FALL 2020 CS 395T

TEXAS
The University of Texas at Austin

Language Generation Grounded in Images

CS 395T: Topics in Natural Language Processing
10/1/2020

Prateek Chaudhry and Quang Duong, The University of Texas at Austin

Clue: Cross-modal Coherence Modeling for Caption Generation [ACL 2020]

By Malihe Alikhani, Piyush Sharma, Shengjie Li, Radu Soricut, and Matthew Stone

Image Captioning Task

- Given an image, generate a natural language description of the content observed in the image

by Joi Ito

the trail climbs steadily uphill most of the way.

by Danail Nachev

the stars in the night sky.

by Justin Higuchi

musical artist performs on stage during festival.

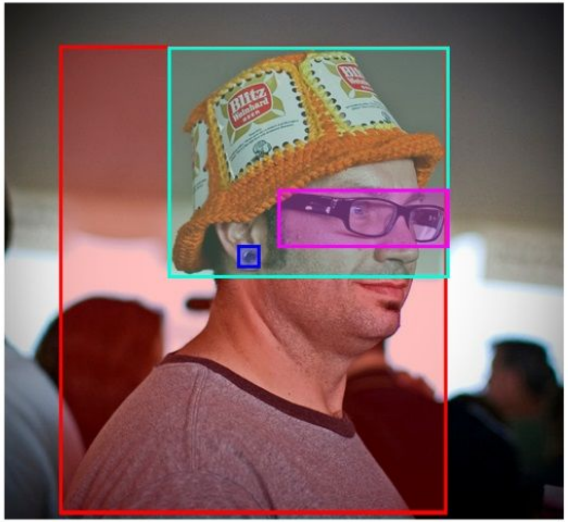
by Viaggio Routard

popular food market showing the traditional foods from the country.

Image from Conceptual Captions Website

Flickr(30K) Dataset

- 31,783 images from Flickr each with 5 captions [Young et al. 2014]
- Added correspondences [Plummer et al. 2017]

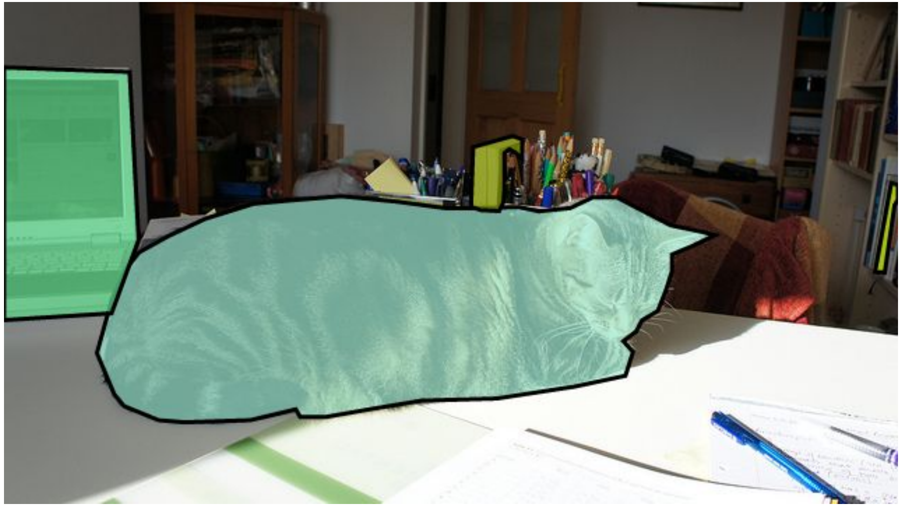


A man with pierced ears is wearing glasses and an orange hat.
A man with glasses is wearing a beer can crotched hat.
A man with gauges and glasses is wearing a Blitz hat.
A man in an orange hat starring at something.
A man wears an orange hat and glasses.

Image from [Plummer et al. 2017]

MS COCO Dataset

- Microsoft Common Objects in COntext
- ~330K images with 5 captions per image
- [Lin et al. 2014]



a cat is laying on a table near a laptop and papers
there is a cat laying on the table enjoying the sun
a cat is on papers on a computer desk.
a close up of a cat laying on a desk
a cat lying in the sun on a table.

Image from MS COCO Website

Motivation: Object Hallucination

- Datasets are too small for training robust models
- Captions “hallucinate” things that aren’t there
- [Rohrbach et al. 2018]

NBT: A woman talking on a cell phone while sitting on a *bench*.
CIDEr: **0.87**, METEOR: 0.23, SPICE: **0.22,** CHs: **1.00**, CHi: **0.33**

TopDown: A woman is talking on a cell phone.
CIDEr: 0.54, METEOR: **0.26**, SPICE: 0.13, CHs: **0.00**, CHi: **0.00**

Image from [Rohrbach et al. 2018]

Object Hallucination Examples

TopDown: A pile of luggage sitting on top of a ***table***.
NBT: Several pieces of luggage sitting on a ***table***.

TopDown: A group of people sitting around a ***table*** with laptops.
NBT: A group of people sitting around a ***table*** with laptop.

TopDown: A kitchen with a stove and a ***sink***.
NBT: A kitchen with a stove and a ***sink***.

TopDown: A couple of cats laying on top of a ***bed***.
NBT: A couple of cats laying on top of a ***bed***.

TopDown: A cat sitting on top of a ***laptop computer***.
NBT: A cat sitting on a table next to a ***computer***.

TopDown: A brown dog sitting on top of a ***chair***.
NBT: A brown and white dog sitting under an ***umbrella***.

TopDown: Aa man and a woman are playing with a ***frisbee***.
NBT: A man riding a skateboard down a street.

TopDown: A man standing on a beach holding a ***surfboard***.
NBT: A man standing on top of a sandy beach.

Image from [Rohrbach et al. 2018]

Conceptual Captions Dataset

- 3.3M Image-Caption pairs generated from web images and associated alt text
- Diverse range of image and caption styles
- Reduced object hallucination
- [Sharma et al. 2018]

Alt-text: A Pakistani worker helps to clear the debris from the Taj Mahal Hotel November 7, 2005 in Balakot, Pakistan.

Conceptual Captions: a worker helps to clear the debris.

Alt-text: Musician Justin Timberlake performs at the 2017 Pilgrimage Music & Cultural Festival on September 23, 2017 in Franklin, Tennessee.

Conceptual Captions: pop artist performs at the festival in a city.

Image from [Sharma et al. 2018]

Conceptual Captions Data Extraction

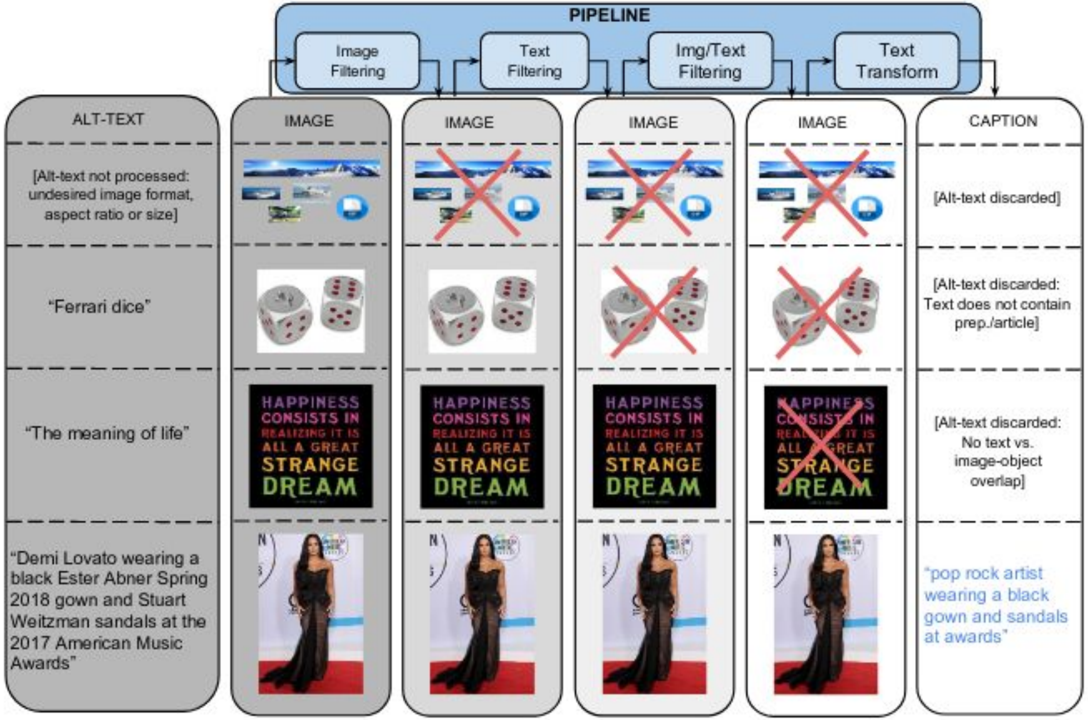


Image from [Sharma et al. 2018]

Context Hallucination

- Conceptual Captions Dataset has lower quality image-caption pairs
  - Contextual Background
    - "This is the new general manager of the team"
  - Subjective Evaluation
    - "This is stylish"
- Metrics compare against only the captions and disregards images
  - How can we tie the caption to the image?

