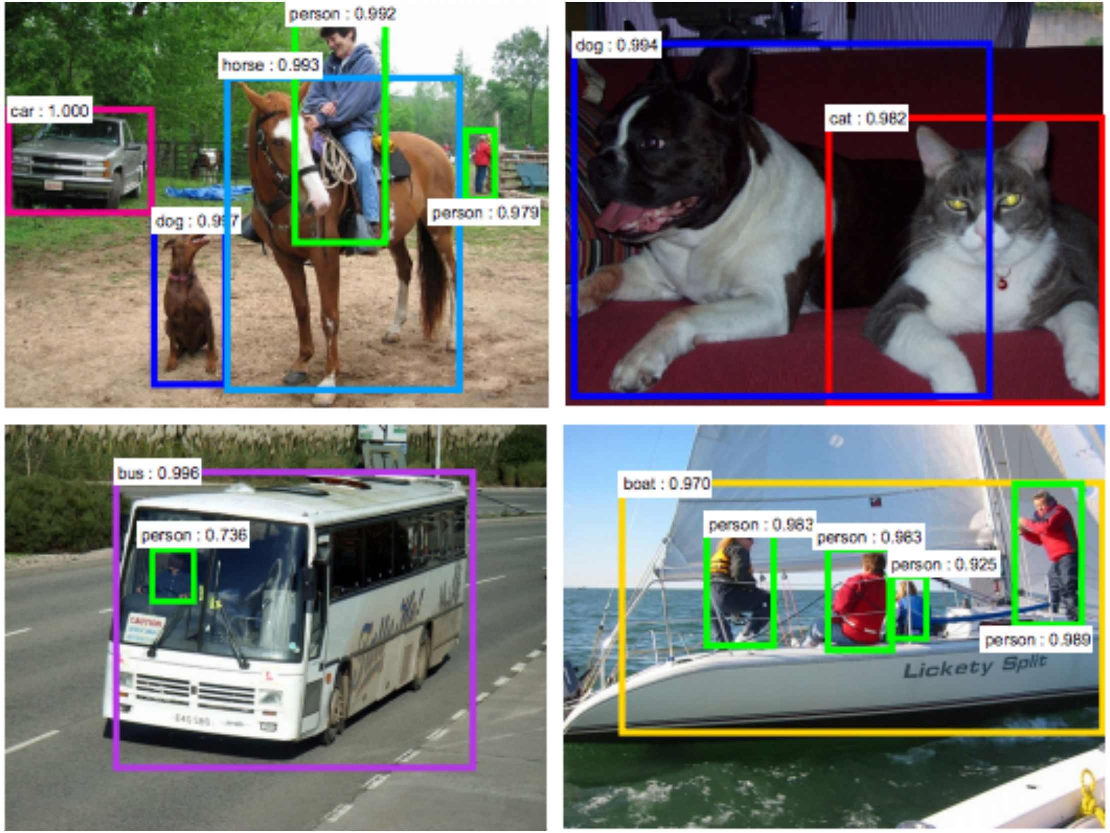


Stacking for KBP (ACL’15)

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).

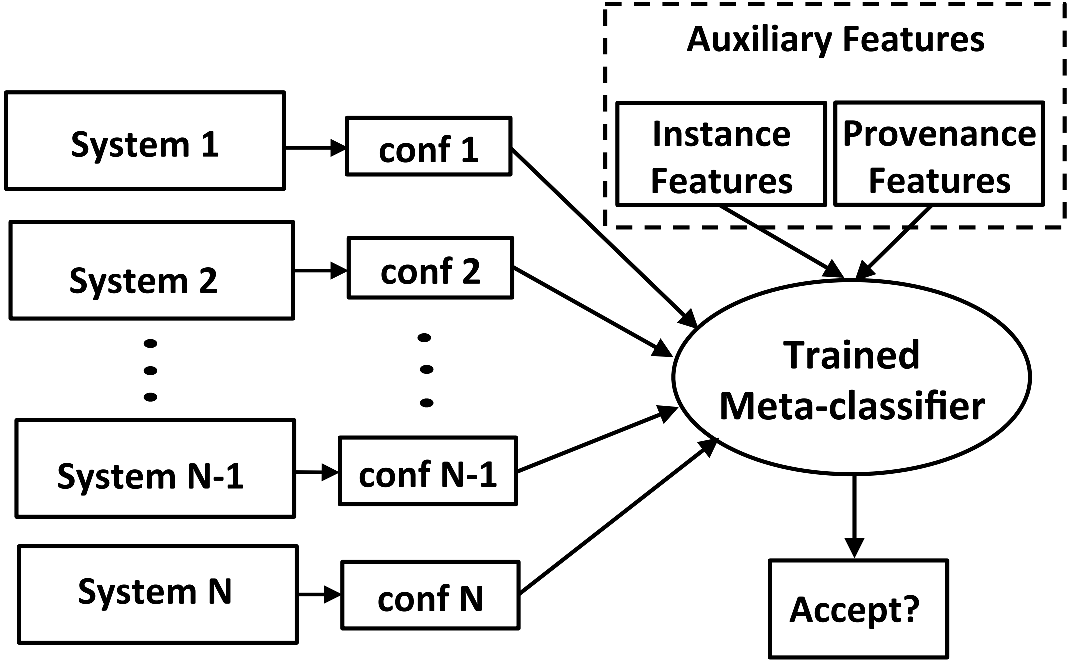
ImageNet Object Detection



(Rajani and Mooney, IJCAI'17)

Stacking with Auxiliary Features (SWAF)

- Stacking using two types of auxiliary 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)

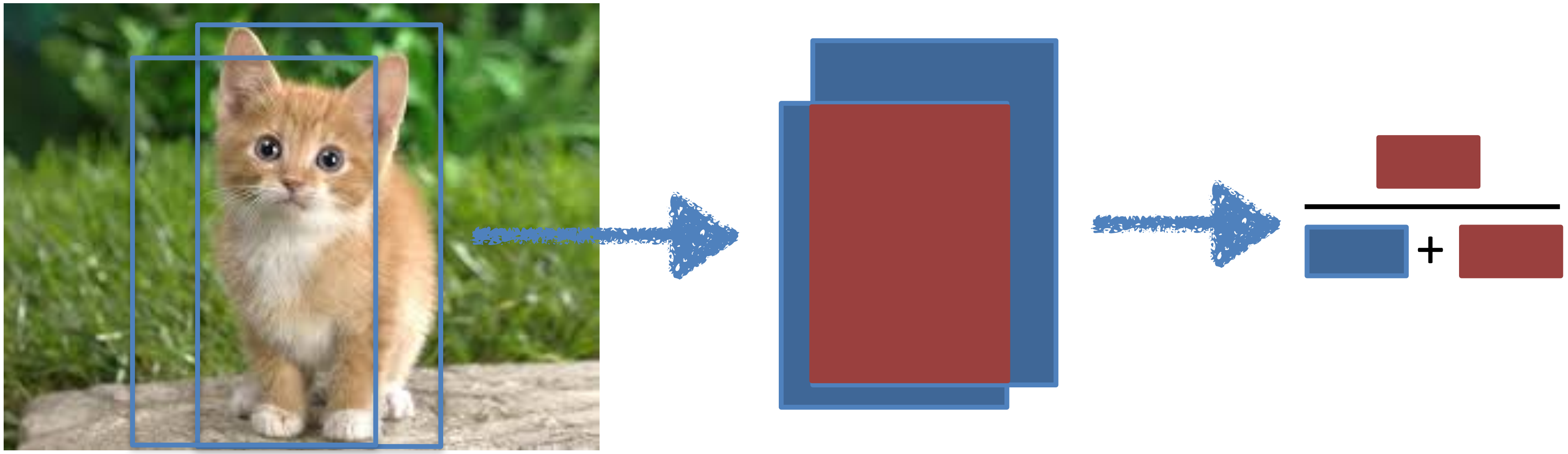
Provenance Features

- 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.

