Multimodal CL and NLP: Combining Language and Vision for (Computational) Semantics

Carina Silberer

IMS, University of Stuttgart

CL Fall School 2024, Passau
16 - 20 September 2024

Outline

Outline

Semantic Role Labeling

Question: Which frames are evoked, what are the respective arguments and their roles (FrameNet)?

Who hosts the UEFA Euro2020 European Championship this year?

⇒ hosts: Frame **Provide_lodging**:

- Core Frame Elements:
  - [the UEFA Euro2020 European Championship] **Lodger**
  - [Who] **Host**
- Non-Core FEs: [this year] **Time**

⇒ the UEFA Euro2020 European Championship: Frame **Social_event**:

- Core Frame Elements: [UEFA Euro2020 European Championship] **Social_event** (Unexpressed)
- Non-Core FEs:
  - [Who] **Host**
  - [this year] **Time**

Semantic Role Labeling

Semantic Roles

- General term for thematic role sets of different size, roles of different level of abstractness
- Generalise over different surface realisations of predicate argument structures
- Used as shallow semantic representation
- Allow for simple inferences, to use as intermediate language

Semantic Role Labeling

Task

- Given a *sentence*
- Find the semantic roles of each argument of each predicate in a sentence
- Underlying subtasks:
  1. target (predicate) identification
  2. frame disambiguation
  3. syntactic argument (text spans) identification

FrameNet versus PropBank

Example adapted from SLP3 Jurafsky & Martin

[You]	can't [blame]	the program	[for being shallow]
COGNIZER	TARGET	EVALUAEE	REASON
[The San Francisco Examiner]	issued	[a special edition]	[yesterday]
ARG0	TARGET	ARG1	ARGM-TMP

Semantic Role Labeling: Early Neural Approach

Neural Semantic Role Labeling with Dependency Path Embeddings

PathLSMT (Roth & Lapata, 2016)

- Applicable to PropBank and FrameNet
- Based on Dependency Parse Paths

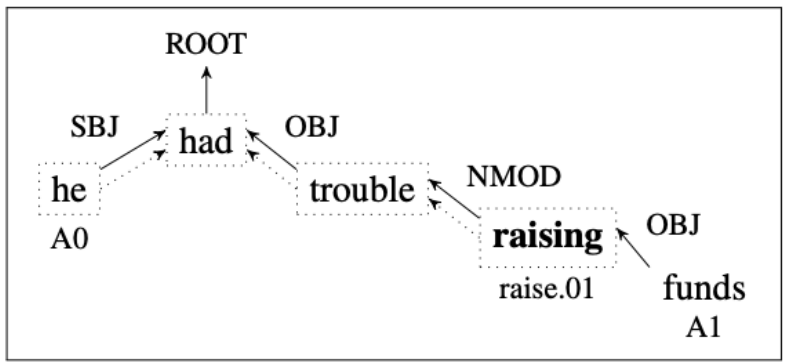


Figure 1: Dependency path (dotted) between the predicate *raising* and the argument *he*.

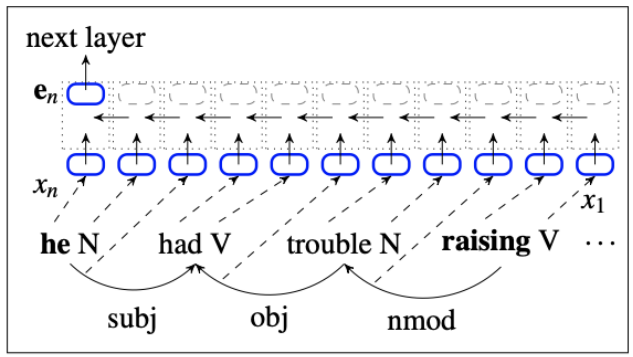


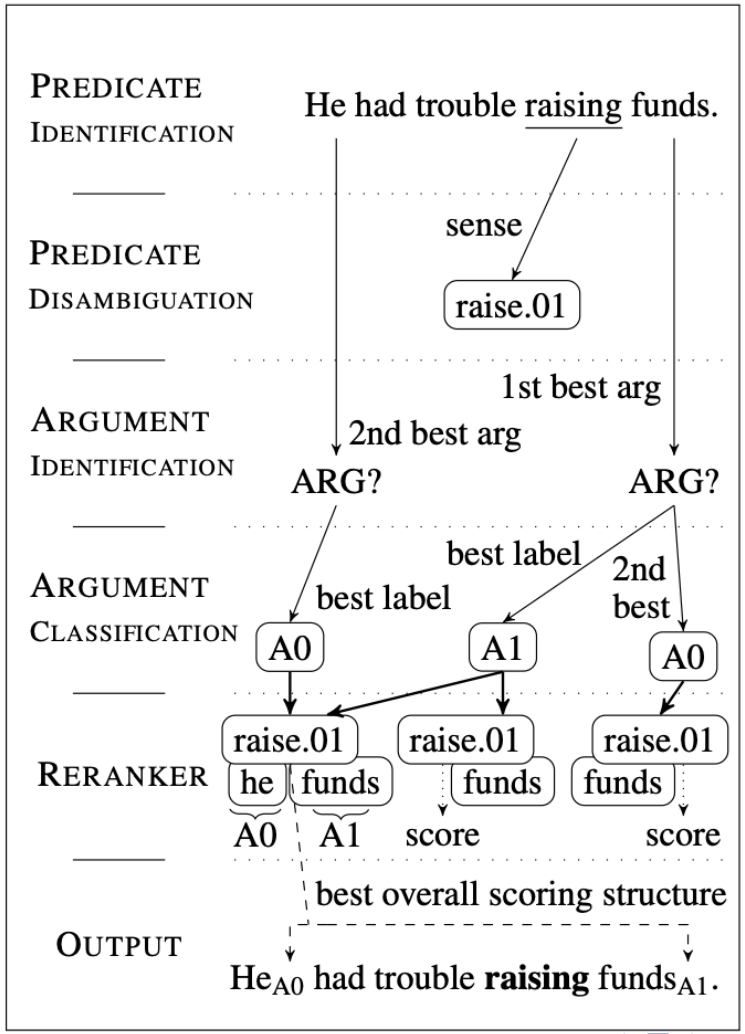
Figure 2: Example input and embedding computation for the path from *raising* to *he*, given the sentence *he had trouble raising funds*. LSTM time steps are displayed from right to left.