Related Work - Metric

- CHAIR [Rohrbach et al. 2018] - Metric penalizing hallucinations by checking against reference caption and if objects are actually in the image

Related Work - Dataset

- CITE [Alikhani et al. 2019] - Image-Caption discourse coherence relations for a multi-modal recipe dataset
  - Crowd-Sourced various questions examining simple relations between instructions and images

Discourse

- Examines language “beyond the sentence”
  - Extension of “grammaticality” to the inter-sentence level
  - Analyzing how sentences (or other discourse units) relate to one another

Discourse Coherence Relations

- Describes relationships between discourse units (such as sentences or clauses)
  - **Contingency**
    - I was tired because I just ran 5 miles.
  - **Comparison**
    - She aced the test, but he barely passed.
  - **Expansion**
    - He likes cats. In particular, he loves ragdolls.
  - **Temporal**
    - They studied at the library. Afterwards, they went home.

Multi-Modal Coherence Relations

- Images as discourse unit
- Relations tie images and captions together
  - **Visible**, **Subjective**, **Action**, **Story**, **Meta**, **Irrelevant**

Visible, Action, Subjective

(b) CAPTION: young happy boy swimming in the lake.

Image from [Alikhani et al. 2020]

Visible

- Text restates what can be found in the image
- A person on a mountain trail

Image from [Alikhani et al. 2020]

Subjective

- Text is reaction / evaluation of image content
- A beautiful and stunning mountain range

Image from [Alikhani et al. 2020]

Action

- Text describes process occurring within image
- A person hiking up a mountain trail

Image from [Alikhani et al. 2020]

Story

- Text describes circumstances within image
- A person approaching their campsite

Image from [Alikhani et al. 2020]

Meta

- Text discusses the manner in which the image was taken or created (When, Where, and How subcategories)
- A landscape of a person at 1550 elevation on the slopes

Image from [Alikhani et al. 2020]

Irrelevant

- Text has nothing to do with the image
- A walrus rolls over for a tasty treat at SeaWorld

Image from [Alikhani et al. 2020]

Clue Dataset

- 5,000 image-caption pairs from Conceptual Captions
- 5,000 image-caption pairs from the top 5 image captioning models on 1,000 images from the Open Images Dataset [Kutzenova et al. 2020]

Clue Dataset Relation Labeling

- Crowd-sourced non-expert labeling was unsatisfactory for general relations
- Relations for each image-caption pair labeled by experts (2 undergrad linguistics students) via authors' annotation interface
  - Cohen's $\kappa = 0.81$ indicating high agreement

Clue Dataset Relation Distribution

- Mostly **Visible** relations
- Models biased towards more **Meta** and **Irrelevant** (increased context hallucination)

	Visible	Subjective	Action	Story	Meta	Irrelevant
Ground-truth	64.97%	9.77%	18.77%	29.84%	24.59%	3.09%
Model output	69.72%	1.99%	11.22%	17.19%	58.94%	16.97%
Ground-truth + Model	66.91%	6.58%	15.68%	24.67%	38.65%	8.77%

Table 1: Distribution of coherence relations over the ground-truth and the model outputs.

Clue Dataset **Meta** Sub-Categories

- Models learn to generate locations and how things occur
- Not as good for **Temporal** relations

	When	How	Where
Ground-truth	33.74%	64.40%	28.60%
Model output	21.75 %	72.84%	41.03%

Table 2: Distribution of fine-grain relations in the Meta category over the ground-truth and the model outputs.

Clue Dataset Relation Co-Occurrence

- Significant Visible / Meta overlap
  - Increases by 32% in model output

		Subjective	Action	Story	Meta
Ground Truth	Visible	3.96%	16.71%	8.08%	22.49%
	Subjective		0.72%	2.96%	1.25%
	Action			2.72%	9.13%
	Story				2.89%
		Subjective	Action	Story	Meta
Model	Visible	1.01%	10.62%	9.67%	54.55%
	Subjective		0.00%	1.49%	0.76%
	Action			2.12%	7.96%
	Story				8.06%

Table 1: Rates of co-occurrences of different labels in groundtruth and the model outputs.

Clue Dataset Source / Relation Distribution

- Most of Getty Images and Picdn have **Visible** relations
- Daily Mail, a news source, has a lot of **Story** relations

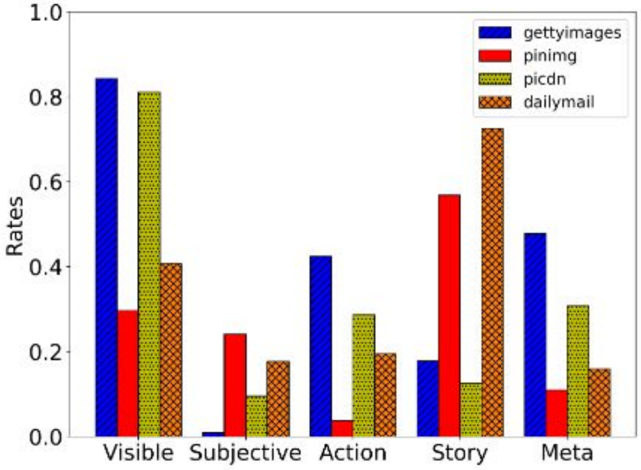


Figure 4: Different resources have different kinds image–caption pairs. The graph shows the distribution of labels in the top four domains present in the Conceptual Captions dataset.

Figure from [Alikhani et al. 2020]

Multi-Label Relation Prediction

- Given an image and a caption, predict all of the coherence relations for that image-caption pair
- 80 / 20 Train-Test split with 5-Fold Cross Validation

Multi-Label Models

1. **SVM**: 1-5 n-gram BoW classifier using text only
2. **GloVe** [Pennington et al. 2014] **Encoder**: LSTM + BN + FC + Tanh
3. **BERT** [Devlin et al. 2018] **Encoder**: Sentence embedding + <CLS>
4. **ResNet-50** [He et al. 2016] **Encoder**: ResNet + BN + FC + ReLU
5. **GloVe + ResNet-50**: (2) + (4)
6. **BERT + ResNet-50**: (3) + (4)

Not entirely clear how (5) and (6) are constructed

Multi-Label Results

	Visible	Subjective	Action	Story	Meta	Irrelevant	Weighted
SVM (text-only)	0.83	0.12	0.32	0.21	0.19	0.00	0.48
GloVe (text-only)	0.80	0.44	0.58	0.57	0.44	0.08	0.63
BERT (text-only)	0.82	0.35	0.62	0.62	0.44	0.06	0.65
GloVe + ResNet	0.81	0.36	0.58	0.60	0.45	0.07	0.64
BERT + ResNet	0.83	0.36	0.69	0.62	0.44	0.06	0.67

Table 3: The F_1 scores of the multi-class classification methods described in Section 4.1; 80-20 train-test split; 5-fold cross validation.