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

(Rajani and Mooney, IJCAI'17)

Object Detection Provenance Features



(Rajani and Mooney, IJCAI'17)

Slot Filling Results

- 2016 SF — 8 shared systems

Approach	Precision	Recall	F1
Oracle voting (>=4)	0.191	0.379	0.206
Top ranked system (Zhang et al., 2016)	0.265	0.302	0.260
Stacking	**0.311**	0.253	0.279
Stacking + instance features	0.257	0.346	0.295
Stacking + provenance features	0.252	0.377	0.302
SWAF	0.258	**0.439**	**0.324**

(Rajani and Mooney, IJCAI’17)

Entity Linking Results

- 2016 EDL — 6 shared systems

Approach	Precision	Recall	F1
Oracle voting (>=4)	0.588	0.412	0.485
Top ranked system (Sil et al., 2016)	0.717	0.517	0.601
Stacking	0.723	0.537	0.616
Stacking + instance features	0.752	0.542	0.630
Stacking + provenance features	**0.767**	0.544	0.637
SWAF	0.739	**0.600**	**0.662**

(Rajani and Mooney, IJCAI'17)

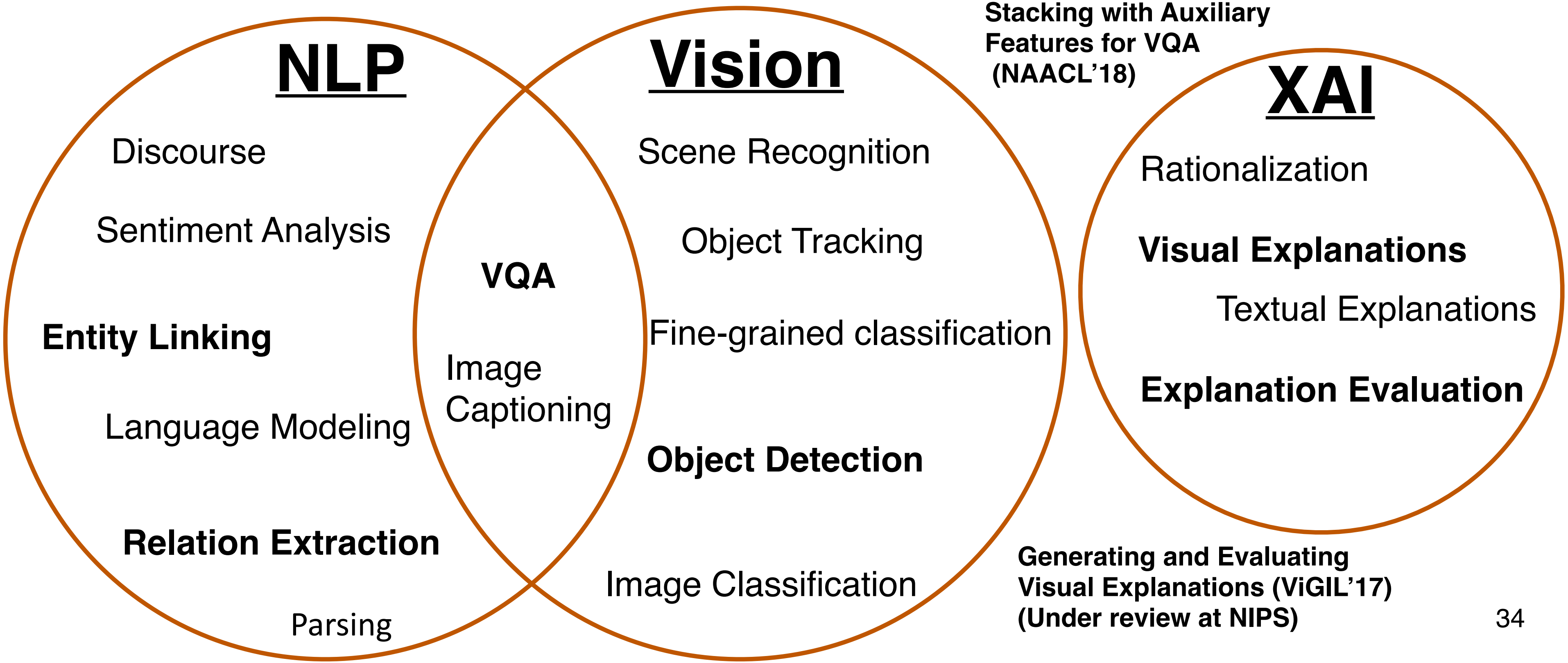
Object Detection Results

- 2015 ImageNet object detection— 3 shared systems

Approach	Mean AP	Median AP
Oracle voting (>=1)	0.366	0.368
Best standalone system (VGG + selective search)	0.434	0.430
Stacking	0.451	0.441
Stacking + instance features	0.461	0.45
Stacking + provenance features	0.502	0.494
SWAF	**0.506**	**0.497**

Since Proposal

Stacking with Auxiliary Features for VQA (NAACL'18)



Generating and Evaluating Visual Explanations (ViGIL'17) (Under review at NIPS)

Since Proposal

Stacking with Auxiliary Features for VQA (NAACL'18)

NLP

Discourse

Sentiment Analysis

Entity Linking

Language Modeling

Relation Extraction

Parsing

VQA

Image Captioning

Vision

Scene Recognition

Object Tracking

Fine-grained classification

Object Detection

Image Classification

XAI

Rationalization

Visual Explanations

Textual Explanations

Explanation Evaluation

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)