Semantic Role Labeling: Early Neural Approach

Neural Semantic Role Labeling with Dependency Path Embeddings

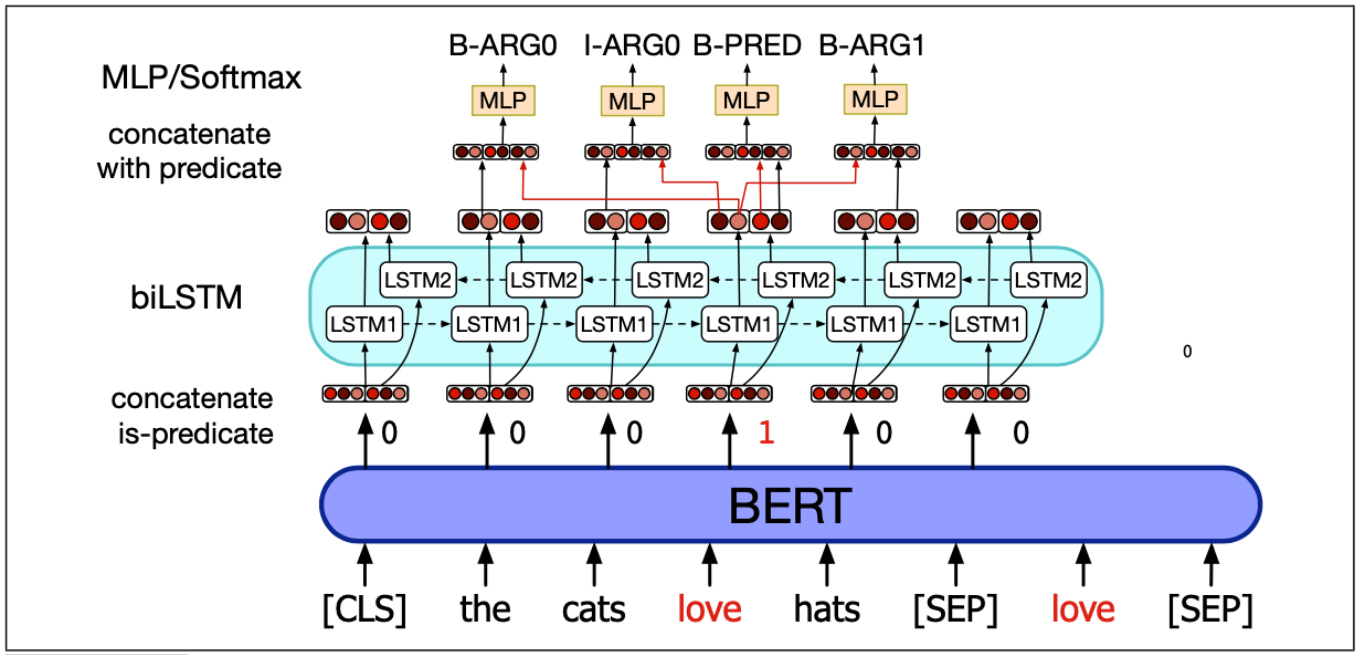
PathLSMT (Roth & Lapata, 2016)

- Applicable to PropBank and FrameNet



Semantic Role Labeling

Using Transformers



The Case of Implicit Semantic Roles

Question: Which vehicle doe the athletes row?

Some athletes are rowing on the water.

- FrameNet Frame: Operate_vehicle
- Roles: Area, Driver, Vehicle

⇒ Which argument fills the Core FE Vehicle?

The Case of Implicit Semantic Roles

Question: Who has to mix What with the dry ingredients into What?

Mix with the dry ingredients carefully.

- FrameNet Frame: Cause_to_amalgamate
- Roles: Agent, Part_1 and Part_2 or Parts, Whole

⇒ Which argument fills the Core FE Part_1?
⇒ Which argument fills the Core FE Whole?
⇒ Which argument fills the Core FE Agent?

The Case of Implicit Semantic Roles [Petruck, 2019]

Semantic Role Labeling is defined on sentence-level.

Some arguments are realised *implicitly*

- Called *Null Instantiations* (NIs) in FrameNet

The Case of Implicit Semantic Roles [Petruck, 2019]

Semantic Role Labeling is defined on sentence-level.

Some arguments are realised *implicitly*

- Called *Null Instantiations* (NIs) in FrameNet
- Defnite NIs: Core REs, realised in or inferrable from discourse
- Indefinite NIs: licensed by lexical specification
  e.g., intransitively used transitive verbs, such as eat, row
- Constructional NIs: licensed by grammatical constructions
  e.g., Agent in passive sentences, subject in imperative

Visual Semantic Roles

Question: Which argument fills the Core FE vehicle?

Some athletes are rowing on the water.

- Frame: Operate_vehicle
- Roles: Area, Driver, Vehicle

Semantic Understanding of Visual Situations

Problem Conceptualisations

Generally: Recognising the activity, event or situation that takes place in an image, the involved actors and objects, and the roles they play.

- Grounded Semantic Role Labeling / Visual Semantic Role Labeling
- Situation Recognition
- Grounded Situation Recognition

Visual Semantic Role Labeling

[Gupta and Malik, 2015, Yang et al., 2016, Silberer and Pinkal, 2018]

Also known as *Grounded VSRL*

Task Definition

Given an image, determine the semantic roles of each participant (filler) of each frame (situation or event) that the image evokes.

- Subtasks:
  1. **Frame identification**
     cf. linguistic SRL using FrametNet: an additional frame identification step is required
     cf. Computer Vision task of activity recognition

Visual Semantic Role Labeling

[Gupta and Malik, 2015, Yang et al., 2016, Silberer and Pinkal, 2018]

Also known as *Grounded VSRL*

Task Definition

Given an image, determine the semantic roles of each participant (filler) of each frame (situation or event) that the image evokes.

- Subtasks:
  1. **Frame identification**
     cf. linguistic SRL using FrameNet: an additional frame identification step is required
     cf. Computer Vision task of activity recognition
  2. **Role Fillers/Referents identification**
     Which objects in the image play a role in the situation/event (the identified frame)?
     ⇒ Localise the participating objects in the image

Visual Semantic Role Labeling

[Gupta and Malik, 2015, Yang et al., 2016, Silberer and Pinkal, 2018]

Also known as *Grounded VSRL*

Task Definition

Given an image, determine the semantic roles of each participant (filler) of each frame (situation or event) that the image evokes.