Multi-Label Results

- Both BERT and GloVe models outperform the SVM baseline by a significant margin
- Results only slightly change when ResNet-50 image encoder is added to the text encoder

Single-Label Prediction

- Goal: Generate captions with a desired coherence relation
  - Need to distinguish different coherence relations for co-occurring types
- Reduce multiple coherence relation labels down to a single label and predict that

Single-Label Mapping

- Reduce each image down to one label
  - If it contains **Meta**, set the relation to **Meta**
  - If it contains **Visible**, but not **Meta** or **Subjective**, then set it to **Visible**
  - Otherwise randomly sample from the image's relations
- 3910 pairs with 3400 / 510 train-test split

Single-Label Models

- **BERT + ResNet-50**
- **BERT + GraphRise** [Juan et al. 2020]
  - GraphRise is pre-trained on 260M images with 40M labels and outputs 64-D representation
- **USE** [Cer et al. 2018] **+ GraphRise**
  - Universal Sentence Encoder produces a 512-D representation of the sentence
- The above are fed into a 3 layer / 256 neuron fully connected network with ReLU activations + Softmax into 6 classes
- Dropout of 0.5 and trained with Adam with learning rate of 1e-6

Single-Label Results

	Visible	Subjective	Action	Story	Meta	Irrelevant	Weighted
Ground-truth Distribution	46.65%	7.07%	1.31%	19.09%	23.42%	2.46%	
BERT + ResNet	0.64	0.26	0.02	0.52	0.46	0.07	0.52
BERT + GraphRise	0.59	0.15	0.00	0.42	0.34	0.00	0.45
USE + GraphRise	0.69	0.45	0.00	0.57	0.48	0.00	**0.57**

Table 4: The F_1 scores of coherence relation classifiers **with label mapping**. The aggregated Weighted scores use the numbers in the first row as weights.

Single-Label Results

- New distribution of relations is less skewed towards **Visible**
- BERT + GraphRise does worse than BERT + ResNet across the board
- USE + GraphRise does the best overall
- Both GraphRise networks have 0 F_1 scores for **Action** and **Irrelevant**
- Multi-Label baseline models removed
  – Namely, no more text-only models

Coherence-Aware Caption Generation

- Generate captions for the rest of the Conceptual Captions dataset that do not have coherence relation labels
  - Predict the relation with USE + GraphRise for the image-caption pair
  - Feed it into the generation process as a target relationship

Coherence-Aware Caption Model

- USE + GraphRise for coherence labels with NONE for Coherence-Agnostic evaluation

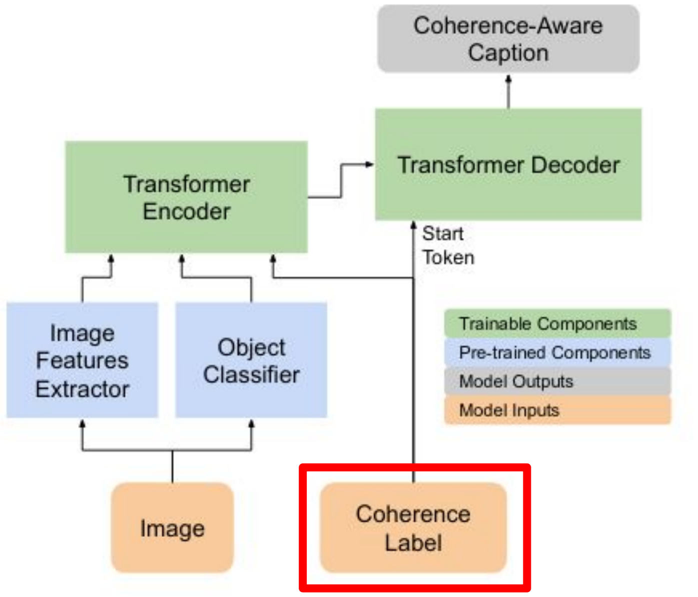


Figure 5: Coherence-aware image captioning model

Figure from [Alikhani et al. 2020]

Coherence-Aware Caption Model

- GraphRise as Image Feature Extractor
- Object labels from Google Cloud Vision API embedded like word2vec [Mikolov et al. 2013] for co-occurring objects in web pages

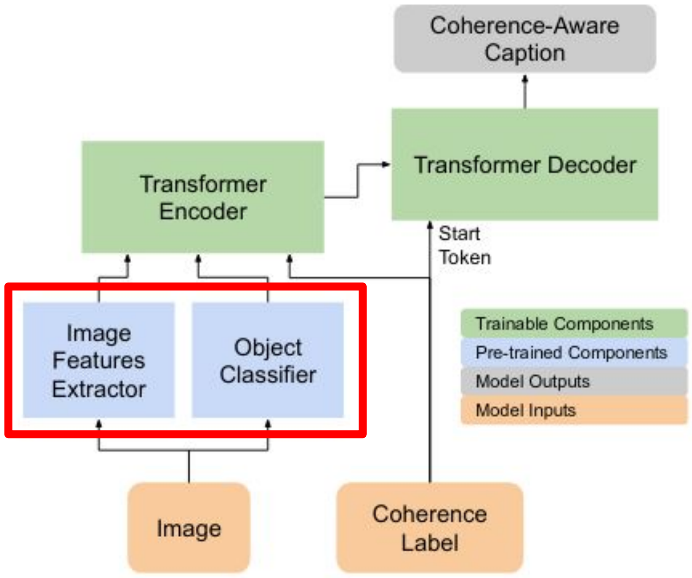


Figure 5: Coherence-aware image captioning model

Figure from [Alikhani et al. 2020]

Coherence-Aware Caption Model

- Transformer [Vaswani et al. 2017] with 6 enc/dec layers, 8 attention heads, 512-D embedding space

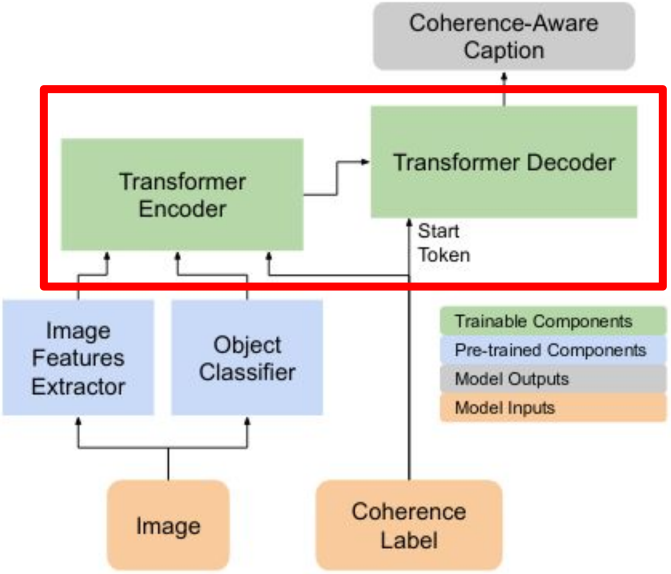


Figure 5: Coherence-aware image captioning model

Figure from [Alikhani et al. 2020]

Coherence-Aware Caption Results

	Coherence agnostic	Visible coherence-aware	Subjective coherence-aware	Story coherence-aware	Meta coherence-aware
Visible	**52.1%**	**79.9%**	31.7%	25.0%	42.80%
Subjective	11.4%	2.6%	24.4%	2.6%	1.9%
Action	10.7%	10.8%	6.3%	8.8%	11.4%
Story	**51.3%**	16.0%	**45.0%**	**58.8%**	17.34%
Meta	31.2%	32.8%	15.1%	17.7%	**46.5%**
Irrelevant	12.2%	12.3%	10.7%	9.9%	21.40%
When	9.5%	5.6%	4.1%	17.7%	9.6%
How	21.3%	21.3%	9.6%	25.0%	30.26%
Where	5.3%	8.6%	4.1%	8.8%	16.6%

Table 5: The distribution of coherence relations in image–caption pairs when captions are generated with the discourse–aware model vs the discourse agnostic model (the mode of the distribution in bold).

Coherence-Aware Caption Results

- Expert evaluation over 1500 image-caption pairs with 300 in each category
- The target coherence relation is generated more often when comparing the coherence-aware model with the coherence agnostic model
- Action / Irrelevant - aware models are left out, likely because they are poorly predicted

Coherence-Aware Caption Examples

(a) coherence-aware *Meta*: A girl in the winter forest.
coherence–agnostic: beautiful girl in a red dress.

(b) coherence-aware *Visible*: the pizza at restaurant is seen.
coherence–agnostic: the best pizza in the world.

(c) coherence-aware *Subjective*: beautiful chairs in a room.
coherence–agnostic: the living room of the home.

(d) coherence-aware *Story*: how to spend a day.
coherence–agnostic: dogs playing on the beach.