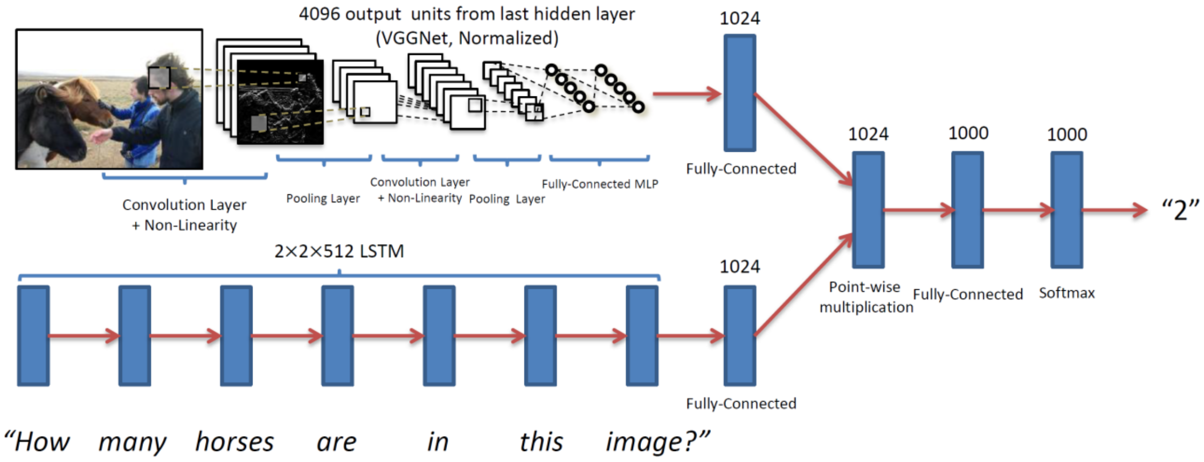
Visual-Question and Only-Question

What is in the child's mouth?	her thumb it's thumg thumb	candy cookie lollipop	What is the animal in the water?	dog dog dog	duck duck guppy
What is the child harnessed to?	her thumb high chair seat	bike child seat seat	How many people are present?	15 15 15	2 3 3

(Rajani and Mooney, NAACL'18)

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)



(Rajani and Mooney, NAACL'18)

SWAF for VQA

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

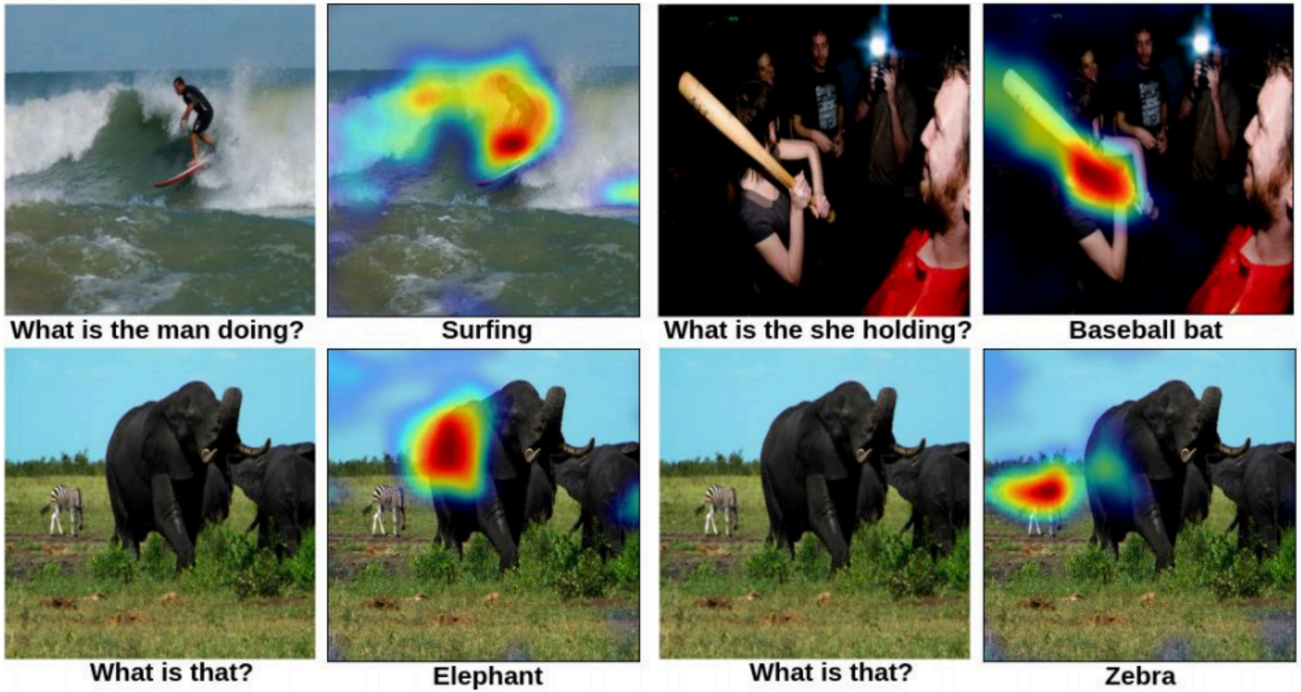
1. Question & Answer types
   - Question prefixes — "What is the color of the vase?"
   - Answer types — yes/no, number and other
2. Question Features
   - BOW representation of words in the question
3. Image Features
   - VGGNet's *fc7* layer

(Rajani and Mooney, NAACL'18, XAI'17)

Visual Explanation as Auxiliary Features

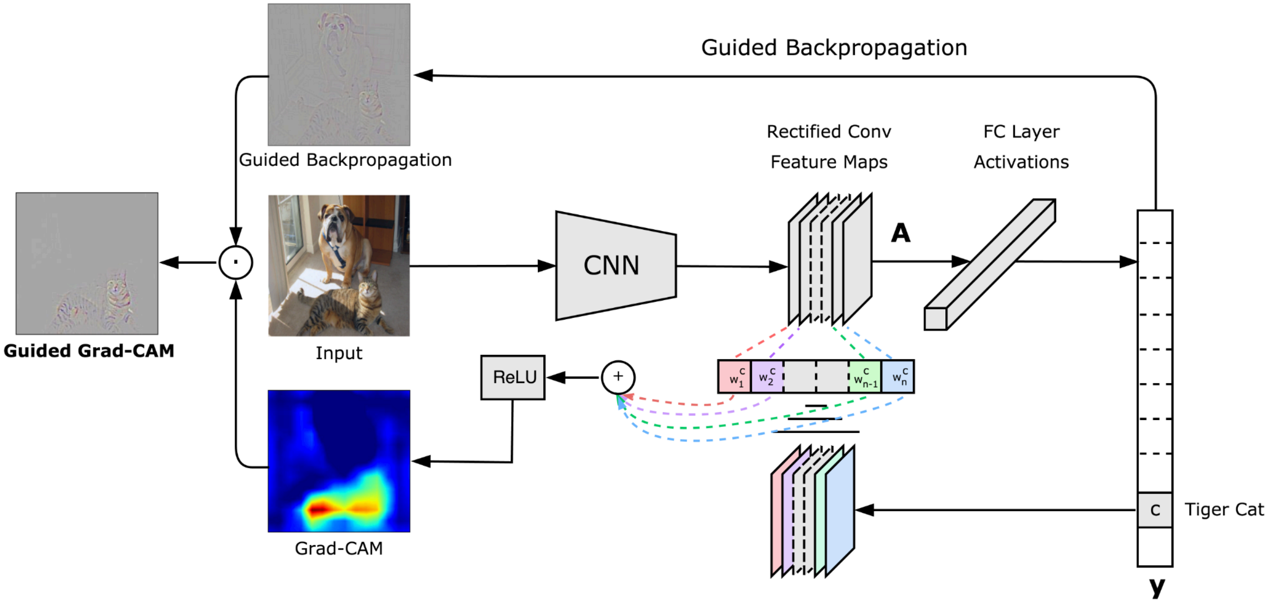
- DNNs attend to relevant regions of image while doing VQA (Goyal et al., 2016).
- The parts of images that the models focus on can be viewed as a ***visual explanation***.
- We use *heat-maps* to visualize explanations in images.
- Enable the stacker to learn to rely on systems that *"look"* at the right region of the image while predicting the answer.

