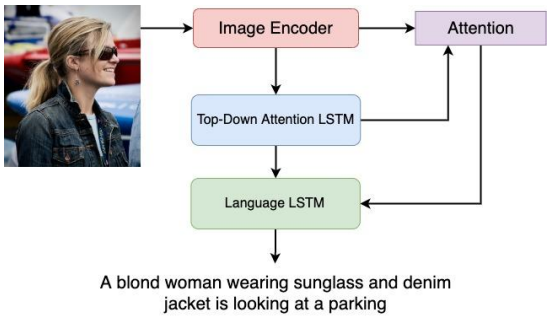
ReCAP: Role Enhanced Caption Generation

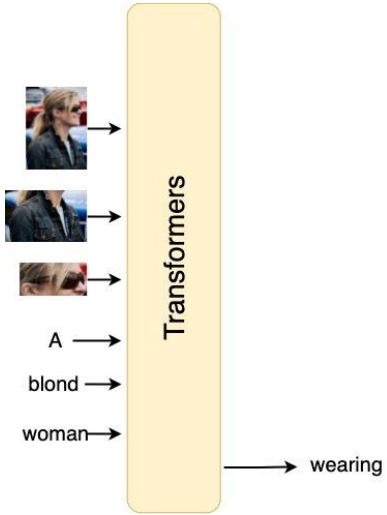
Abhidip Bhattacharyya, Martha Palmer, Christoffer Heckman

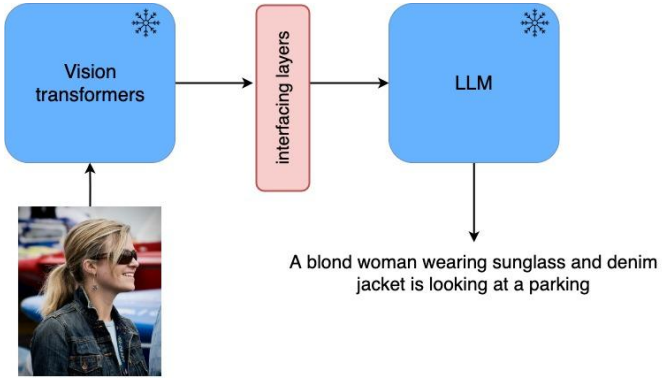
LREC-COLING 2024 - The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation

Caption Generation

Current methods

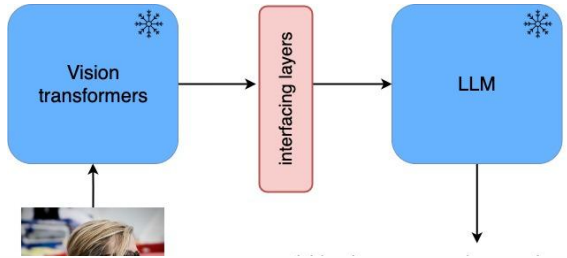






LLMs and V+L models

1. Internet scale data
2. Weak alignments and transfer learning
3. Huge models
4. Incontext and few shot learning

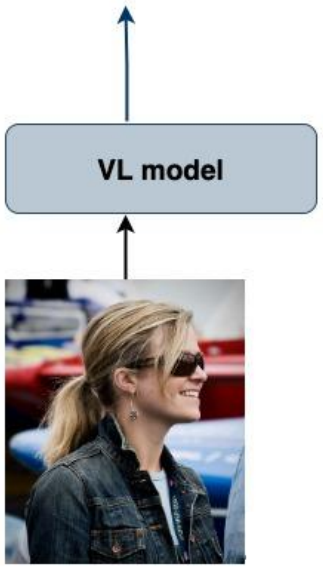


nd denim
g

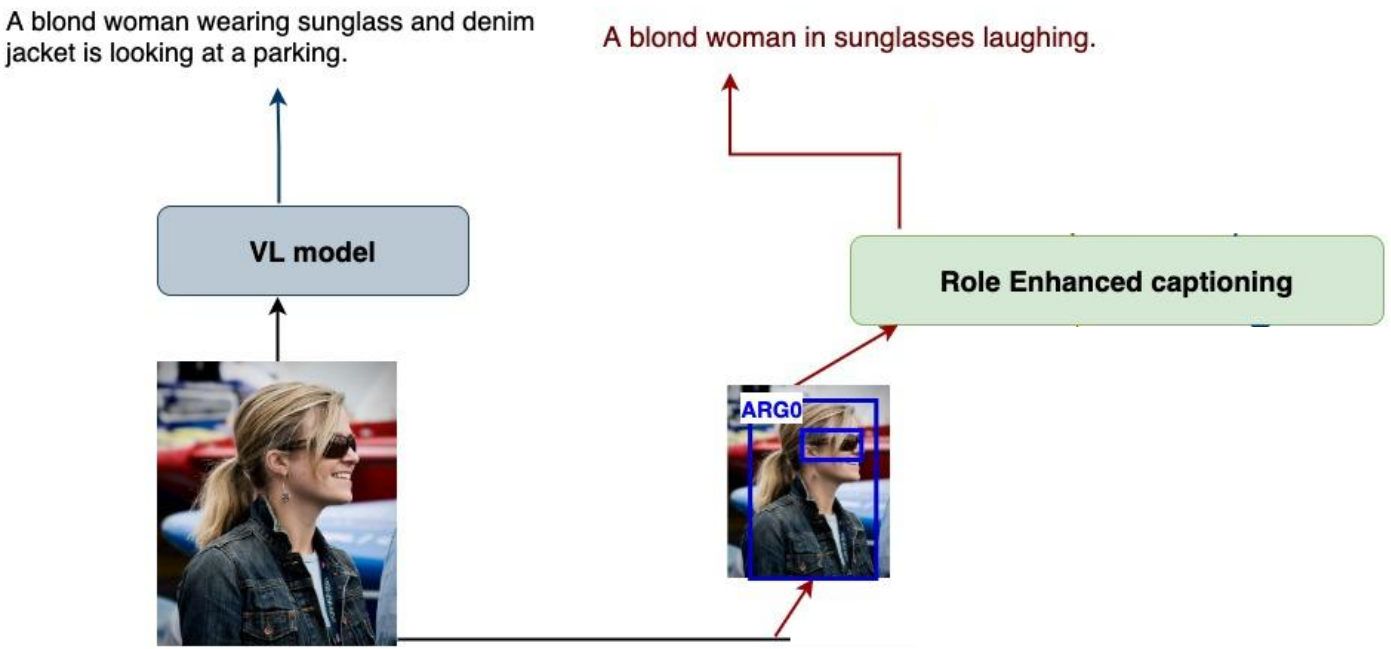
1. Black-box
2. Controllability ???
3. Interpretability ???