Figure 6: Captions generated by the coherence-aware and coherence-agnostic models. (Photo credits: YesVideo; TinnaPong; Sok Chien Lim; GoPro)

Figure from [Alikhani et al. 2020]

Human Evaluation - “Good”

- Asked humans to determine if **Visible** outputs were “good” or not
- 86% of the coherence-aware outputs were “good” vs 74% of the coherence-agnostic approach outputs
- Under data for the Conceptual Caption Workshop at CVPR 2019, SOTA models obtain 67% “good” ratings

Human Evaluation - Preference

- Humans choose between coherence-agnostic and coherence-aware outputs
- 68.2% of the coherence-aware outputs were preferred as opposed to 31.8% of the coherence-agnostic

Human Evaluation - Quality / Relevance

- Humans were asked to evaluate quality and relevance to the image on a Likert scale
- Coherence-Aware: 3.44 Quality / 4.43 Relevance
- Coherence-Agnostic: 2.83 Quality / 4.40 Relevance
- Quality not reflected in CIDEr scores:
  - Coherence-Aware: 0.958
  - Coherence-Agnostic: 0.964

Discussion

- How often were the predicted coherence relation labels wrong? Would the coherence-aware expert labeled distribution change if they used gold label coherence relation labels?
- What would happen if the multi-label models were used and multiple coherence labels were fed into the coherence-aware generation approach?
  - At the very least, the **Action** F_1 score for the multi-label model was non-zero

Discussion

- Why was **Action** not preferred in the single-label mapping? The resulting distribution dropped **Action** prevalence from 18.77% to 1.31%
- Why did they only use 3910 of the 10K coherence relation labeled samples that they had?
- Why did they do human evaluation on the **Visible** coherence relation only? While **Subjective** would be harder to evaluate, **Action**, **Story**, and **Meta** would be similar to **Visible** to evaluate.

Grounded Situation Recognition [ECCV 2020]

By Sarah Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi

Situation Recognition

- From **language to structure**
- Given an image, generate a structured summary
  - main activity
  - participating actors, objects, substances, and locations
  - roles of participants

JUMPING			
ROLE	VALUE	ROLE	VALUE
AGENT	BOY	AGENT	BEAR
SOURCE	CLIFF	SOURCE	ICEBERG
OBSTACLE	-	OBSTACLE	WATER
DESTINATION	WATER	DESTINATION	ICEBERG
PLACE	LAKE	PLACE	OUTDOOR

Source: Situation Recognition: Visual Semantic Role Labeling for Image Understanding, Yatskar et. al

Situation Recognition

FrameNet

- Linguist-authored verb lexicon
- **Verb ⟷ Frame**: Set of semantic roles
- Describe possible situations
- QA, Information Extraction, Semantic Role Labeling
- Built/Being built for many languages
  - French, Spanish, Chinese, Japanese, Korean, Portuguese, Swedish, German

JUMPING			
ROLE	VALUE	ROLE	VALUE
AGENT	BOY	AGENT	BEAR
SOURCE	CLIFF	SOURCE	ICEBERG
OBSTACLE	-	OBSTACLE	WATER
DESTINATION	WATER	DESTINATION	ICEBERG
PLACE	LAKE	PLACE	OUTDOOR

Source: Situation Recognition: Visual Semantic Role Labeling for Image Understanding, Yatskar et. al

Situation Recognition

imSitu dataset (Yatskar et. al.)

- Filter verbs from FrameNet for describing image events
- Collect related images from Google search
- Annotate roles using crowdsourcing

JUMPING

ROLE	VALUE	ROLE	VALUE
AGENT	BOY	AGENT	BEAR
SOURCE	CLIFF	SOURCE	ICEBERG
OBSTACLE	-	OBSTACLE	WATER
DESTINATION	WATER	DESTINATION	ICEBERG
PLACE	LAKE	PLACE	OUTDOOR

Source: Situation Recognition: Visual Semantic Role Labeling for Image Understanding, Yatskar et. al

Situation Recognition

What situation recognition answers:

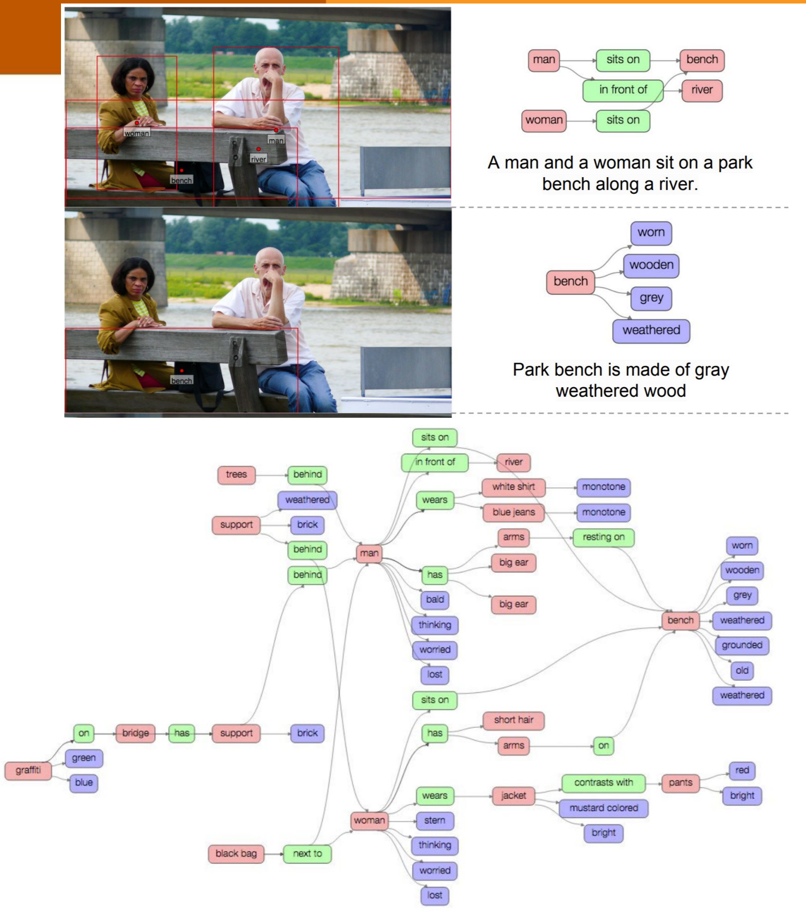
- what is happening?
- who are participants?
- what are their roles?

What it does not answer:

- **where** are the entities in the image? → **Grounded Situation Recognition (GSR)**

Related Work

- Flickr30k Entities
  - Grounded captioning
  - Human centric
- v-COCO [Gupta et. al., 2015]
  - Much smaller scale
- Visual Genome
  - Dense scene graphs
  - Most relations are binary and positional

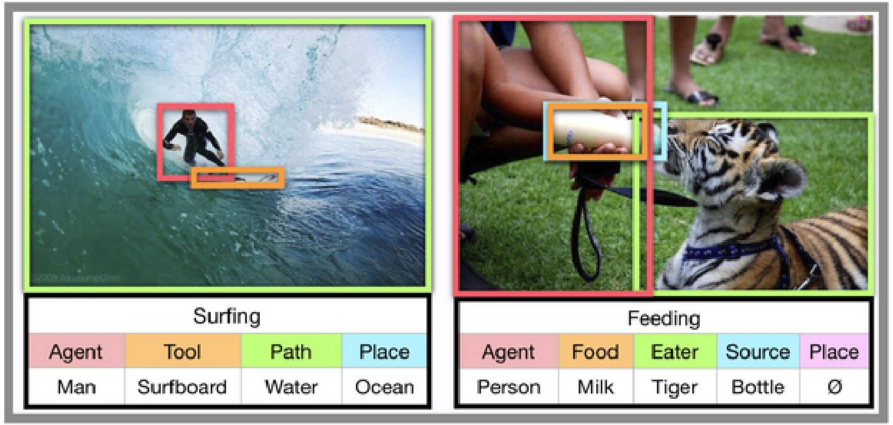


Source: Visual Genome Connecting Language and Vision Using Crowdsourced Dense Image Annotations, Krishna, Ranjay, et al.