Visual Explanation



Generating Visual Explanation

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



(Rajani and Mooney, NAACL'18, XAI'17)

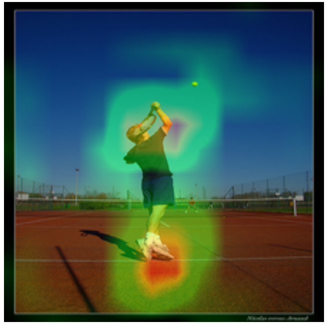
Generating Visual Explanation Features

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

Q: What sport is this?

MCB

HieCoAtt



LSTM

A: Tennis

A: Baseball

(Rajani and Mooney, NAACL’18, XAI’17)

Generating Visual Explanation Features

Q: What is the kid doing? **A:** Skateboarding

LSTM MCB HieCoAtt

(Rajani and Mooney, NAACL’18, XAI’17)

Generating Visual Explanation Features

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

LSTM MCB HieCoAtt

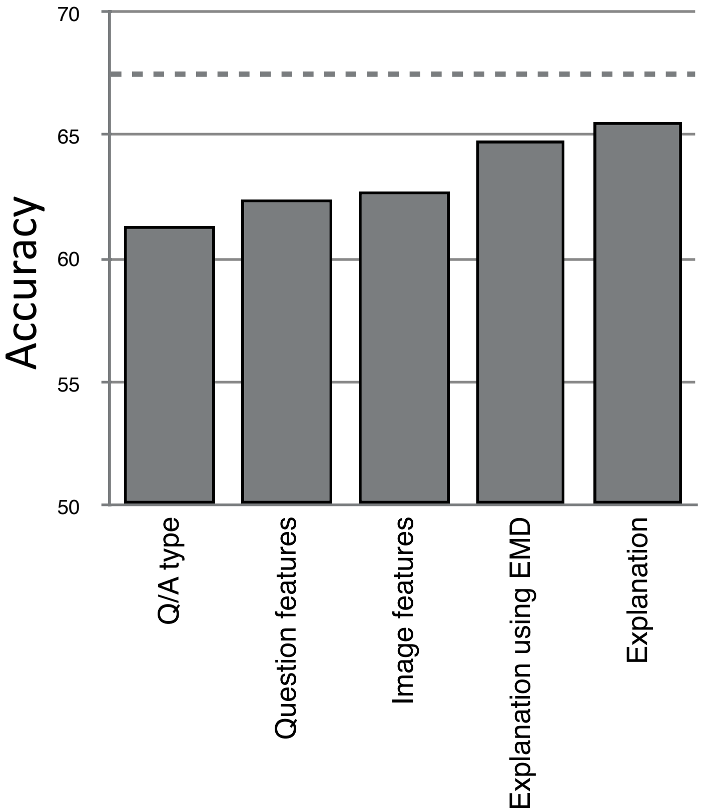
(Rajani and Mooney, NAACL'18)

VQA Results

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

(Rajani and Mooney, NAACL'18)

Feature Ablation Analysis



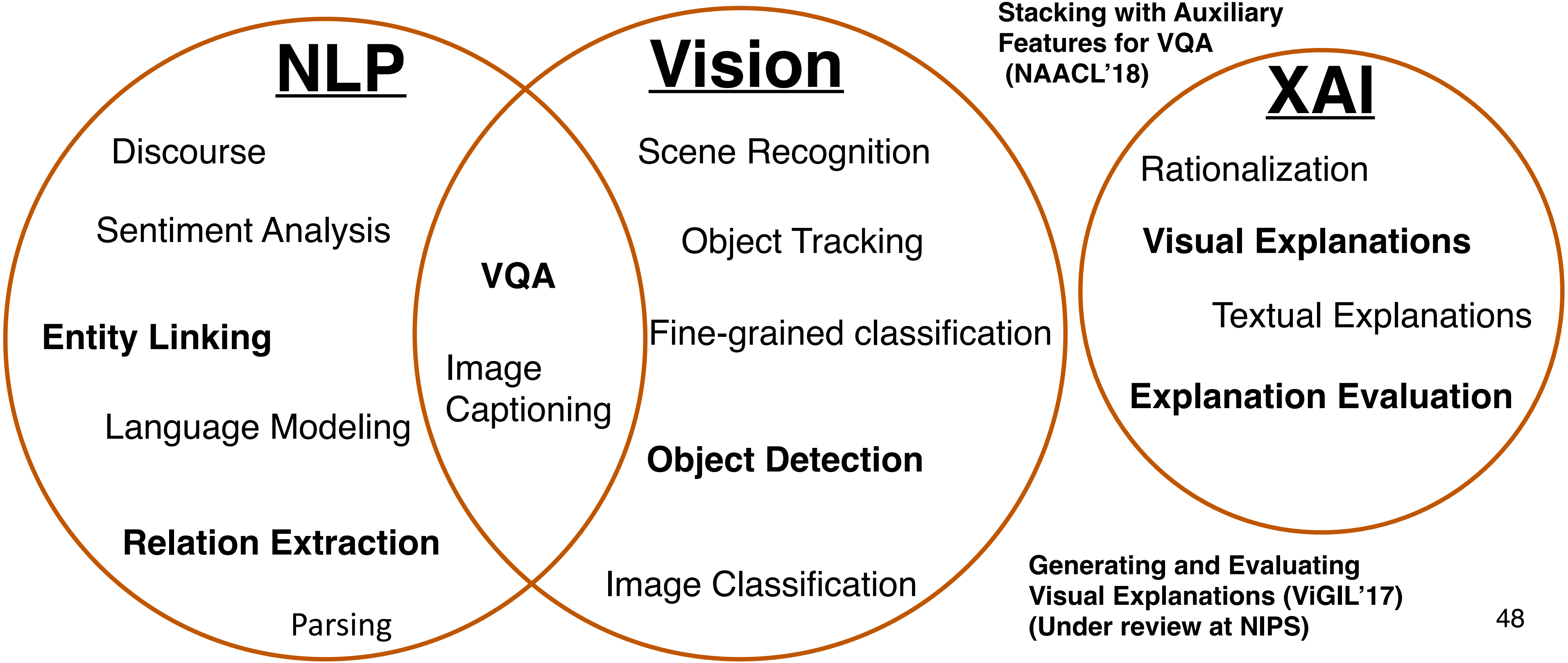
(Rajani and Mooney, NAACL'18)

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.