- Subtasks:
  1. **Frame identification**
     cf. linguistic SRL using FrameNet: an additional frame identification step is required
     cf. Computer Vision task of activity recognition
  2. **Role Fillers/Referents identification**
     Which objects in the image play a role in the situation/event (the identified frame)?
     ⇒ Localise the participating objects in the image
  3. **Semantic Role determination**
     Which semantic role does each object play?
     ⇒ Classify the localised objects
- Subtasks 2 and 3: cf. Computer Vision task of object detection

Visual Semantic Role Labeling (VSRL)

[Gupta and Malik, 2015, Yang et al., 2016, Silberer and Pinkal, 2018]

Also known as *Grounded VSRL*

Task Definition

Given an image, determine the semantic roles of each participant (filler) of each frame (situation or event) that the image evokes.

ARREST		PLACING	
r_1, r_2	Authorities	r_1	Agent
r_5	Suspect	r_5	Theme
r_3	Place	r_3	Place
		r_4	Goal

[Silberer and Pinkal, 2018]

Situation Recognition

[Yatskar et al., 2016]

VSRL and *Situation Recognition* are defined slightly differently:

Task Definition: Grounded Situation Recognition

Given an image, determine the frame the image evokes, and the participants (fillers) of the semantic roles of the frame.

- Subtasks:
  1. **Frame identification**
     ↔ One single ground truth frame is assumed
  2. **Semantic Role Filler determination**
     For each role associated with the frame, determine the WordNet synset of its filler
     ↔ No localisation of the objects in the image (i.e., no explicit grounding of roles)

Situation Recognition

[Yatskar et al., 2016]

VSRL and *Situation Recognition* are defined slightly differently:

Task Definition: Situation Recognition

Given an image, determine the frame the image evokes, and the participants (fillers) of the semantic roles of the frame.

CLIPPING			
ROLE	VALUE	ROLE	VALUE
AGENT	MAN	AGENT	VET
SOURCE	SHEEP	SOURCE	DOG
TOOL	SHEARS	TOOL	CLIPPER
ITEM	WOOL	ITEM	CLAW
PLACE	FIELD	PLACE	ROOM

[Yatskar et al., 2016]

Visual Semantic Role Labeling

"Grounded Situation Recognition" [Pratt et al., 2020a]

Task Definition:

Given an image, determine

- the frame (verb) the image evokes, and e.g., hitchhiking
- the participants (fillers) of the semantic roles of the frame:
  - WordNet synset and e.g., agent: n10287213
  - localisation in image e.g., agent: [195, 282, 380, 554]

Catching			
Agent	Caught Item	Tool	Place
Bear	Fish	Mouth	River

Jumping				
Agent	Source	Destination	Obstacle	Place
Female Child	Sofa	Sofa	Ø	Living Room

Kneading		
Agent	Item	Place
Person	Dough	Kitchen

[Pratt et al., 2020a]

Visual Semantic Role Labeling

"Grounded Situation Recognition" [Pratt et al., 2020a]

Task Definition:

Given an image, determine

- the frame (verb) the image evokes, and e.g., hitchhiking
- the participants (fillers) of the semantic roles of the frame:
  - WordNet synset and e.g., agent: n10287213
  - localisation in image e.g., agent: [195, 282, 380, 554]

```
"hitchhiking_238.jpg": {
"verb": "hitchhiking",
"height": 683,
"width": 512,
"bb": {"place": [-1, -1, -1, -1], "agent": [195, 282, 380, 554]},
"frames": [{"place": "n03519981", "agent": "n10287213"},
           {"place": "n03519981", "agent": "n10287213"},
           {"place": "n04096066", "agent": "n10287213"}]
}
```

Data item in SWiG dataset

(Ungrounded) Situation Recognition

ImSitu Dataset [Yatskar et al., 2016]

Demo: vision-explorer.allenai.org/situation_recognition

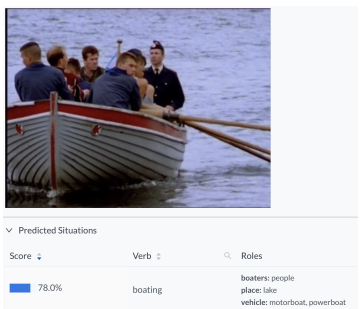
Predicted Situations

Score	Verb	Roles
78.0%	boating	**boaters:** people **place:** lake **vehicle:** motorboat, powerboat

(Ungrounded) Situation Recognition

ImSitu Dataset [Yatskar et al., 2016]

Demo: vision-explorer.allenai.org/situation_recognition



14.5%	rowing	**agent:** athlete, jock **place:** river **vehicle:** boat
3.3%	rafting	**agent:** people **place:** river
2.4%	floating	**agent:** people **medium:** water, H2O **place:** river **tool:** boat
1.8%	boarding	**agent:** people **place:** body of water, water **vehicle:** boat

Visual Semantic Role Labeling

"Grounded Situation Recognition" [Pratt et al., 2020a]

Situation With Groundings (SWiG) Dataset [Pratt et al., 2020a]

- Created on top of ImSitu [Yatskar et al., 2016]
- 451,916 noun slots, of which
  435,566 are filled with synset (noun)
- Bounding box annotations were crowdsourced (AMT)
  - each role annotated by 3 workers
  - final box results from averaging the 3 annotated extents
- Code and data: SWiG (link)

Visual Semantic Role Labeling

SWiG Dataset for Grounded Situation Recognition

[Pratt et al., 2020a]