Grounded Situation Recognition

Given an image, produce three outputs:

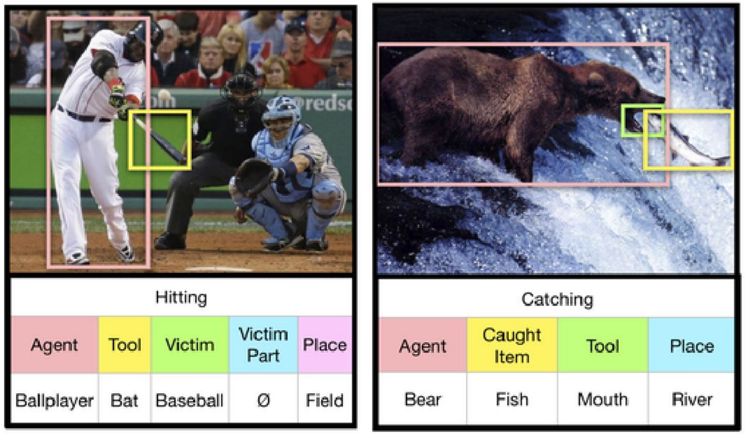
- Verb
  - classify among 504 verbs
- Frame
  - 1 to 6 semantic roles associated with the verb
  - match roles with related nouns
- Groundings
  - Identify bounding boxes for identified nouns



SWiG : Situations With Groundings

Dataset for GSR

- Retain from imSitu
  - images
  - frame annotations (three for each image)
  - data splits
- Obtain bounding boxes
  - using AMT
  - each role annotated by three workers
  - use average for truth value

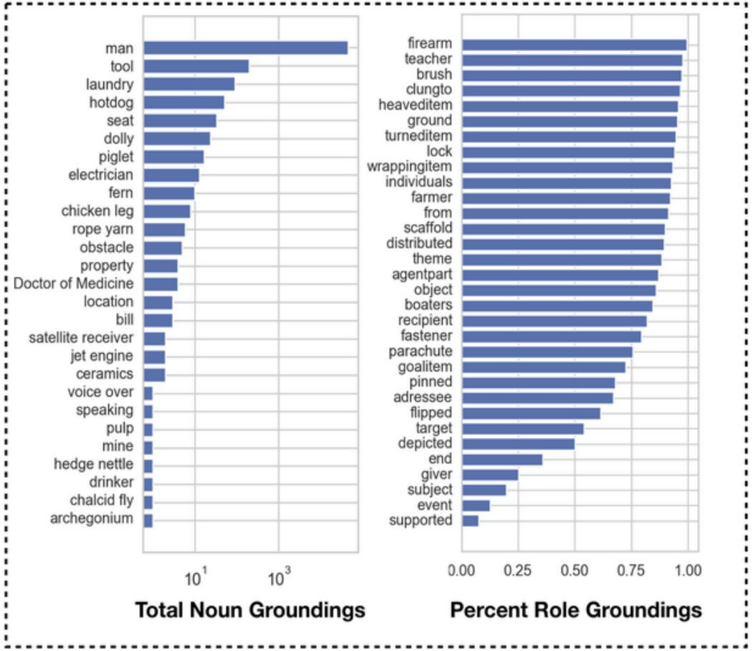


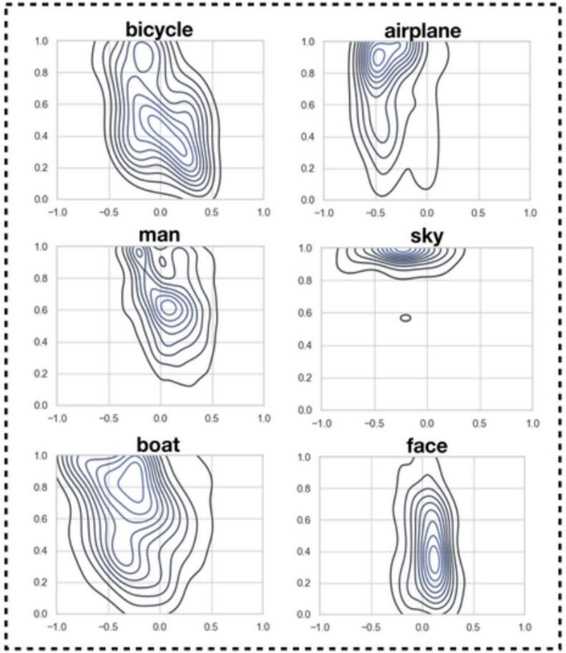
SWiG analysis

- 126,102 images
- 504 verbs
- ~10,000 nouns
- 451,916 noun slots
    - 435,566 are non empty
    - 278,336 (63.9%) have bounding boxes
    - Missing boxes for objects not visible or ‘Place’

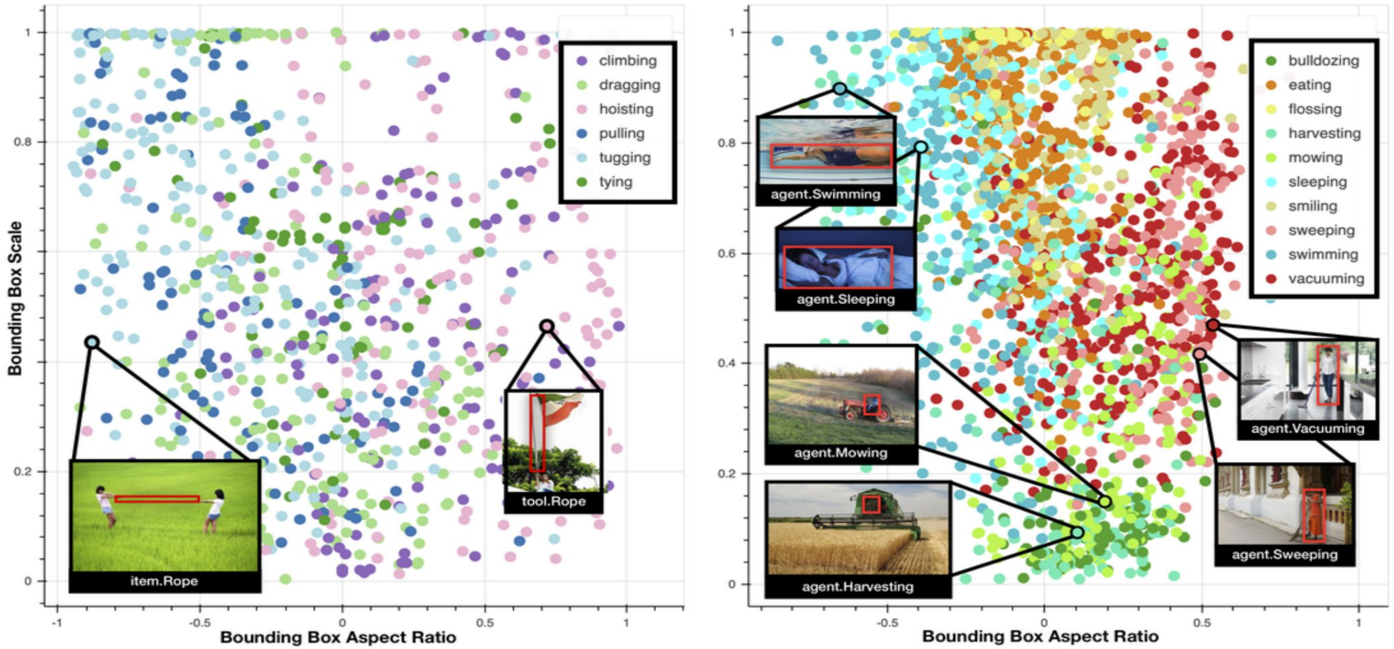
SWiG analysis

- Long tail noun groundings
- Different roles have different ratios of occurrences grounded
- Some nouns have strong priors w.r.t. scale and aspect ratio while many are diverse