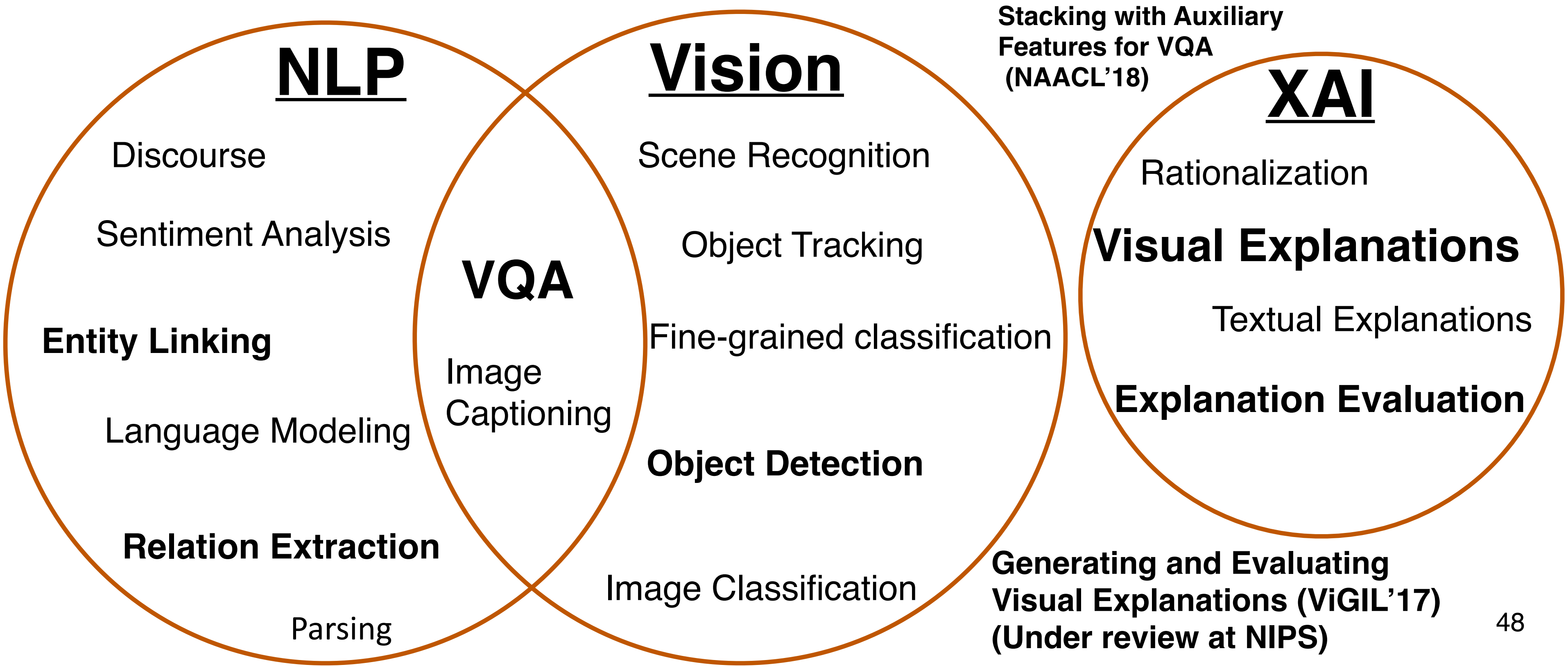
Since Proposal

Stacking with Auxiliary Features for VQA (NAACL'18)

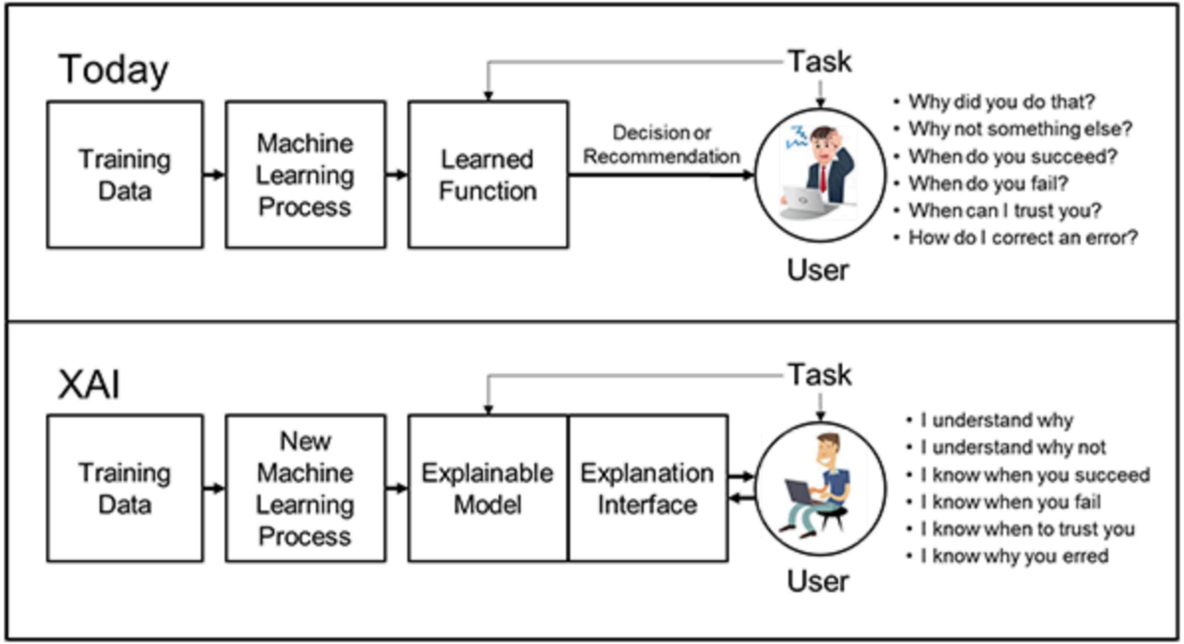


Generating and Evaluating Visual Explanations (ViGIL'17) (Under review at NIPS)

Since Proposal



Explainable AI (XAI)



- Generate explanations for an ensemble.
- Evaluate explanations.

(Rajani and Mooney, ViGIL'17 & BC)

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.

(Rajani and Mooney, ViGIL'17 & BC)

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)

(Rajani and Mooney, ViGIL'17 & BC)

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

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

$$\text{subject to } \sum_{k \in K} w_k = 1$$

(Rajani and Mooney, ViGIL’17 & BC)

WA Example

Q: What color is the umbrella? **A:** Yellow

$\frac{1}{3}\Big[$ w_1

LSTM

$+\, w_2$

HieCoAtt

$+\, w_3$

MCB

$\Big] =$

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.