Role Enhanced Caption Generation

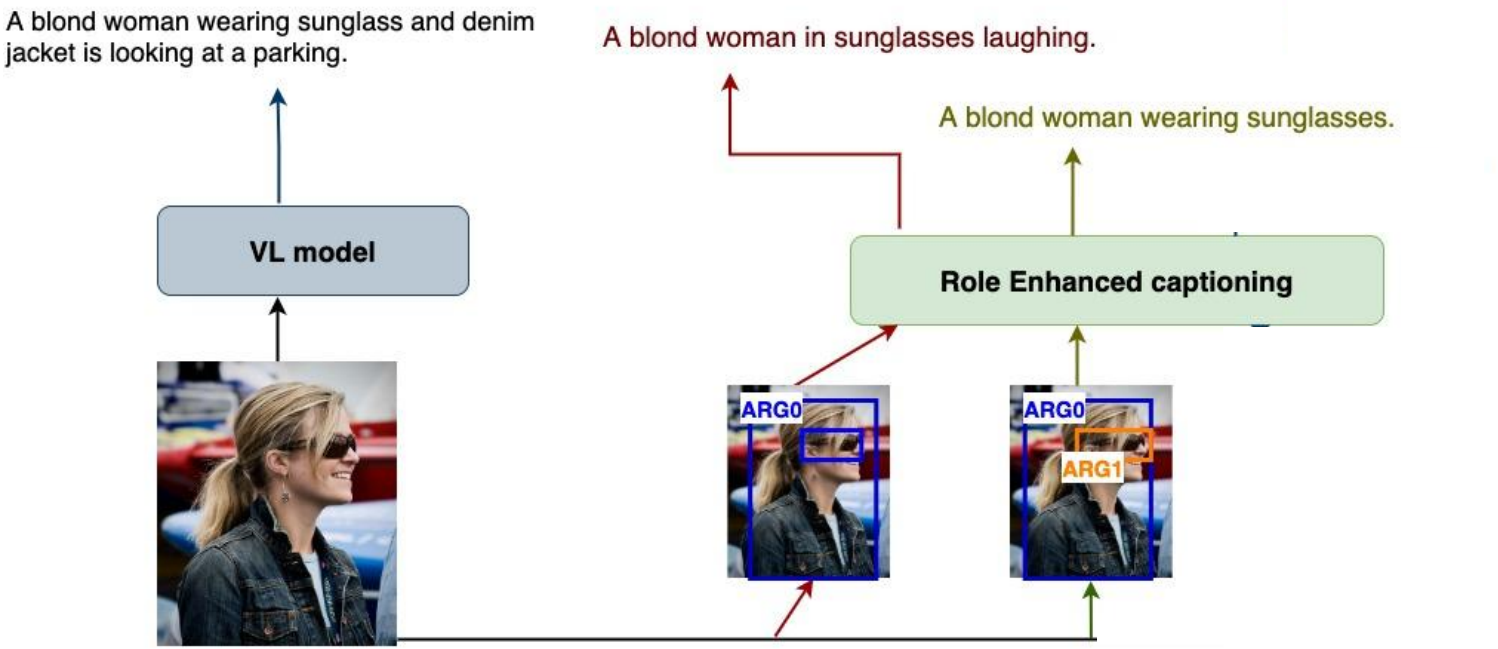
A blond woman wearing sunglass and denim jacket is looking at a parking.