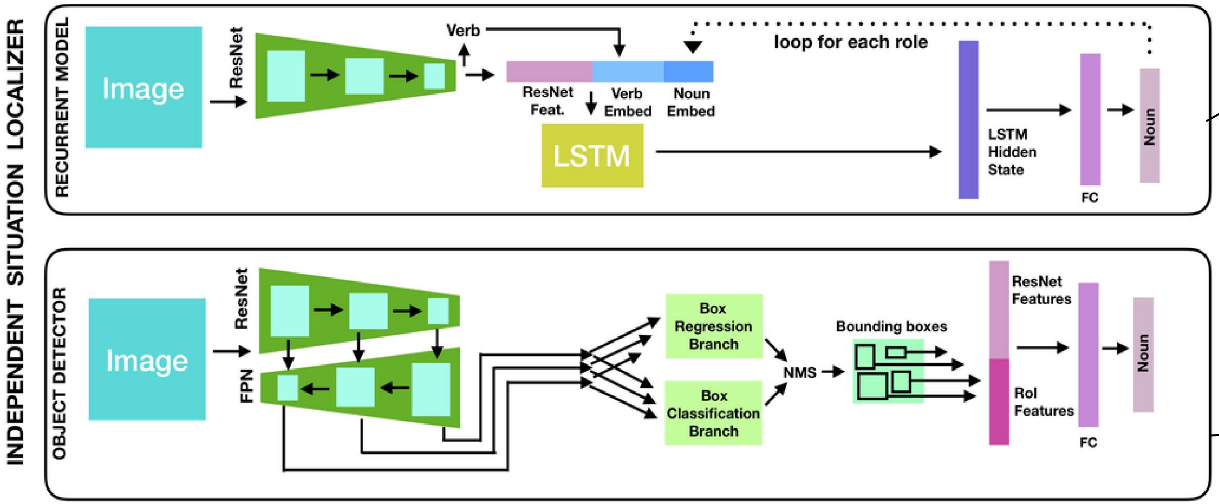
SWiG analysis



Verbs give strong priors

Models

ISL: Independent Situation Localization



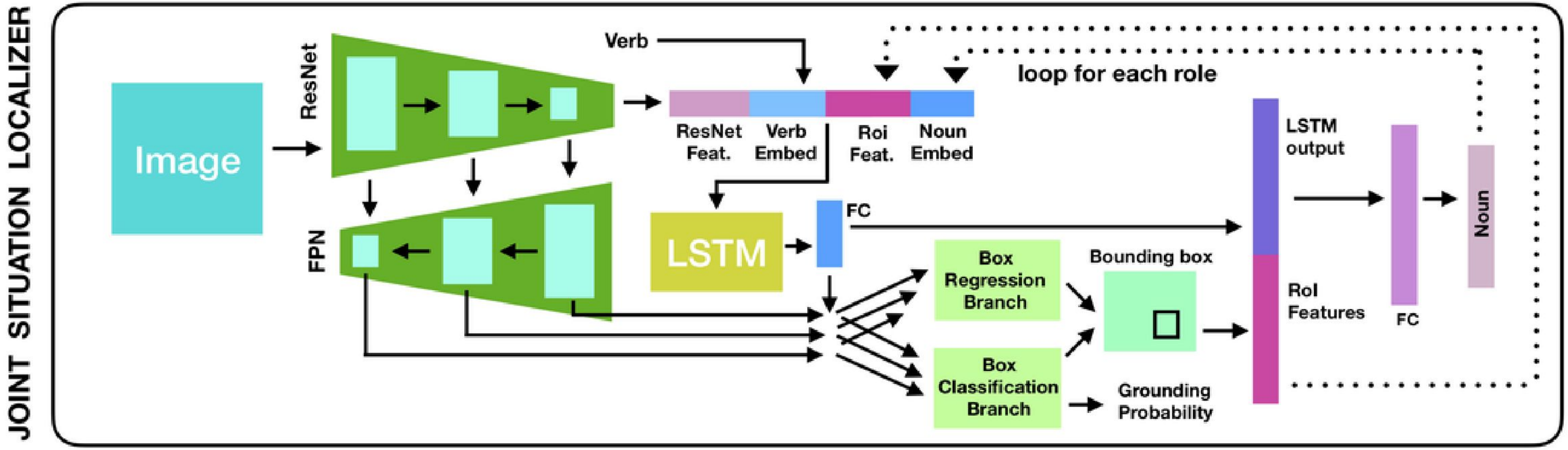
Existing model for situation recognition
Mallya, A., Lazebnik, S.: Recurrent models for situation recognition (2017)
(with updated CNN, label smoothing, hyperparameter tuning)

RetinaNet for object detection
Lin, T.Y. et. al.: Focal loss for dense object detection (2017)

- Situation recognition and object detection run independently
- Each generated noun matched with bounding box which scores that noun the most