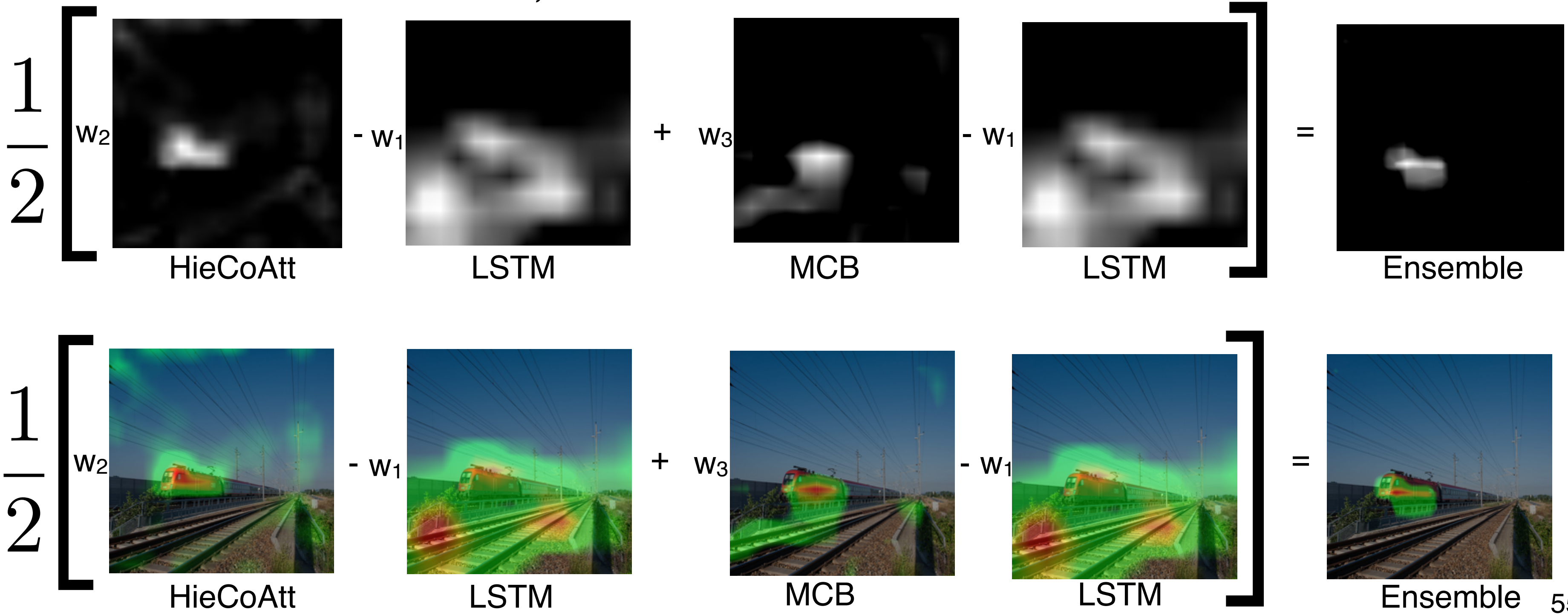
$$E_{i,j} = \begin{cases} \frac{1}{|K|} \sum_{k \in K} \sum_{m \in M} \overbrace{w_k A^k_{i,j} - w_m I^m_{i,j}}^{\text{p}}, & \text{if } p \geq t \\ 0, & \text{otherwise} \end{cases}$$

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

PWA Example

(Rajani and Mooney, ViGIL'17 & BC)

Q: The car in front of the train is what color? **A:** Red
HieCoAtt, MCB answer: red and **LSTM answer:** white

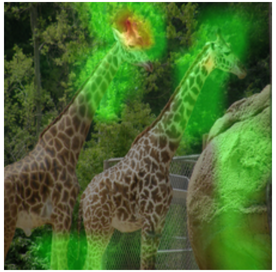


(Rajani and Mooney, ViGIL'17 & BC)

PWA Example

Q: What direction are the giraffe looking? **A:** Right
LSTM, HieCoAtt answer: right and **MCB answer:** left

$\frac{1}{2}\Big[$ w_1



LSTM

$-\ w_3$

MCB

$+\ w_2$

HieCoAtt

$-\ w_3$

MCB

$\Big] =$

Ensemble

(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

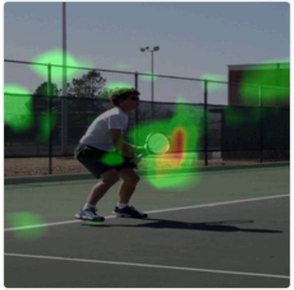
Question: What sport is he playing? **Answer:** tennis

Option A

Option B

Option C

Option D



(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 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.

(Rajani and Mooney, ViGIL'17 & BC)

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 that 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.

Reference Image

Question:
Is there a giraffe eating?

Answer:
yes

Option A

Option B

Option C
Cannot decide

Option D
Incorrect answer

Submit

(Rajani and Mooney, ViGIL'17 & BC)

Comparison Results

Approach	Ensemble	Single System	Cannot decide
	Ensemble (WA)		
LSTM	**58**	36	3
HieCoAtt	**62**	27	6
MCB	52	41	2
	Ensemble (PWA)		
LSTM	**64**	28	3
HieCoAtt	**69**	26	1
MCB	**61**	35	1

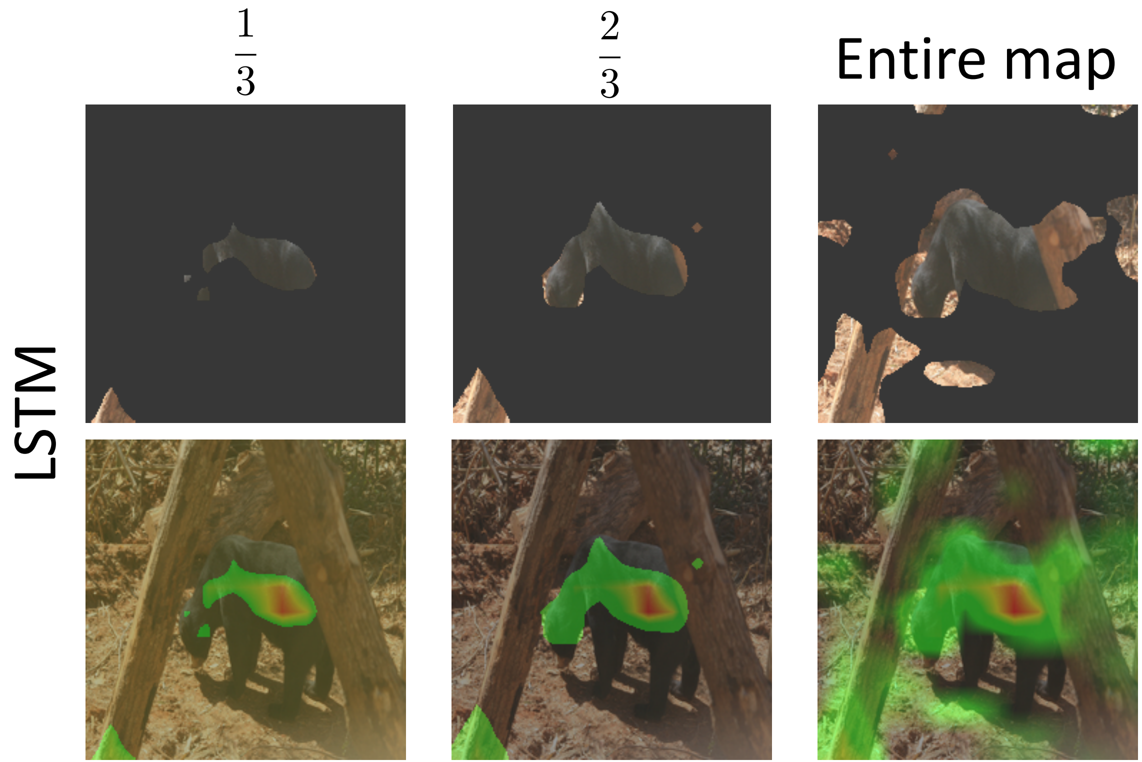
(Rajani and Mooney, ViGIL'17 & BC)