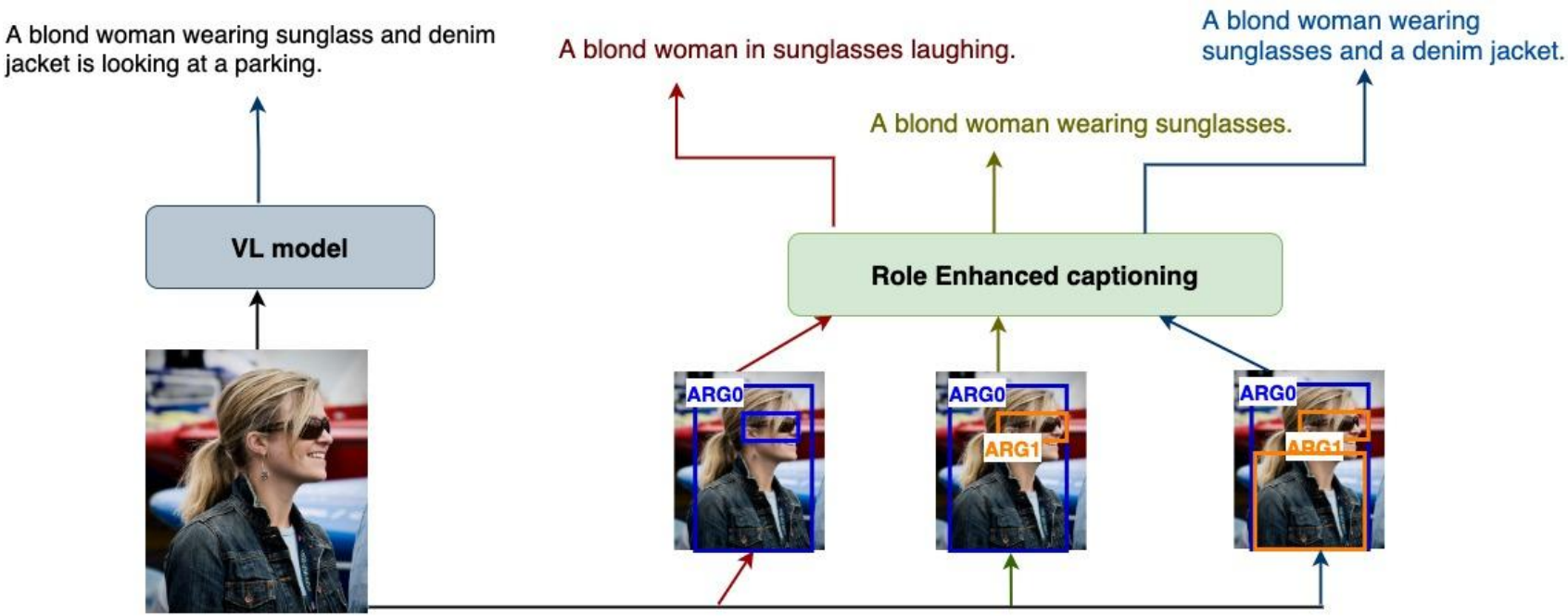
Role Enhanced Caption Generation



Role Enhanced Caption Generation



Role Enhanced Caption Generation



Semantic Role Labeling

- ‘Who’ is doing ‘what’ to ‘whom’, ‘where’, ‘when’ and ‘how’
- Predicate-arguments structures of sentence

Carl	gave	food	to	his pet
Carl	gave	his pet		food
Food	was given	to	his pet	by Carl

Semantic Role Labeling

- ‘Who’ is doing ‘what’ to ‘whom’, ‘where’, ‘when’ and ‘how’
- Predicate-arguments structures of sentence

[Carl]_who **[**gave**]_V** [food]_what to **[**his pet**]_**whom

[Carl]_who **[**gave**]_V** **[**his pet**]_whom** [food]_what

[Food]_what was **[**given**]_V** to **[**his pet**]_**whom by [Carl]_who

SRL: Propbank…

ARG0	agent	ARG3	starting point, benefactive, attribute
ARG1	patient	ARG4	ending point
ARG2	instrument, benefactive, attribute	ARGM	modifier

Table 1.1: List of arguments in PropBank

Role Enhanced Caption Generation

- We have used semantic roles as control signals to generate focused image descriptions.

Role Enhanced Caption Generation

- We have used semantic roles as control signals to generate focused image descriptions.
- SRL-informed models can generate diverse captions based on different linguistic foci.

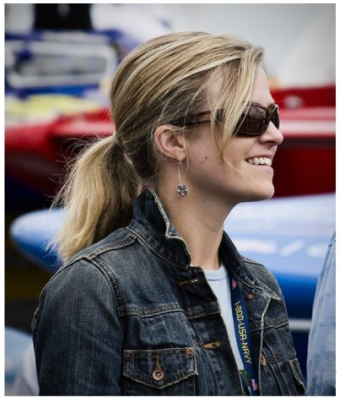
Role Enhanced Caption Generation

- We have used semantic roles as control signals to generate focused image descriptions.
- SRL-informed models can generate diverse captions based on different linguistic foci.
- The explanation of the generated caption stems from its grounding in linguistics, which can be attributed to the provided SRL annotation of the image.

Role Enhanced Caption Generation

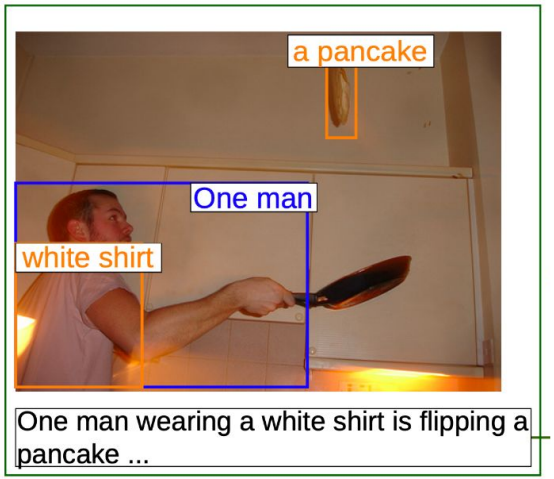
1. Generate caption based on the input image and SRL annotation of bounding boxes.
2. Can generate different Captions for the same image based on SRL
    a. Different predicates
    b. Different participants
    c. Different ARGM roles
3. SRL as cue
    a. Control signal to diversify
    b. Control amount of information

ReCAP: Architecture



Data Preparation

image caption pair with entity mapping



ReCAP and Reward function

1. Semantic roles as hint of predicate argument structure
2. Model should reward caption that follows the predicate argument structure

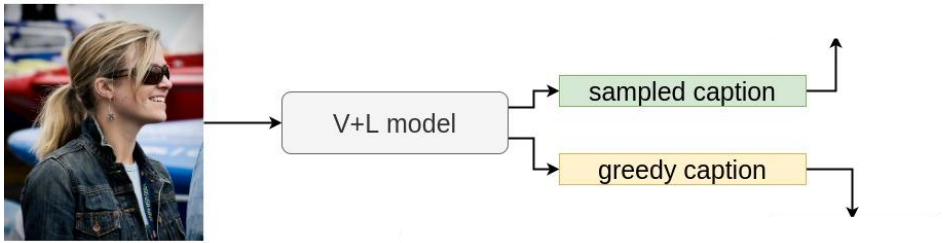
ReCAP and Reward function

1. Semantic roles as hint of predicate argument structure
2. Model should reward caption that follows the predicate argument structure
    a. **How to score abidance of predicate argument structure?**

ReCAP and Reward function

1. Semantic roles as hint of predicate argument structure
2. Model should reward caption that follows the predicate argument structure
    a. **How to score abidance of predicate argument structure?**