Code and data: SWiG

Demo: https://prior.allenai.org/projects/gsr

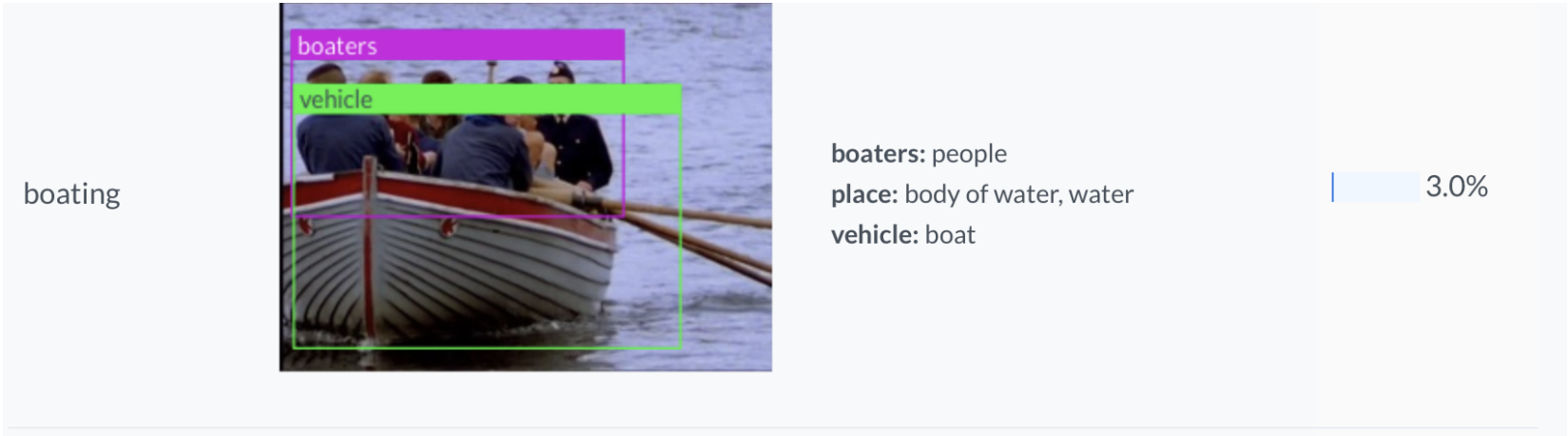
Visual Semantic Role Labeling

SWiG Dataset for Grounded Situation Recognition

[Pratt et al., 2020a]

Code and data: SWiG

Demo: https://prior.allenai.org/projects/gsr

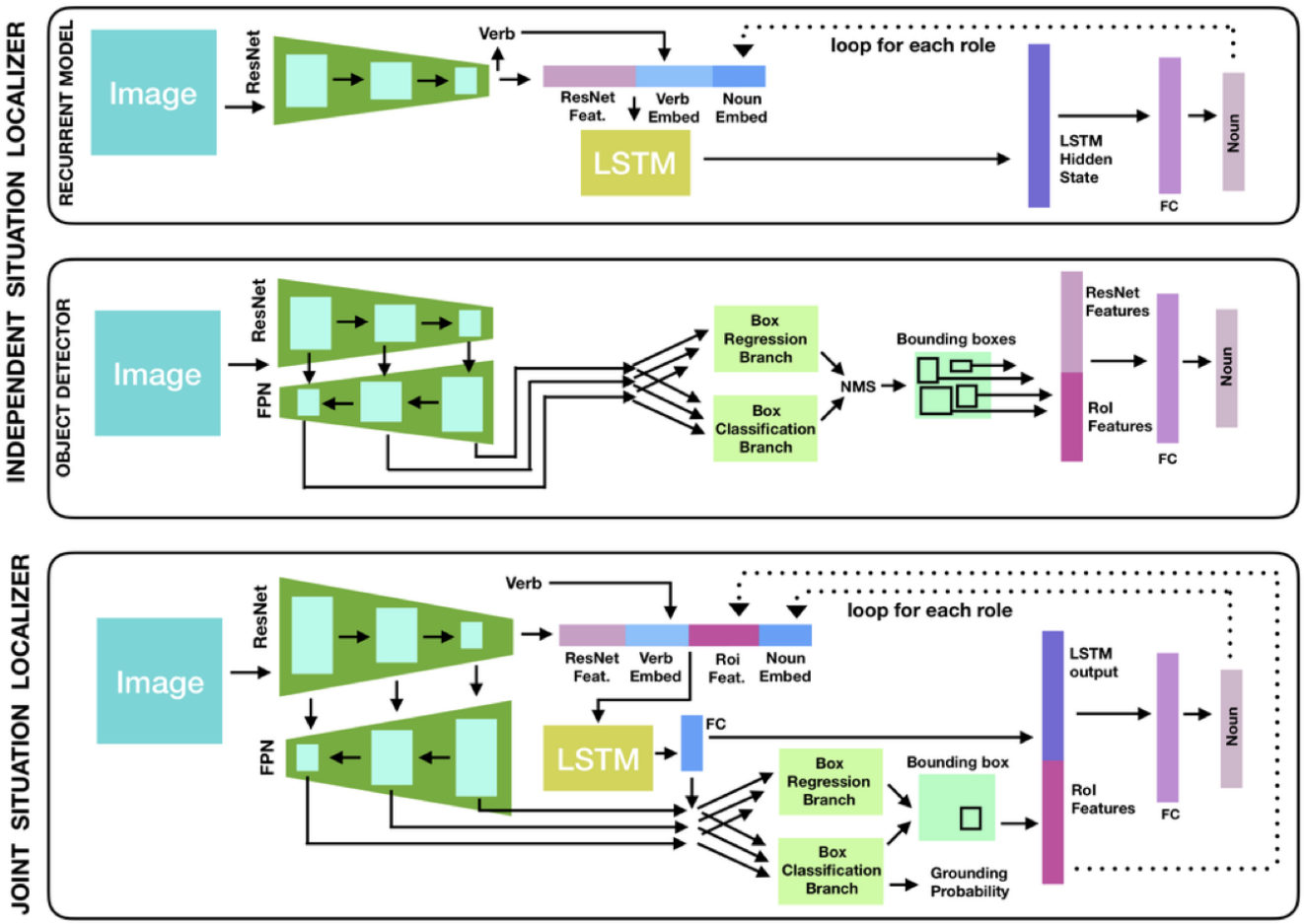


Visual Semantic Role Labeling

"Grounded Situation Recognition" [Pratt et al., 2020a]

Methods: ISL and JSL [Pratt et al., 2020a]

Demo: https://prior.allenai.org/projects/gsr



Visual Semantic Role Labeling

"Grounded Situation Recognition" [Pratt et al., 2020a]

SWiG Task Definition: Details

Given an image, determine the frame the image evokes, and the participants (fillers) of the semantic roles of the frame.

- Subtasks:
  1. **Frame identification**
     ↔ One single ground truth frame is assumed
  2. **Semantic Role Filler determination**
     For each role associated with the frame, determine the filler noun (WordNet synset)

Visual Semantic Role Labeling

"Grounded Situation Recognition" [Pratt et al., 2020a]

SWiG Task Definition: Details

Given an image, determine the frame the image evokes, and the participants (fillers) of the semantic roles of the frame.