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)

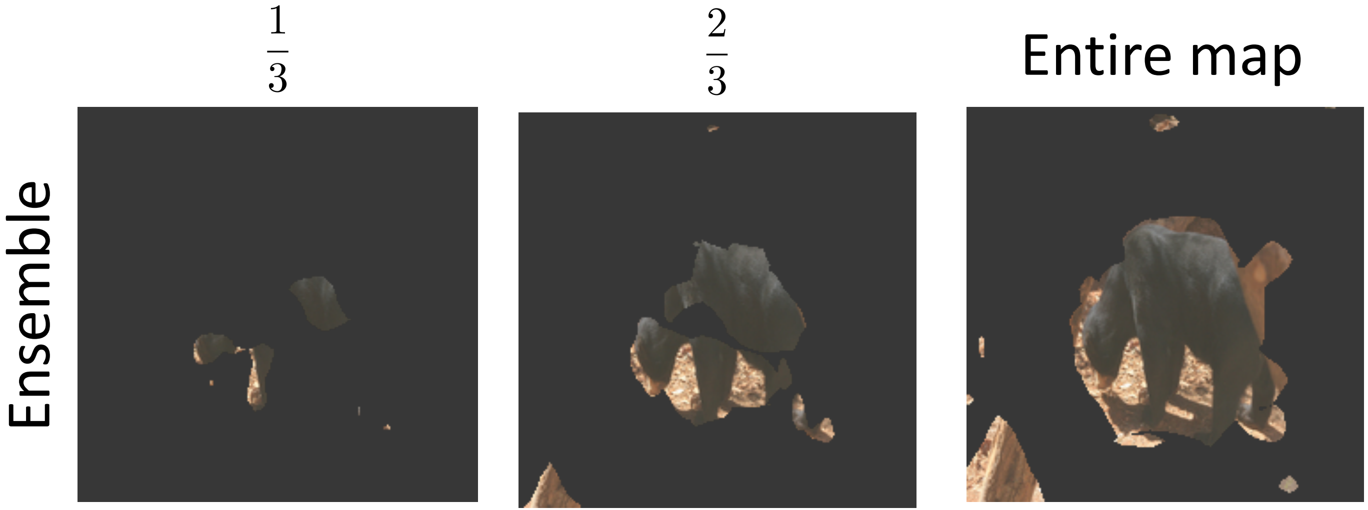
Uncovering Evaluation



Q: What color is the bear? **Answer options:** 1. Brown 2. Black 3. White 4. Still cannot decide

(Rajani and Mooney, ViGIL'17 & BC)

Uncovering Evaluation



Q: What color is the bear? **Answer options:** 1. Brown 2. Black 3. White 4. Still cannot decide

(Rajani and Mooney, ViGIL'17 & BC)

Uncovering Results

System	One-third	Two-thirds	Entire map
Ensemble (PWA)	**29**	**35**	**69**
Ensemble (WA)	17	28	64
LSTM	10	22	42
HieCoAtt	9	19	38
MCB	11	20	46

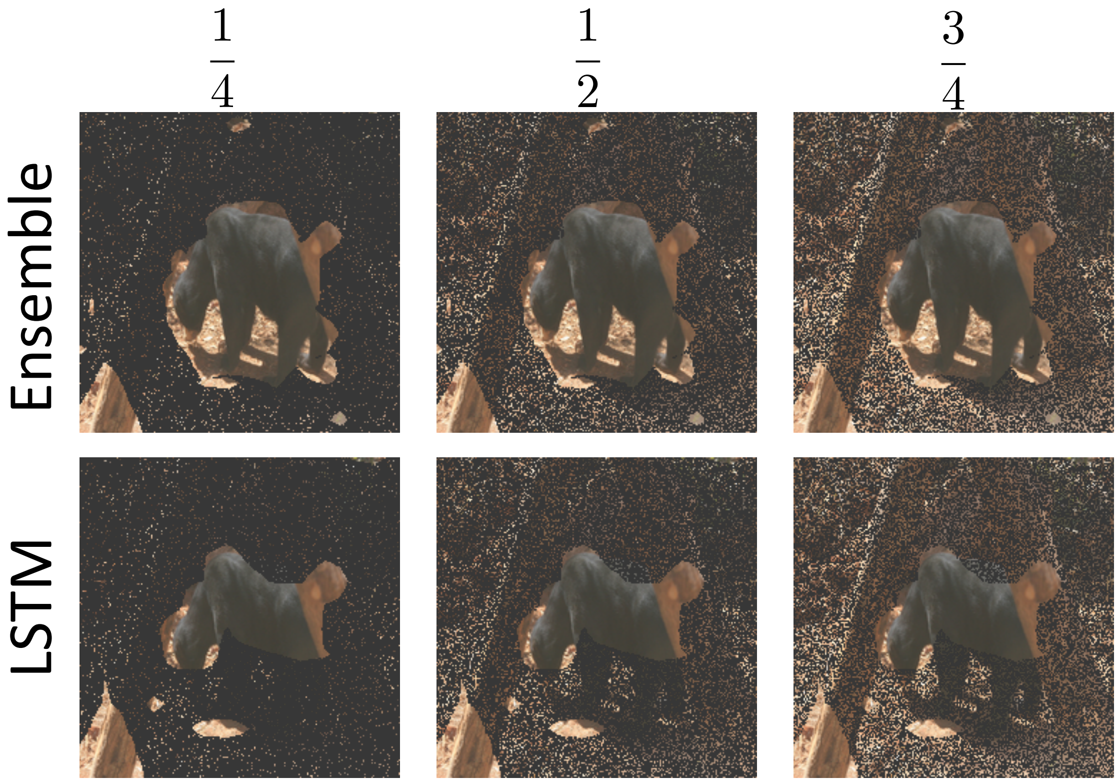
(Rajani and Mooney, ViGIL'17 & BC)

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? **Answer options:** 1. Brown 2. Black 3. White 4. Still cannot decide

(Rajani and Mooney, ViGIL'17 & BC)

Normalized Uncovering Evaluation

$\frac{1}{4}$

$\frac{1}{2}$

$\frac{3}{4}$

Q: How many seats are open? **Answer options:** 1. One 2. Two 3. Three 4. Still cannot decide

(Rajani and Mooney, ViGIL'17 & BC)