Typical Reward function



Smatch Reward function

Reference caption → Smatch scoring → $r(y^s_{1:T})$

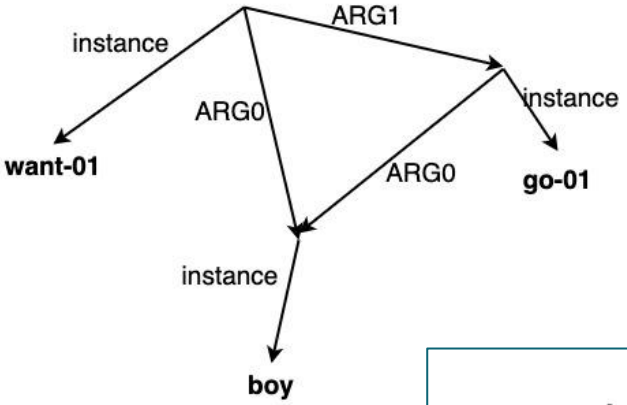
ReCAP → sampled caption → Smatch scoring → $r(y^s_{1:T})$

ReCAP → greedy caption → Smatch scoring → $r(\hat{y}_{1:T})$

Reference caption → Smatch scoring → $r(\hat{y}_{1:T})$

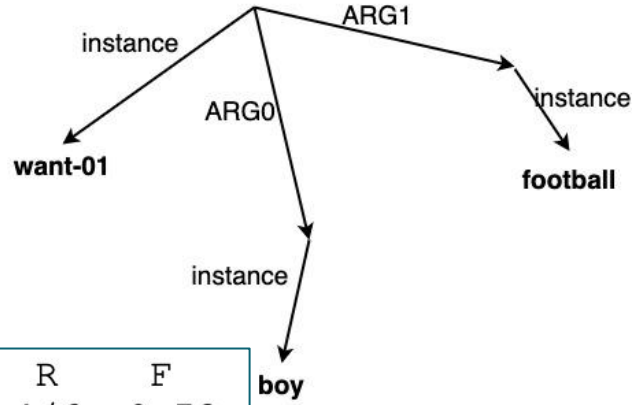
$$\nabla_\theta \mathcal{L}_r(\theta) \approx -(r(y^s_{1:T}) - r(\hat{y}_{1:T}))\nabla_\theta \log P_\theta(y_{1:T})$$

Smatch



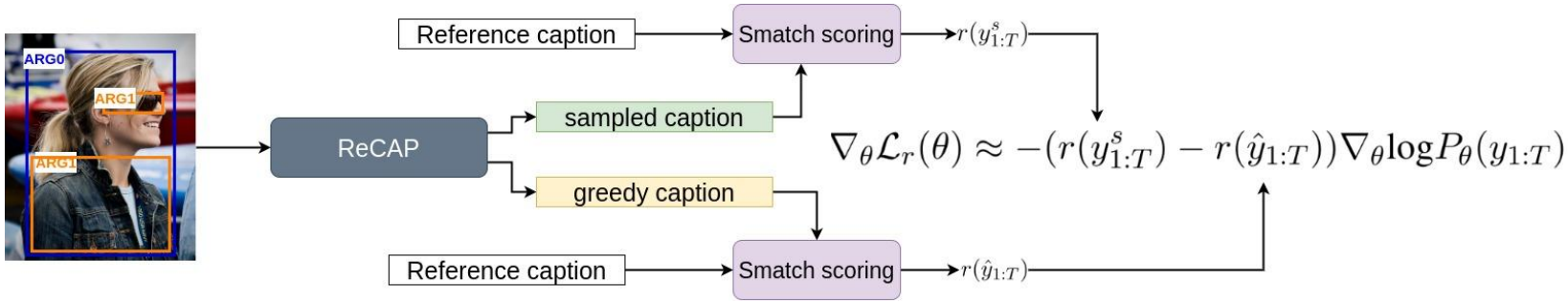
instance(a, want-01)
∧
instance(b, boy) ∧
instance(c, go-01) ∧
ARG0(a, b) ∧
ARG1(a, c) ∧
ARG0(c, b)

	M	P	R	F
x=a, y=b, z=c:	4	4/5	4/6	0.73
x=a, y=c, z=b:	1	1/5	1/6	0.18
x=b, y=a, z=c:	0	0/5	0/6	0.00
x=b, y=c, z=a:	0	0/5	0/6	0.00
x=c, y=a, z=b:	0	0/5	0/6	0.00
x=c, y=b, z=a:	2	2/5	2/6	0.36
smatch score:				0.73



instance(x, want-01) ∧
instance(y, boy) ∧
instance(z, football) ∧
ARG0(x, y) ∧
ARG1(x, z)

ReCAP and Abstract Meaning Representation



We use Smatch score as reward signal-

- Parse sentences with SPRING Parsers to generate AMR

ReCAP and Abstract Meaning Representation

1. Semantic roles as hint of predicate argument structure
2. Model should reward caption that follows the predicate argument structure
    a. **How to score abidance of predicate argument structure?**

We use Smatch score as reward signal-

- Parse sentences with SPRING Parsers to generate AMR
- Calculate Smatch score

Quantitative Results

Model	Smatch	CIDEr	BLEU-4	METEOR	ROUGE
BLIP (base_coco) (Li et al., 2022)	0.33	0.761	0.274	0.237	0.507
BLIP(large_coco) (Li et al., 2022)	0.33	0.793	0.289	0.243	0.518
BLIP2(finetuned t5 xl) (Li et al., 2023)	0.35	0.947	0.337	0.264	0.557
TDBU (Anderson et al., 2018)	0.31	0.53	0.253	0.214	0.512
+SRL	0.33	0.384	0.206	0.185	0.481
RECAP (ours)					
VL-BART (Cho et al., 2021a)	0.35	0.66	0.27	0.23	0.50
+SRL	0.4	0.95	0.12	0.17	0.35
+CLIP-S(Cho et al., 2022a)	0.4	0.945	0.118	0.167	0.346
+amr scst	0.51	0.857	0.104	0.177	0.344
+amr scst −CLIP-S	0.51	0.823	0.099	0.175	0.341