- Subtasks:
  1. **Frame identification**
     ↔ One single ground truth frame is assumed
  2. **Semantic Role Filler determination**
     For each role associated with the frame, determine the filler noun (WordNet synset)
  3. **Role grounding**
     Localise the filler objects, if possible (box coordinates)
     ↔ Some fillers may be *ungrounded*

Visual Semantic Role Labeling

"Grounded Situation Recognition" [Pratt et al., 2020a]

Experiments: Example Prediction & Evaluation

⇒ Ground truth (3 annotations):

Verb: whisking

agent	`[man, cook, chef]`
item	`[egg, egg, egg]`
container	`[bowl, bowl, bowl]`
place	`[kitchen, chopping_board, kitchen]`

(gt bounding boxes not shown)

⇒ Prediction:

Verb: whisking

agent	`person`
item	`batter`
container	`bowl`
place	`table`

Visual Semantic Role Labeling

"Grounded Situation Recognition" [Pratt et al., 2020a]

Experiments: Localisation Errors

Localization Errors

Painting			
Agent	Item	Tool	Place
Woman	Wall	Paint Roller	Room

Blocking			
Blocker	Blocked	Tool	Place
Football Player	Football Player	Arm	Football Field

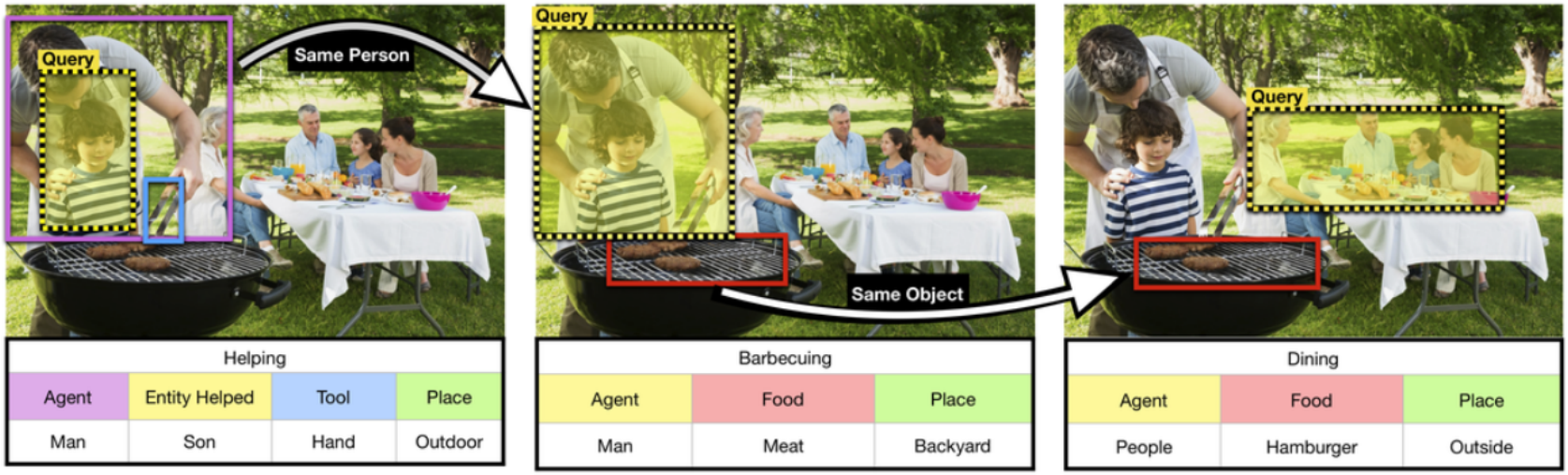
Signing			
Agent	Signed Item	Tool	Place
Person	Document	Pen	Ø

Visual Semantic Role Labeling

"Grounded Situation Recognition" [Pratt et al., 2020a]

Results: Grounded Semantic Chaining

Multiple, interrelated situations may be evoked by an image



Visual Semantic Role Labeling

"Grounded Situation Recognition" [Pratt et al., 2020a]

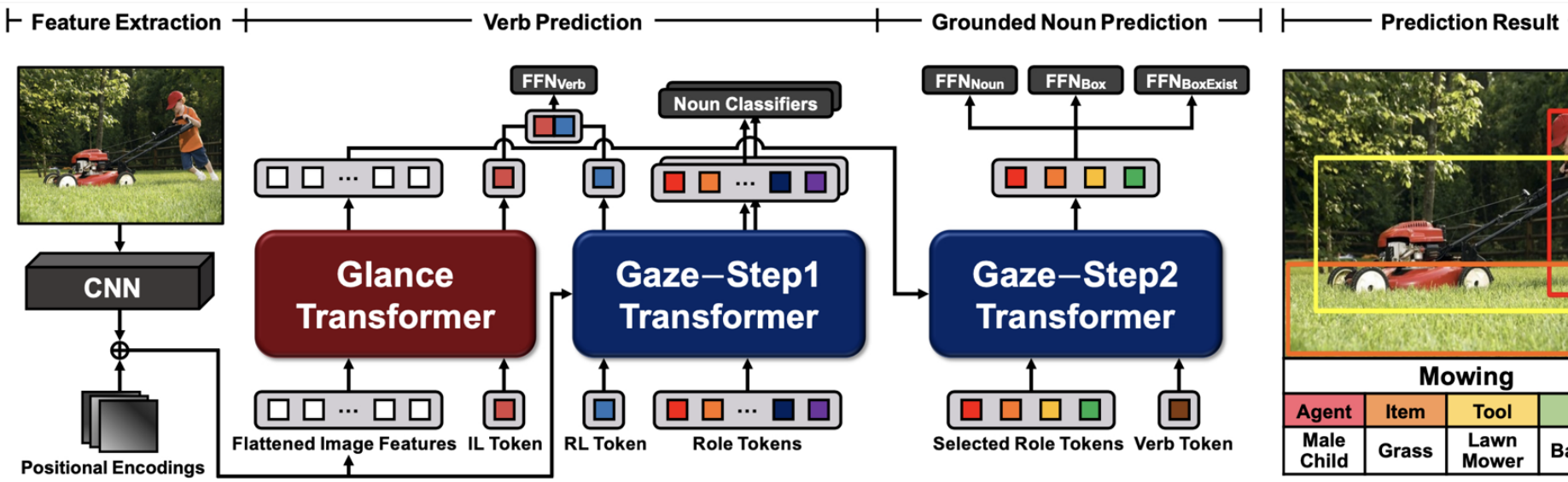
Summary

- Promissing results on grounded/visual semantic role labeling
- Can be useful for deeper semantic understanding of visuals with respect to the shown scene overall (situations and their relation), the participants and their verbalisation therein (by nouns)
- However, all subtasks figure difficult – still a huge gap to satisfying results:
  situation prediction (frames), filler naming (noun prediction), and filler/noun localisation

Visual Semantic Role Labeling

"Collaborative Transformers for Grounded Situation Recognition" [Cho et al., 2022]