Normalized Uncovering Results

System	One-fourth	One-half	Three-fourths
Ensemble (PWA)	**23**	**38**	**76**
Ensemble (WA)	21	34	71
LSTM	10	24	65
HieCoAtt	10	23	57
MCB	12	25	64

(Rajani and Mooney, ViGIL'17 & BC)

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

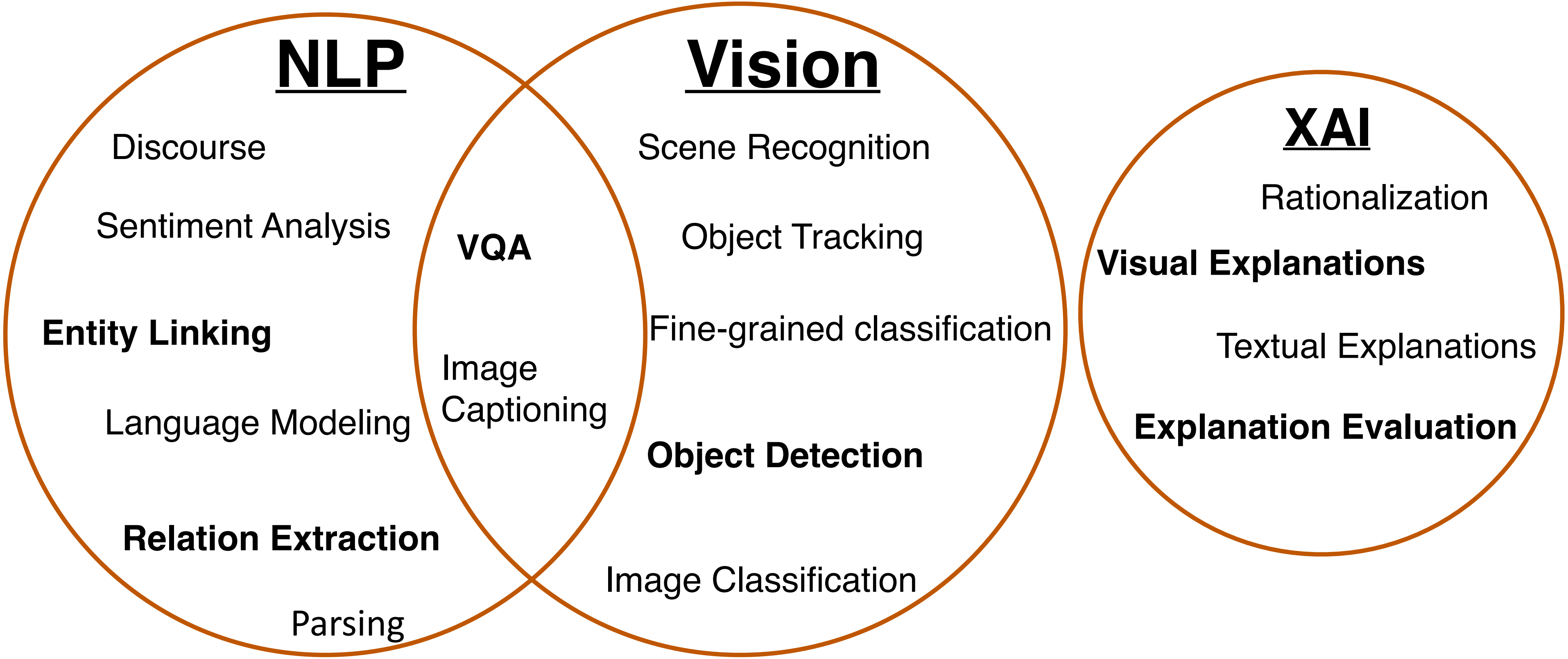
Conclusion

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!


NLP
Discourse
Sentiment Analysis
Entity Linking
Language Modeling
Relation Extraction
Parsing
VQA
Image Captioning
Vision
Scene Recognition
Object Tracking
Fine-grained classification
Object Detection
Image Classification
XAI
Rationalization
Visual Explanations
Textual Explanations
Explanation Evaluation

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In The IEEE International Conference on Computer Vision (ICCV), December 2015.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In Proccedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 3319–3327, 2017.

Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? Computer Vision and Image Understanding, 163:90–100, 2017.

Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal Compact Bilinear pooling for Visual Question Answering and Visual Grounding. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP2016), 2016.

Yash Goyal, Akrit Mohapatra, Devi Parikh, and Dhruv Batra. Towards Transparent AI Systems: Interpreting Visual Question Answering Models. arXiv preprint arXiv:1608.08974, 2016.

Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. Generating Visual Explanations. arXiv preprint arXiv:1603.08507, 2016.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In Advances In Neural Information Processing Systems (NIPS2016), pages 289–297, 2016.

Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, editors, Advances in Neural Information Processing Systems 27, pages 1682–1690. Curran Associates, Inc., 2014.

Hyeonwoo Noh, Paul Hongsuck Seo, and Bohyung Han. Image question answering using convolutional neural network with dynamic parameter prediction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR2016), pages 30–38, 2016.

Nazneen Fatema Rajani and Raymond J. Mooney. Combining Supervised and Unsupervised Ensembles for Knowledge Base Population. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP-16), 2016.

Nazneen Fatema Rajani and Raymond J. Mooney. Ensembling visual explanations for vqa. In Proceedings of the NIPS 2017 workshop on Visually-Grounded Interaction and Language (ViGIL), December 2017.

Nazneen Fatema Rajani and Raymond J. Mooney. Stacking With Auxiliary Features. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI2017), Melbourne, Australia, August 2017.

Nazneen Fatema Rajani and Raymond J. Mooney. Stacking With Auxiliary Features for Visual Question Answering. In Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018.

Nazneen Fatema Rajani*, Vidhoon Viswanathan*, Yinon Bentor, and Raymond J. Mooney. Stacked Ensembles of Information Extractors for Knowledge-Base Population. In Association for Computational Linguistics (ACL2015), pages 177–187, Beijing, China, July 2015.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In The IEEE International Conference on Computer Vision (ICCV2017), Oct 2017.

Suhr, Alane, et al. "A corpus of natural language for visual reasoning." *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*

I-Jeng Wang, Edwina Liu, Cash Costello, and Christine Piatko. JHUAPL TACKBP2013 slot filler validation system. In TAC2013, 2013.

David H. Wolpert. Stacked Generalization. Neural Networks, 5:241–259, 1992.

Backup Slides

(Rajani and Mooney, ViGIL'17 & BC)

WA Example (forced version)

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

$\frac{1}{3}\Big[\, w_1$

forced LSTM

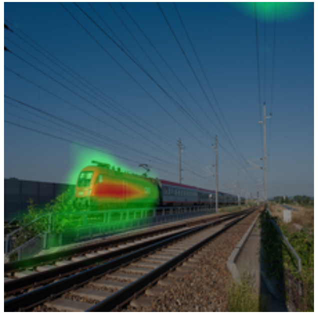
$+\, w_2$

HieCoAtt

$+\, w_3$

MCB

$\Big] =$



Ensemble