Quantitative Results

Model	Smatch	CIDEr	BLEU-4	METEOR	ROUGE
BLIP (base_coco) (Li et al., 2022)	0.33	0.761	0.274	0.237	0.507
BLIP(large_coco) (Li et al., 2022)	0.33	0.793	0.289	0.243	0.518
BLIP2(finetuned t5 xl) (Li et al., 2023)	0.35	0.947	0.337	0.264	0.557
TDBU (Anderson et al., 2018)	0.31	0.53	0.253	0.214	0.512
+SRL	0.33	0.384	0.206	0.185	0.481
RECAP (ours)					
VL-BART (Cho et al., 2021a)	0.35	0.66	0.27	0.23	0.50
+SRL	0.4	0.95	0.12	0.17	0.35
+CLIP-S(Cho et al., 2022a)	0.4	0.945	0.118	0.167	0.346
+amr scst	0.51	0.857	0.104	0.177	0.344
+amr scst −CLIP-S	0.51	0.823	0.099	0.175	0.341

Quantitative Results

Model	Smatch	CIDEr	BLEU-4	METEOR	ROUGE
BLIP (base_coco) (Li et al., 2022)	0.33	0.761	0.274	0.237	0.507
BLIP(large_coco) (Li et al., 2022)	0.33	0.793	0.289	0.243	0.518
BLIP2(finetuned t5 xl) (Li et al., 2023)	0.35	0.947	0.337	0.264	0.557
TDBU (Anderson et al., 2018)	0.31	0.53	0.253	0.214	0.512
+SRL	0.33	0.384	0.206	0.185	0.481
RECAP (ours)					
VL-BART (Cho et al., 2021a)	0.35	0.66	0.27	0.23	0.50
+SRL	0.4	0.95	0.12	0.17	0.35
+CLIP-S(Cho et al., 2022a)	0.4	0.945	0.118	0.167	0.346
+amr scst	0.51	0.857	0.104	0.177	0.344
+amr scst −CLIP-S	0.51	0.823	0.099	0.175	0.341

Quantitative Results

Model	Smatch	CIDEr	BLEU-4	METEOR	ROUGE
BLIP (base_coco) (Li et al., 2022)	0.33	0.761	0.274	0.237	0.507
BLIP(large_coco) (Li et al., 2022)	0.33	0.793	0.289	0.243	0.518
BLIP2(finetuned t5 xl) (Li et al., 2023)	0.35	0.947	0.337	0.264	0.557
TDBU (Anderson et al., 2018)	0.31	0.53	0.253	0.214	0.512
+SRL	0.33	0.384	0.206	0.185	0.481
RECAP (ours)					
VL-BART (Cho et al., 2021a)	0.35	0.66	0.27	0.23	0.50
+SRL	0.4	0.95	0.12	0.17	0.35
+CLIP-S(Cho et al., 2022a)	0.4	0.945	0.118	0.167	0.346
+amr scst	0.51	0.857	0.104	0.177	0.344
+amr scst −CLIP-S	0.51	0.823	0.099	0.175	0.341

Quantitative Results

Model	Smatch	CIDEr	BLEU-4	METEOR	ROUGE
BLIP (base_coco) (Li et al., 2022)	0.33	0.761	0.274	0.237	0.507
BLIP(large_coco) (Li et al., 2022)	0.33	0.793	0.289	0.243	0.518
BLIP2(finetuned t5 xl) (Li et al., 2023)	0.35	0.947	0.337	0.264	0.557
TDBU (Anderson et al., 2018)	0.31	0.53	0.253	0.214	0.512
+SRL	0.33	0.384	0.206	0.185	0.481
RECAP (ours)					
VL-BART (Cho et al., 2021a)	0.35	0.66	0.27	0.23	0.50
+SRL	0.4	0.95	0.12	0.17	0.35
+CLIP-S(Cho et al., 2022a)	0.4	0.945	0.118	0.167	0.346
+amr scst	0.51	0.857	0.104	0.177	0.344
+amr scst −CLIP-S	0.51	0.823	0.099	0.175	0.341

Quantitative Results

Model	Smatch	CIDEr	BLEU-4	METEOR	ROUGE
BLIP (base_coco) (Li et al., 2022)	0.33	0.761	0.274	0.237	0.507
BLIP(large_coco) (Li et al., 2022)	0.33	0.793	0.289	0.243	0.518
BLIP2(finetuned t5 xl) (Li et al., 2023)	0.35	0.947	0.337	0.264	0.557
TDBU (Anderson et al., 2018)	0.31	0.53	0.253	0.214	0.512
+SRL	0.33	0.384	0.206	0.185	0.481
RECAP (ours)					
VL-BART (Cho et al., 2021a)	0.35	0.66	0.27	0.23	0.50
+SRL	0.4	0.95	0.12	0.17	0.35
+CLIP-S(Cho et al., 2022a)	0.4	0.945	0.118	0.167	0.346
+amr scst	0.51	0.857	0.104	0.177	0.344
+amr scst −CLIP-S	0.51	0.823	0.099	0.175	0.341

Quantitative Results

Model	Smatch	CIDEr	BLEU-4	METEOR	ROUGE
BLIP (base_coco) (Li et al., 2022)	0.33	0.761	0.274	0.237	0.507
BLIP(large_coco) (Li et al., 2022)	0.33	0.793	0.289	0.243	0.518
BLIP2(finetuned t5 xl) (Li et al., 2023)	0.35	0.947	0.337	0.264	0.557
TDBU (Anderson et al., 2018)	0.31	0.53	0.253	0.214	0.512
+SRL	0.33	0.384	0.206	0.185	0.481
RECAP (ours)					
VL-BART (Cho et al., 2021a)	0.35	0.66	0.27	0.23	0.50
+SRL	0.4	0.95	0.12	0.17	0.35
+CLIP-S(Cho et al., 2022a)	0.4	0.945	0.118	0.167	0.346
+amr scst	0.51	0.857	0.104	0.177	0.344
+amr scst −CLIP-S	0.51	0.823	0.099	0.175	0.341