CoFormer

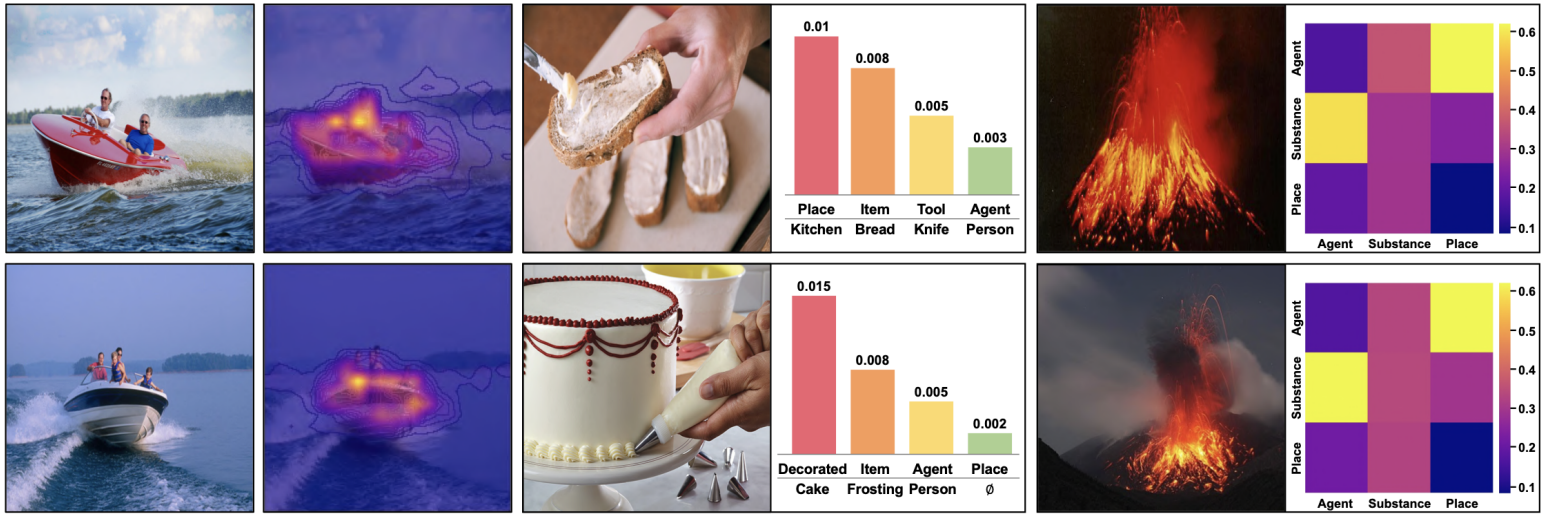


- Two closely interacting modules: Glance transformer + Gaze transformer
- Glance transformer: verb prediction by leveraging the Gaze transformer
- Gaze transformer: entity grounding by focusing only on the entities relevant to the predicted verb

Visual Semantic Role Labeling

"Collaborative Transformers for Grounded Situation Recognition" [Cho et al., 2022]

CoFormer



(a) Attention Scores (IL Token) (b) Attention Scores (RL Token) (c) Attention Scores (Frame-Role Queries)

Figure 7. Attention scores from IL token to image features, from RL token to role features, and on frame-role queries. We visualize the attention scores computed from the last self-attention layer of the encoder in Glance transformer, the encoder in Gaze-S1 transformer, and the decoder in Gaze-S2 transformer, respectively. Higher attention scores are highlighted in red color on images.

Visual Semantic Role Labeling

"Collaborative Transformers for Grounded Situation Recognition" [Cho et al., 2022]

CoFormer

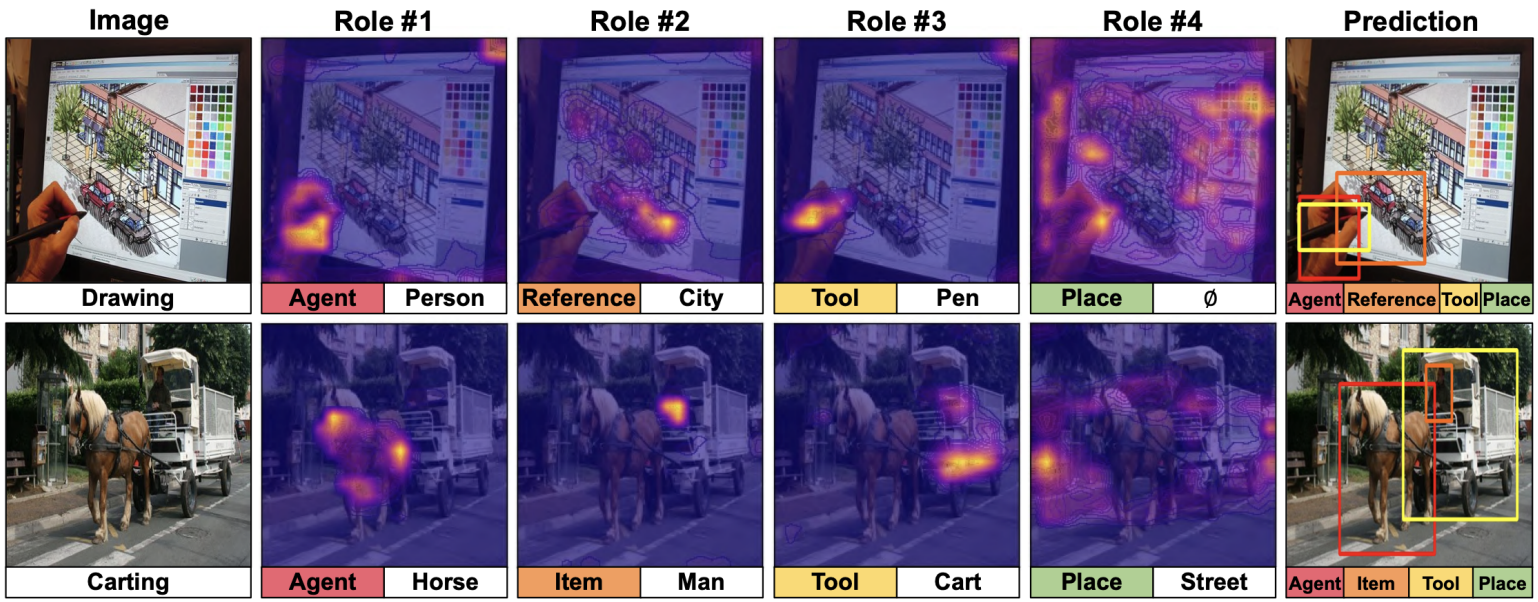


Figure 8. Attentions scores from frame-role queries to image features. We visualize the attention scores computed from the last cross-attention layer of the decoder in Gaze-S2 transformer. Higher attention scores are highlighted in red color on images.