Models

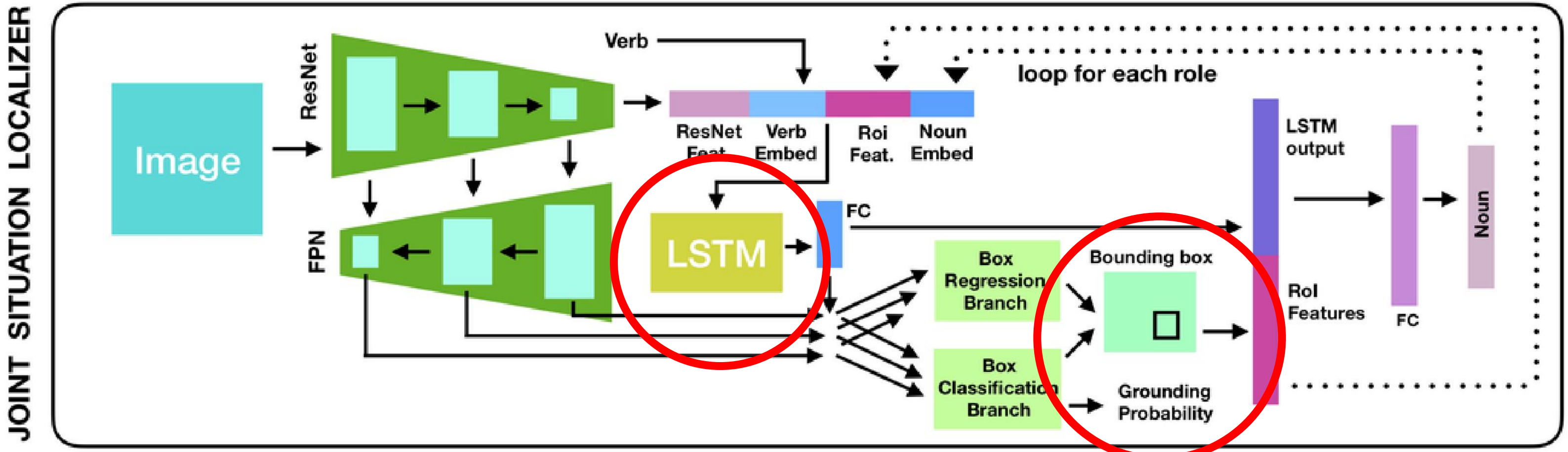
JSL: Joint Situation Localization



Generate nouns and boxes simultaneously

Models

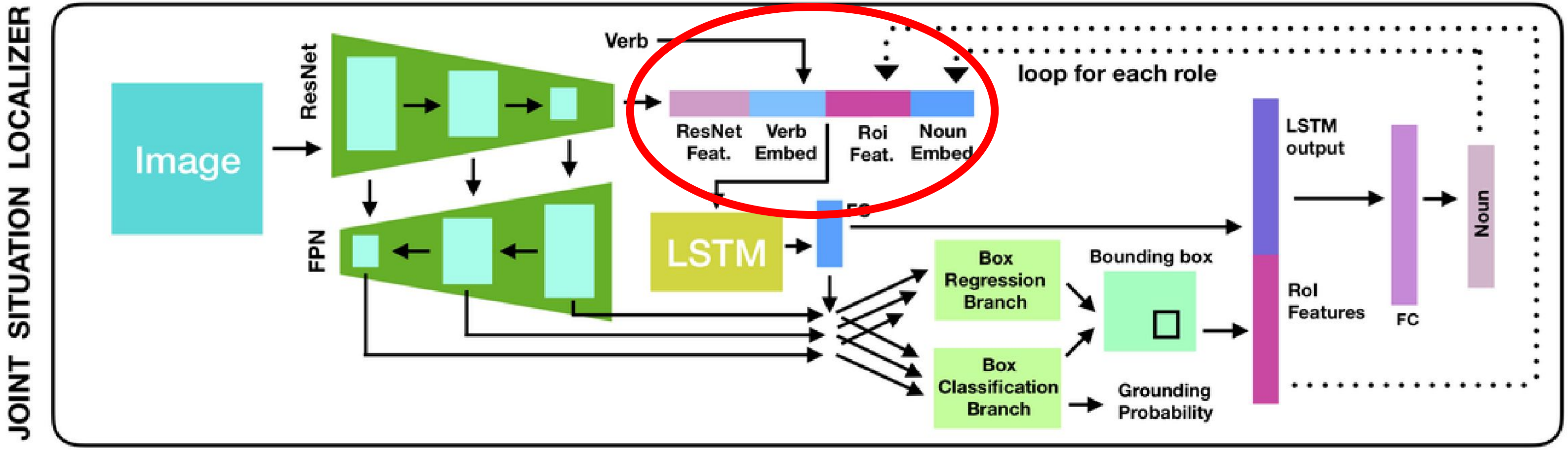
JSL: Joint Situation Localization



JSL localizes objects recurrently while ISL does it in independently

Models

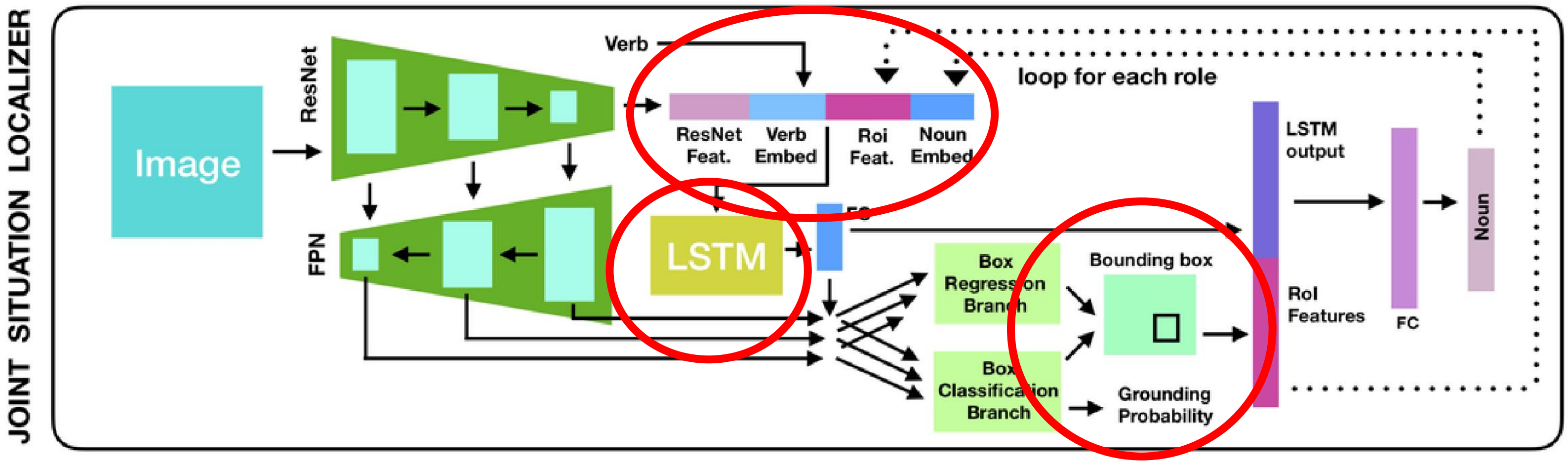
JSL: Joint Situation Localization



Additional input to the LSTM (ResNet features of previous bounding box)

Models

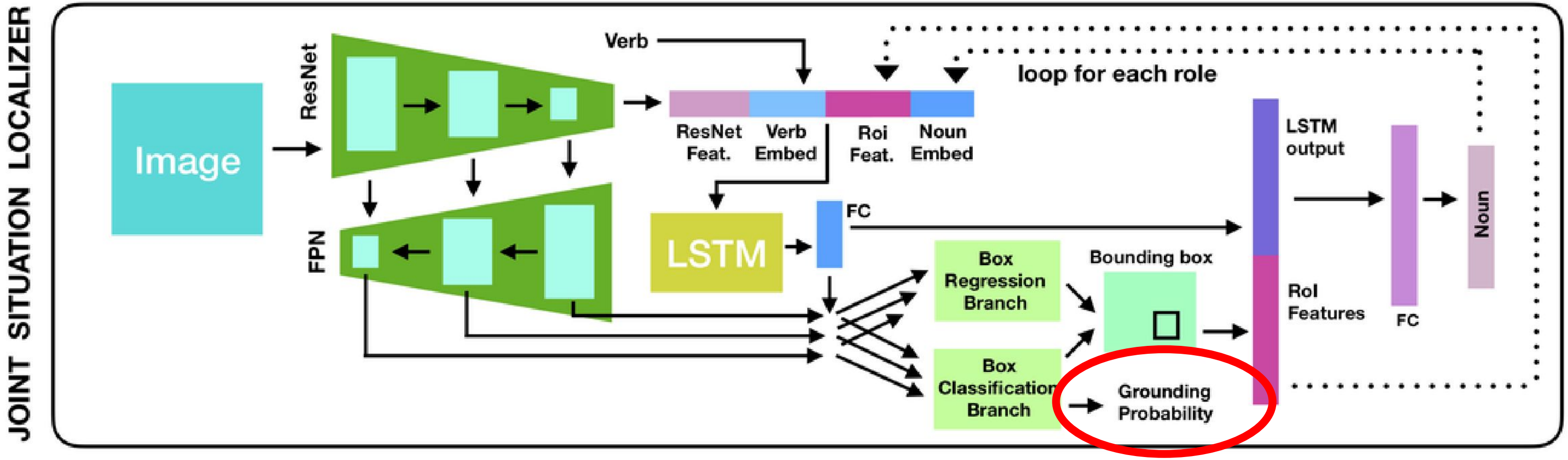
JSL: Joint Situation Localization



Object Detector conditioned on verb and role

Models

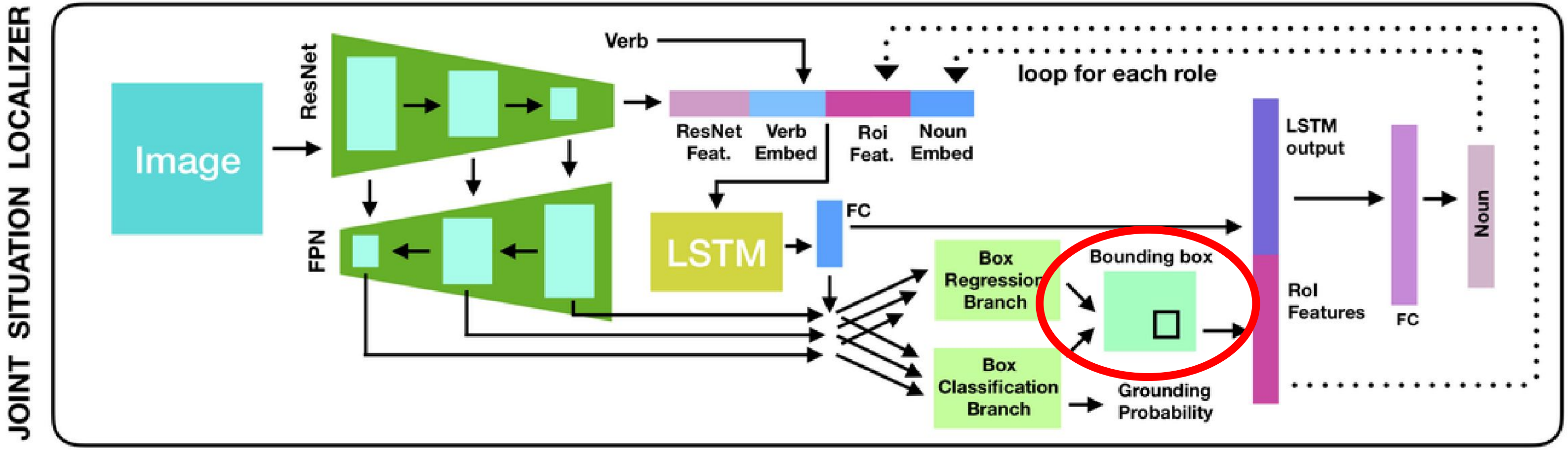
JSL: Joint Situation Localization



JSL explicitly generates grounding probability

Models

JSL: Joint Situation Localization



Only one box generated per noun generation

Experiments

- Both ISL and JSL use ResNet-50 for backbone
- Both have 108M parameters
- Gradient Descent with Adam
- 20h training on 4 24GB TITAN RTX GPUs