Quantitative Results

Model	Smatch	CIDEr	BLEU-4	METEOR	ROUGE
BLIP (base_coco) (Li et al., 2022)	0.33	0.761	0.274	0.237	0.507
BLIP(large_coco) (Li et al., 2022)	0.33	0.793	0.289	0.243	0.518
BLIP2(finetuned t5 xl) (Li et al., 2023)	0.35	0.947	0.337	0.264	0.557
TDBU (Anderson et al., 2018)	0.31	0.53	0.253	0.214	0.512
+SRL	0.33	0.384	0.206	0.185	0.481
RECAP (ours)					
VL-BART (Cho et al., 2021a)	0.35	0.66	0.27	0.23	0.50
+SRL	0.4	0.95	0.12	0.17	0.35
+CLIP-S(Cho et al., 2022a)	0.4	0.945	0.118	0.167	0.346
+amr scst	0.51	0.857	0.104	0.177	0.344
+amr scst −CLIP-S	0.51	0.823	0.099	0.175	0.341

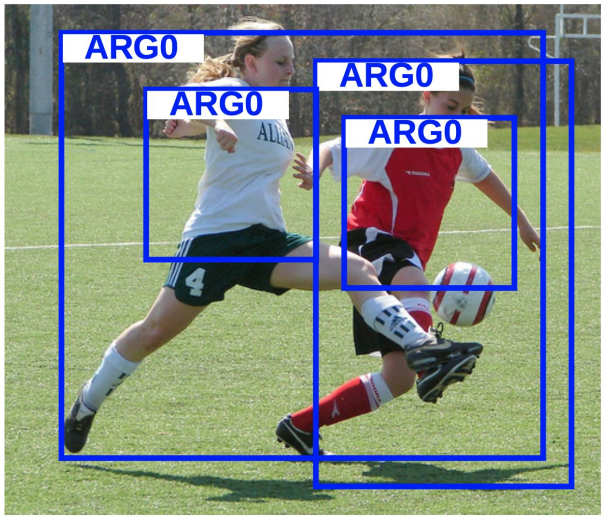
Qualitative Analysis:Arguments accommodation

Two girls are playing soccer

Two girls playing soccer are going for the ball

Qualitative Analysis:Adjoining

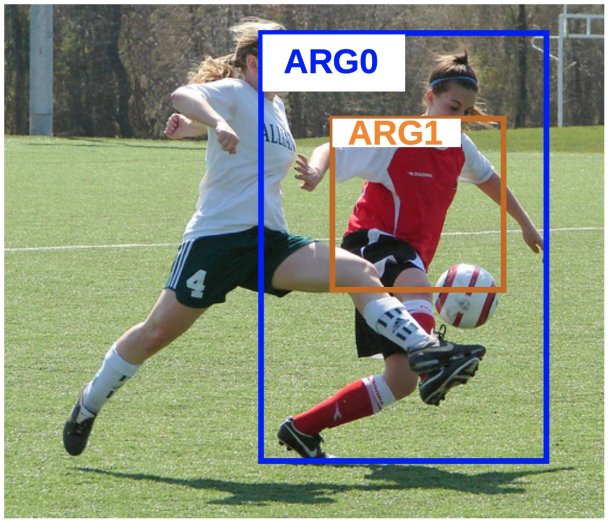
Two girls are playing soccer



Two girls wearing uniforms are playing soccer

Qualitative Analysis:Diversified based on Predicate-Argument

Two girls are playing soccer



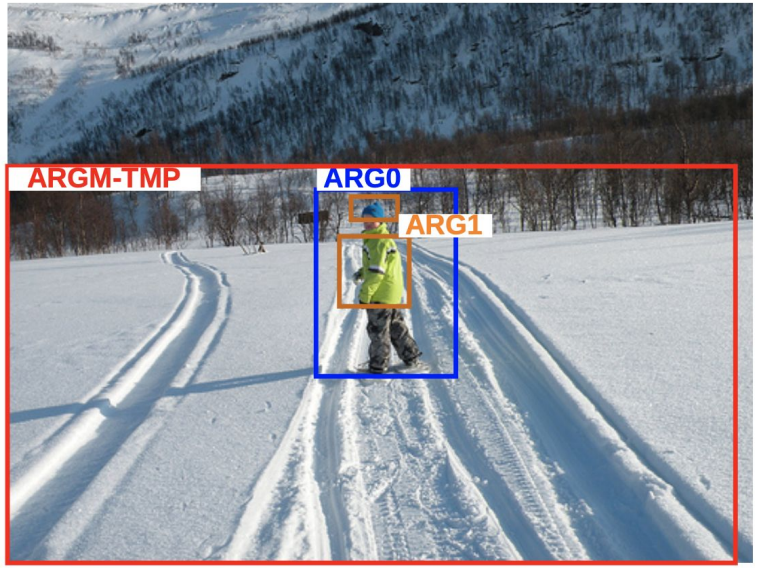
A young girl is wearing a red and white soccer uniform

Qualitative Analysis:Activity vs State

A group of people is walking down a sidewalk lined with trees.

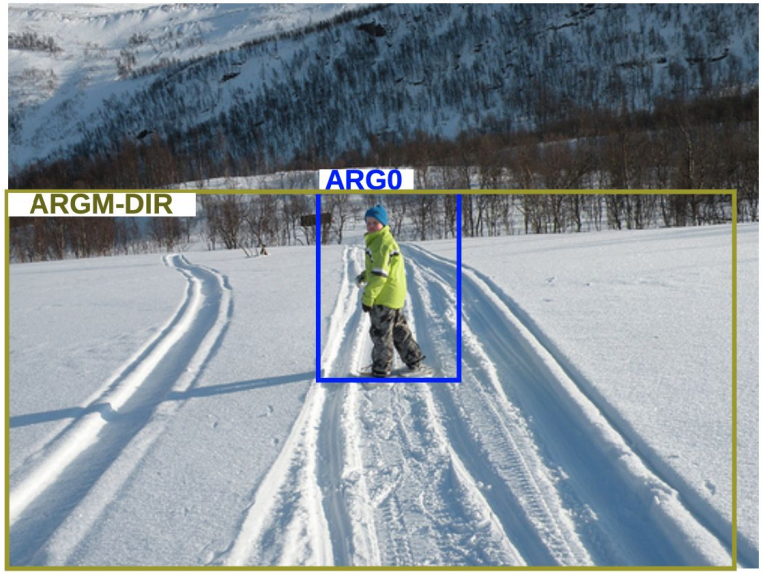
A group of people is standing in a park.