Further References

Image Data (ImSitu/SWiG Data, inter alia)

- Knowledge-Aware Global Reasoning for Situation Recognition [Yu et al., 2023] (TPAMI)
  Addresses problem of long-tail data distribution in terms of "noun" classification
- Attention-Based Context Aware Reasoning for Situation Recognition [Cooray et al., 2020]
- Human-Like Controllable Image Captioning with Verb-Specific Semantic Roles [Chen et al., 2021]
- Teaching Structured Vision & Language Concepts to Vision & Language Models [Doveh et al., 2023]

Visual Semantic Role Labeling in Videos

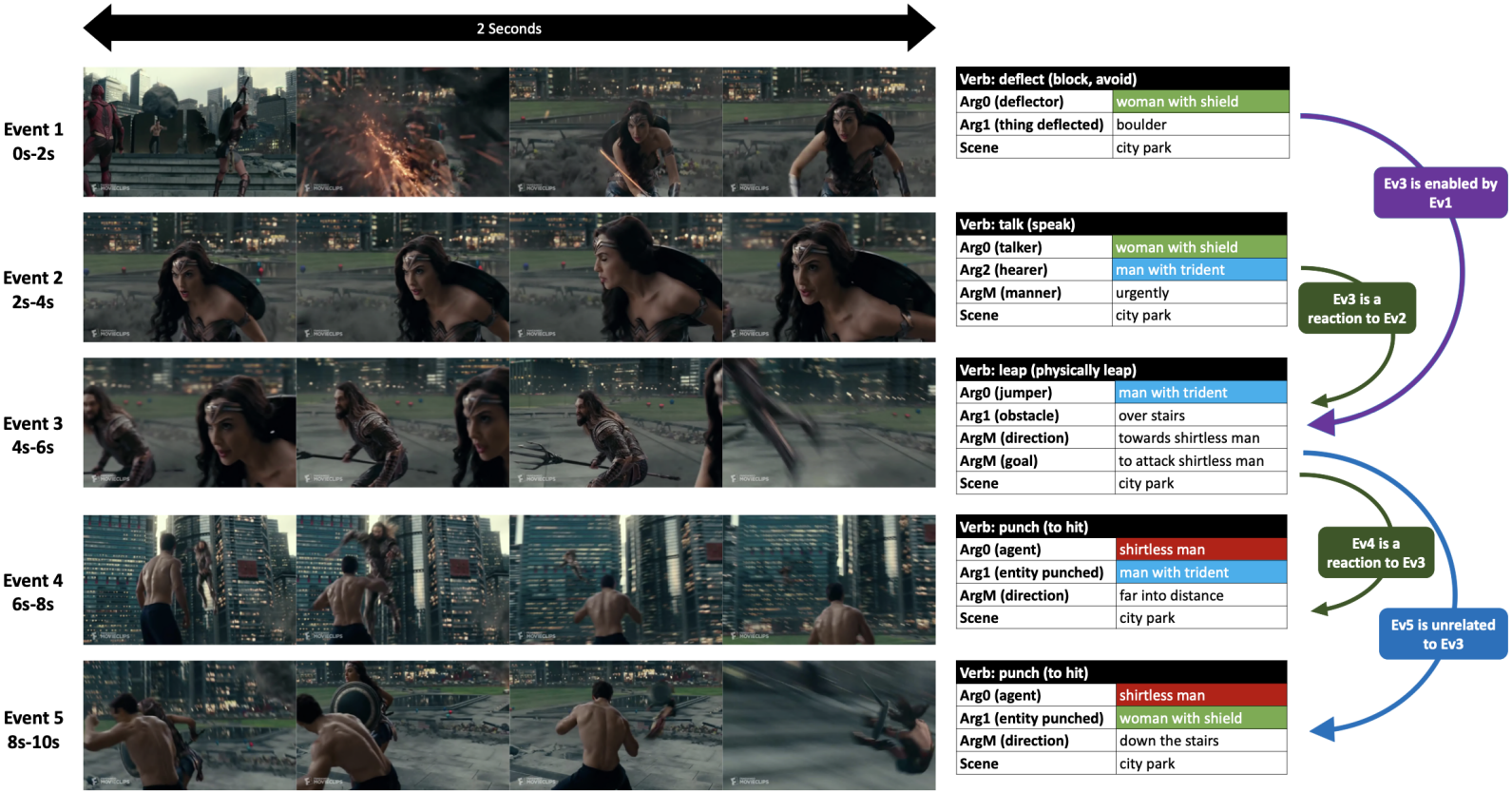
[Regneri et al., 2013, Rohrbach et al., 2016, Yang et al., 2016, Shutova et al., 2017, Sadhu et al., 2020, Sadhu et al., 2021, Khan et al., 2022]

VidSitu

[Sadhu et al., 2021]

- Uses PropBank framework

see also [Sadhu et al., 2020]



[Sadhu et al., 2021]

Visual Semantic Role Labeling in Videos

[Regneri et al., 2013, Rohrbach et al., 2016, Yang et al., 2016, Shutova et al., 2017, Sadhu et al., 2020, Sadhu et al., 2021, Khan et al., 2022]

VidSitu

[Sadhu et al., 2021]

- Uses PropBank framework

see also [Sadhu et al., 2020]

Visual Semantic Role Labeling in Videos

[Regneri et al., 2013, Rohrbach et al., 2016, Yang et al., 2016, Shutova et al., 2017, Sadhu et al., 2020, Sadhu et al., 2021, Khan et al., 2022]

VidSitu [Sadhu et al., 2021]

- Uses PropBank framework see also [Sadhu et al., 2020]
- Grounded Video Situation Recognition [Khan et al., 2022]
  github: `https://zeeshank95.github.io/grvidsitu`