Metrics

- **VERB**: Verb accuracy
- **Value**: Noun accuracy for given role
- **Value-all**: Whole Frame accuracy
- **Grounded-value:** Noun accuracy with grounding
- **Grounded-value-all**: Frame accuracy with grounding

Results

Dev Set

Method	top-1 predicted verb					top-5 predicted verbs					ground truth verbs			
	verb	value	value-all	grnd value	grnd value-all	verb	value	value-all	grnd value	grnd value-all	value	value-all	grnd value	grnd value-all
Prior Models for Situation Recognition														
CRF [60]	32.25	24.56	14.28	-	-	58.64	42.68	22.75	-	-	65.90	29.50	-	-
CRF+Aug [59]	34.20	25.39	15.61	-	-	62.21	46.72	25.66	-	-	70.80	34.82	-	-
RNN w/o Fusion[36]	35.35	26.80	15.77	-	-	61.42	44.84	24.31	-	-	68.44	32.98	-	-
RNN w/ Fusion[36]	36.11	27.74	16.60	-	-	63.11	47.09	26.48	-	-	70.48	35.56	-	-
GraphNet [31]	36.93	27.52	19.15	-	-	61.80	45.23	29.98	-	-	68.89	41.07	-	-
Kernel GraphNet[51]	**43.21**	**35.18**	**19.46**	-	-	**68.55**	**56.32**	**30.56**	-	-	**73.14**	**41.48**	-	-
RNN based models														
RNN w/o Fusion [36]	35.35	26.80	15.77	-	-	61.42	44.84	24.31	-	-	68.44	32.98	-	-
Updated RNN*	38.83	30.47	18.23	-	-	65.74	50.29	28.59	-	-	72.77	37.49	-	-
ISL*	38.83	30.47	18.23	22.47	7.64	65.74	50.29	28.59	36.90	11.66	72.77	37.49	52.92	15.00
JSL*	**39.60**	**31.18**	**18.85**	**25.03**	**10.16**	**67.71**	**52.06**	**29.73**	**41.25**	**15.07**	**73.53**	**38.32**	**57.50**	**19.29**

- JSL improves all metrics

- Joint modeling helps better grounding, grounding metrics improve the most

- Interestingly, joint modeling helps improve frame metrics, indicating better understanding

- JSL achieves SOTA on GroundTruthVerbValue

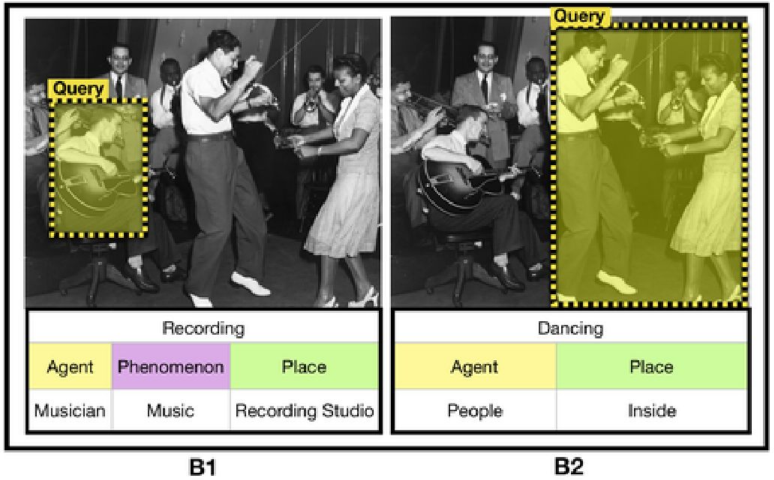
- JSL betters on value but worse on value-all, indicating it can work with partial information

Future Work

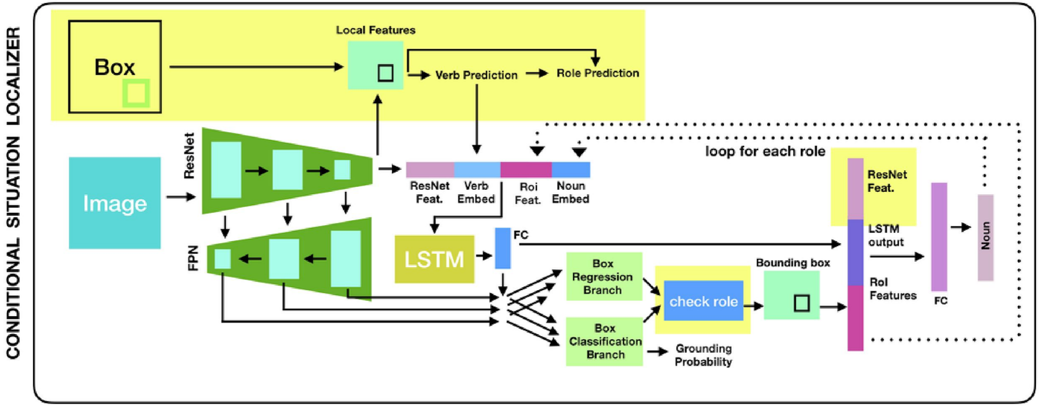
Grounded Semantic aware Image Retrieval

Future Work

Conditional Grounded Situation Recognition



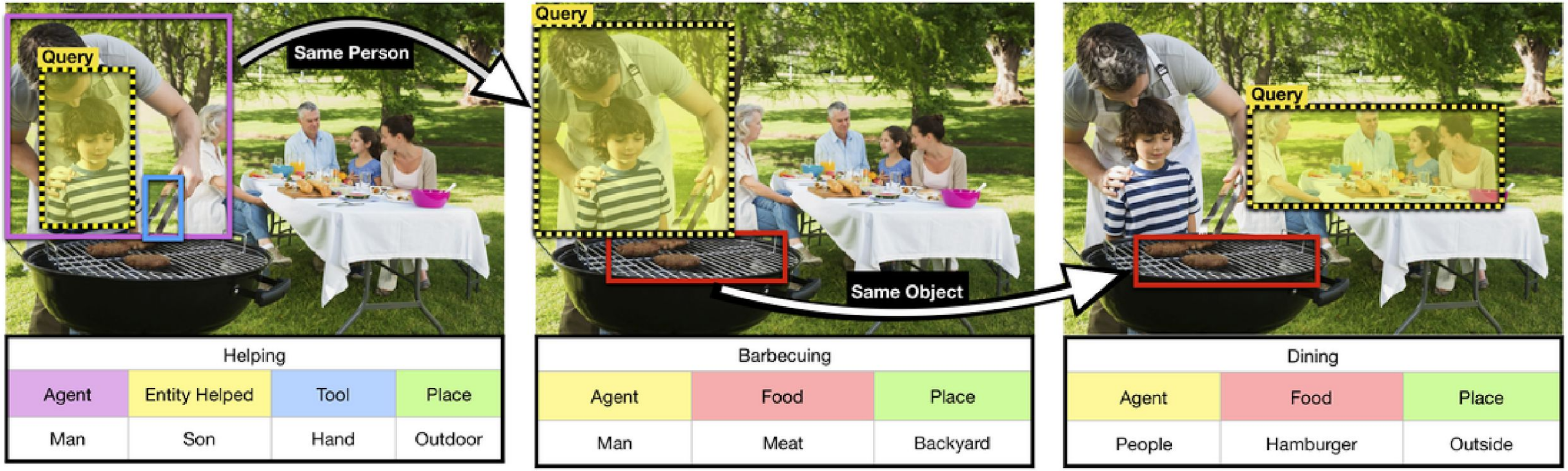
Example



Modified JSL

Future Work

Grounded Semantic Chaining (Future work)



Discussion

- The task seems very dependent on formal lexicons. How can it be made more flexible?
- Is detecting and grounding multiple frames a worthwhile task to pursue? What are the complexities involved?
- What about verbs having different meanings in different context? Current modeling proposes to predict verb first, but in those cases, can predicting 'participants' first be helpful?

Discussion

- Is conditional generation (on verb) the only practical way of generation?
- What is the relevance of this task compared to other image grounding problems? How do we rate them in terms of importance or applications?
- What are the benefits and shortcoming of structural generation vs natural language generation?

YOU
BELONG
HERE

CNS Diversity and Inclusion • cns.utexas.edu/diversity

Questions?