ARGM



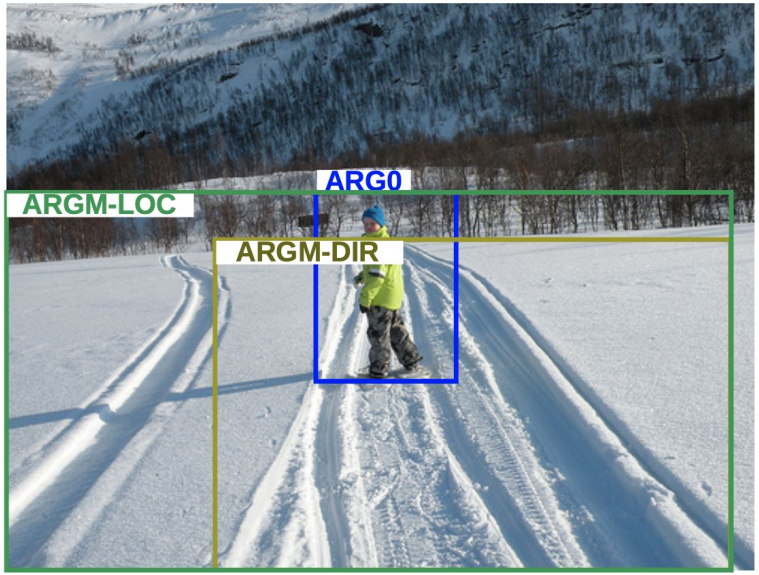
- A person wearing a yellow jacket and blue hat **as he walks down a snow covered road.**

ARGM



- A person wearing a yellow jacket and blue hat **as he walks down a snow covered road.**
- A person is walking **down a snow covered road.**

ARGM



- A person wearing a yellow jacket and blue hat as he walks down a snow covered road.
- A person is walking **down a snow covered road.**
- A person is walking along a trail **in the snow.**

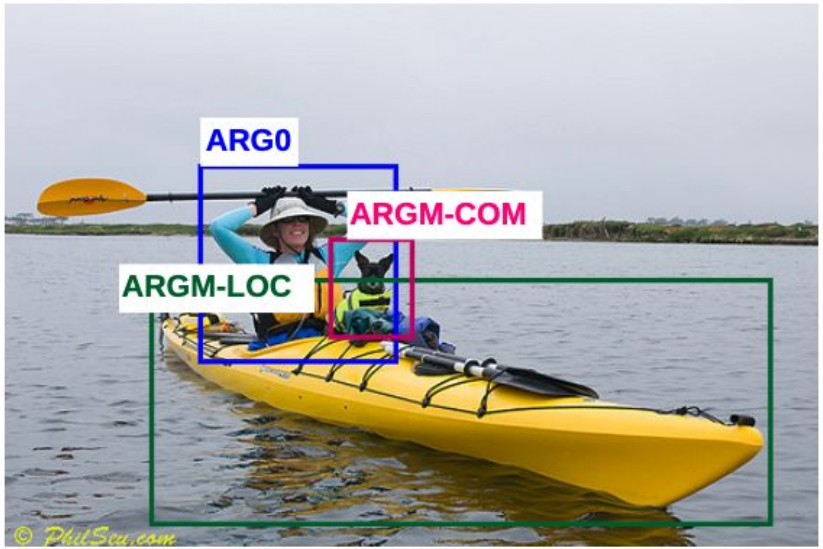
ReCAP: Analysis

1. Generated caption is focused w.r.t given SRL
2. Reflect predicate-argument structures given in the input SRL
3. Can be explained with different linguistic aspects
    a. Adjoining of information
    b. Accommodation arguments
    c. Choose different predicate based on arguments
    d. Activity vs State
    e. ARGM roles

Quantitative Results

Model	Smatch	CIDEr	BLEU-4	METEOR	ROUGE
TDBU (Anderson et al., 2018)	0.31	0.53	0.253	0.214	0.512
+SRL	0.33	0.384	0.206	0.185	0.481
RECAP (ours)					
VL-BART (Cho et al., 2021)	0.35	0.66	0.27	0.23	0.50
+SRL	0.4	0.95	0.12	0.17	0.35