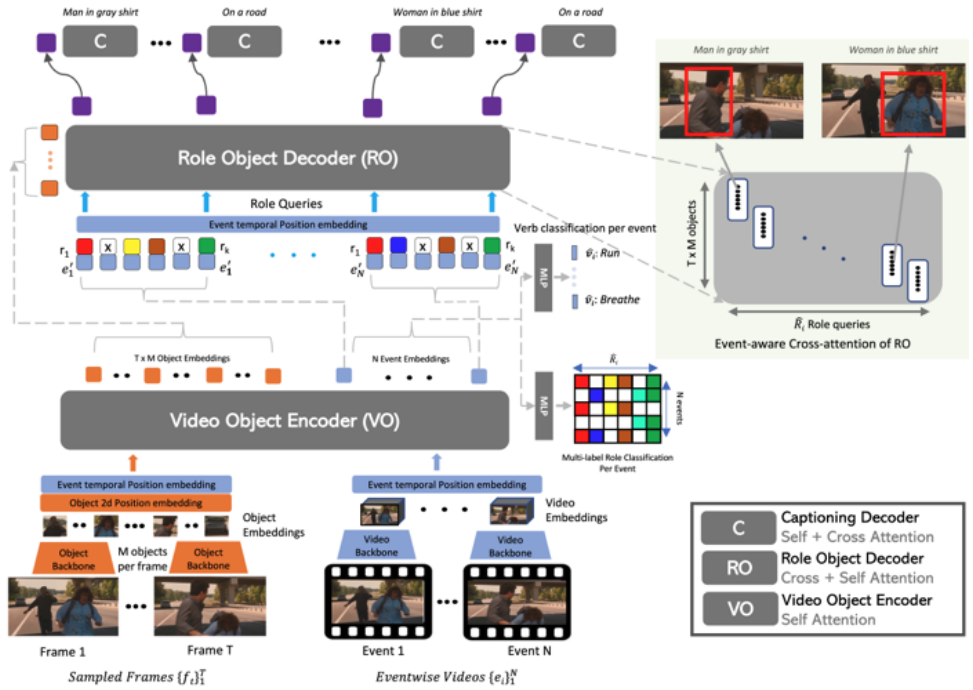
Visual Semantic Role Labeling in Videos

[Regneri et al., 2013, Rohrbach et al., 2016, Yang et al., 2016, Shutova et al., 2017, Sadhu et al., 2020, Sadhu et al., 2021, Khan et al., 2022]

VidSitu [Sadhu et al., 2021]

- Uses PropBank framework see also [Sadhu et al., 2020]
- Grounded Video Situation Recognition [Khan et al., 2022]
  github: `https://zeeshank95.github.io/grvidsitu`

VideoWhisperer



[Khan et al., 2022]

References

Outline

Multimodal Image–Text Summarisation

How does the problem relate to cross-modal relations?

- What kind of relationships are beneficial for multimodal summarisation?
- What kind of relationships make visual information obsolete?

References

- `https://www.ijcai.org/proceedings/2018/0577.pdf`
- `https://arxiv.org/pdf/2403.03823`

References I

- Benchmarking Vision Language Models for Cultural Understanding https://arxiv.org/pdf/2407.10920
- Learning Action and Reasoning-Centric Image Editing from Videos and Simulations https://arxiv.org/pdf/2407.03471
- Reframing linguistic bootstrapping as joint inference using visually-grounded grammar induction models https://arxiv.org/pdf/2406.11977
- Inferring Social Media Users' Mental Health Status from Multimodal Information https://aclanthology.org/2020.lrec-1.772.pdf
- Hate speech https://aclanthology.org/W19-3502.pdf
- SemEval-2022 Task 9: R2VQ – Competence-based Multimodal Question Answering [Pusejovsky et al., 2022] https://competitions.codalab.org/competitions/34056

References II

- Abstractness https://aclanthology.org/2024.cogalex-1.12.pdf
- Sarcasm https://arxiv.org/pdf/1906.01815
- Sarcasm ¡Qué maravilla! Multimodal Sarcasm Detection in Spanish: a Dataset and a Baseline https://arxiv.org/pdf/2105.05542
- Mental health analysis https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7147779/
- YouMakeup: A Large-Scale Domain-Specific Multimodal Dataset for Fine-Grained Semantic Comprehension https://aclanthology.org/D19-1517.pdf
- Bag-of-Lies: A Multimodal Dataset for Deception Detection https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9025340

References I

Cooray, T., Cheung, N.-M., and Lu, W. (2020).
Attention-Based Context Aware Reasoning for Situation Recognition.
In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4736–4745.

Gupta, S. and Malik, J. (2015).
Visual semantic role labeling.
CoRR, abs/1505.04474.

Khan, Z., Jawahar, C., and Tapaswi, M. (2022).
Grounded video situation recognition.
In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K., editors, *Advances in Neural Information Processing Systems*.

Mallya, A. and Lazebnik, S. (2017).
Recurrent models for situation recognition.
In *Proceedings of the IEEE International Conference on Computer Vision*, pages 455–463.

Pratt, S., Yatskar, M., Weihs, L., Farhadi, A., and Kembhavi, A. (2020a).
Grounded situation recognition.
ArXiv, abs/2003.12058.

Pratt, S., Yatskar, M., Weihs, L., Farhadi, A., and Kembhavi, A. (2020b).
Grounded situation recognition.
In *European Conference on Computer Vision*, pages 314–332. Springer.

Pusejovsky, J., Tu, J., Maru, M., Conia, S., Navigli, R., Rim, K., Lynch, K., Brutti, R., and Holderness, E. (2022).
SemEval-2022 Task 9: R2VQ - competence-based multimodal question answering.
In *Proceedings of the 16th Workshop on Semantic Evaluation (SemEval-2022)*.

References II

Sadhu, A., Chen, K., and Nevatia, R. (2020).
Video object grounding using semantic roles in language description.
In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sadhu, A., Gupta, T., Yatskar, M., Nevatia, R., and Kembhavi, A. (2021).
Visual semantic role labeling for video understanding.
In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Shutova, E., Wundsam, A., and Yannakoudakis, H. (2017).
Semantic frames and visual scenes: Learning semantic role inventories from image and video descriptions.
In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 149–154, Vancouver, Canada. Association for Computational Linguistics.

Silberer, C. and Pinkal, M. (2018).
Grounding semantic roles in images.
In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2616–2626.

Yang, S., Gao, Q., Liu, C., Xiong, C., Zhu, S.-C., and Chai, J. (2016).
Grounded semantic role labeling.
In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 149–159.

Yatskar, M., Ordonez, V., Zettlemoyer, L., and Farhadi, A. (2017).
Commonly uncommon: Semantic sparsity in situation recognition.
In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

References III

Yatskar, M., Zettlemoyer, L., and Farhadi, A. (2016).
Situation recognition: Visual semantic role labeling for image understanding.
In *Conference on Computer Vision and Pattern Recognition*.

Yu, W., Wang, H., Li, G., Xiao, N., and Ghanem, B. (2023).
Knowledge-aware Global Reasoning for Situation Recognition.
IEEE Transactions on Pattern Analysis and Machine Intelligence.