Quantitative Results

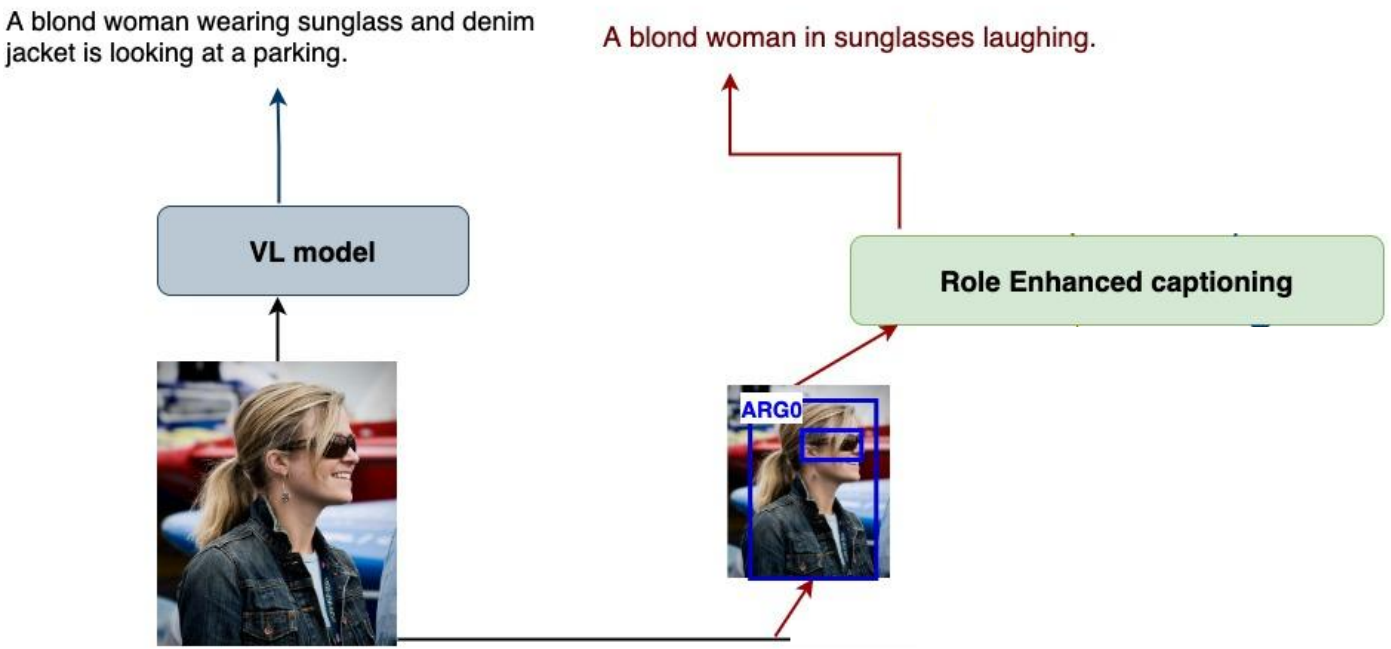


GT sentence: A woman kayaking in a yellow kayak with her dog.

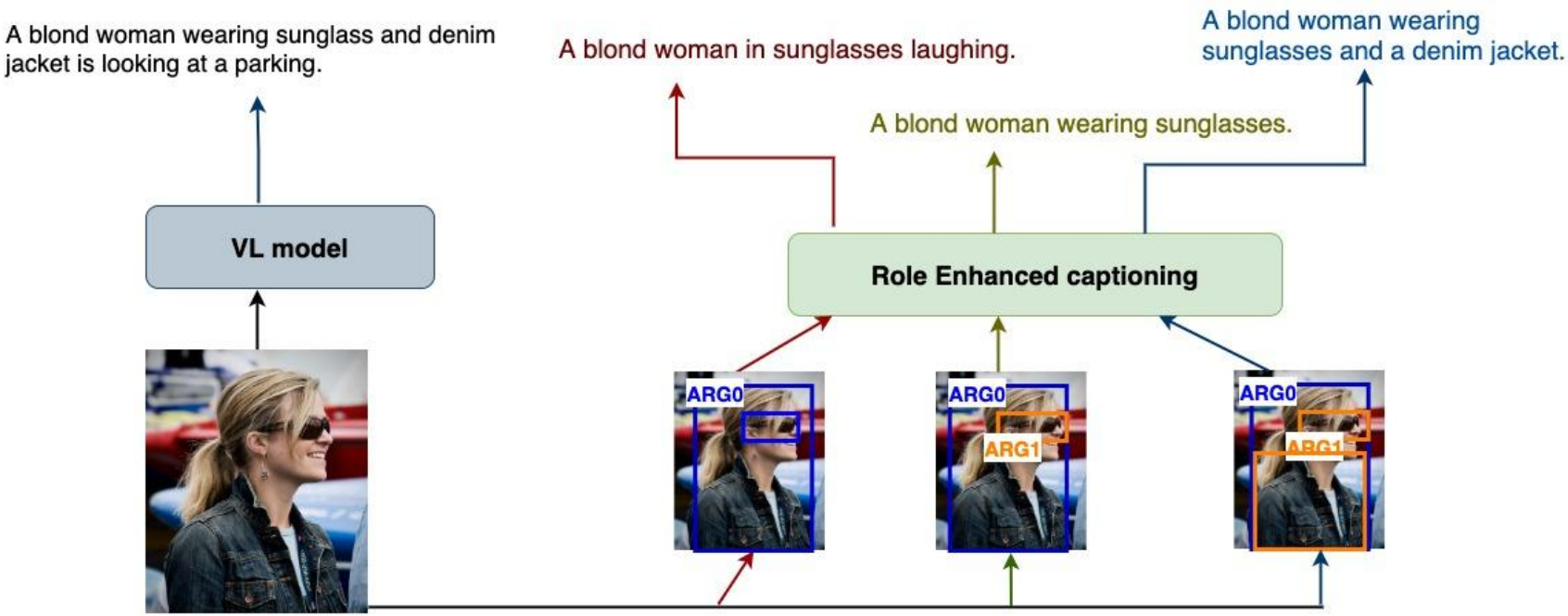
Without SCST: A person kayaking on a lake with a dog.

With AMR SCST:A person in a blue wetsuit kayaking on a lake with a black and yellow dog.

Summary


A blond woman wearing sunglass and denim jacket is looking at a parking.
A blond woman in sunglasses laughing.
VL model
Role Enhanced captioning
ARG0

Summary


A blond woman wearing sunglass and denim jacket is looking at a parking.
VL model
A blond woman in sunglasses laughing.
A blond woman wearing sunglasses.
A blond woman wearing sunglasses and a denim jacket.
Role Enhanced captioning
ARG0
ARG0
ARG1
ARG0
ARG1

Summary

1. Generated caption is focused w.r.t given SRL
2. Reflect predicate-argument structures given in the input SRL
3. Can be explained with different linguistic aspects
    a. Adjoining of information
    b. Accommodation arguments
    c. Choose different predicate based on arguments
    d. Activity vs State
    e. ARGM roles

Summary

Can be explained with different linguistic aspects

a. Adjoining of information
b. Accommodation arguments
c. Choose different predicate based on arguments
d. Activity vs State verbs

Thank you

Questions?