When an Image Tells a Story: The Role of Visual and Semantic Information for Generating Paragraph Descriptions [and much more...]

Nikolai Ilinykh

Centre for Linguistic Theory and Studies in Probability (CLASP)
Department of Philosophy, Linguistics and Theory of Science (FLoV)
University of Gothenburg, Sweden

10^{th} March 2021

Describing images with longer sequences

Describing images with longer sequences[1]

People are standing on the grass behind a concrete patch that looks like it was just set. There are two orange cones in front of the concrete and yellow tape surrounding it. There are three people in yellow vests and white hard hats. There are some people sitting on a bench next to them.

Properties of image paragraphs

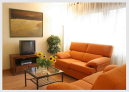
- **informativeness**: descriptions consisting of multiple sentences
- **grounding**: every word can be grounded in something in the image
  - != visual storytelling
- **discourse**: there is some type of an order to the sentences in the paragraph

Why image paragraphs?

It has all started with visual dialogue...[2]

- *Visual Dialogue*[3] is a type of setting in which an artificial agent is required to hold a meaningful dialogue with humans in natural language about visual content
- MeetUp! is a conversational game aimed at modelling natural human-human interaction in a situated setting (e.g. environment is shared between speakers)
- Important properties of MeetUp!: collaborative nature of the task, symmetry between speakers, visual grounding of words in the environment, conversational grounding between speakers (e.g. the meaning is negotiated and established rather than fixed and stated)
- **Very important**: dialogue discourse is actively used to refer to previously mentioned elements (not present in many other visual dialogue settings!)

Visual Dialogue Examples

	Time	Private to A	Public	Private to B
1	(00:00)	bedroom[72]		aprt_building/outdoor[1]
2	(00:00)	You can go [/w]est		
3	(00:01)			You can go [/s]outh
4	(00:05)		A: Hello	
5	(00:08)		B: Hello	
6	(00:11)		A: /i	
7	(00:14)		B: /i	
8	(00:19)		B: /I	
9	(00:23)			B: /l
10	(00:23)			You can go [/s]outh
11	(00:29)		A: I am in a bedroom with a black bed, my exits are west.	
12	(00:34)		A: I am heading west.	
13	(00:37)	$\xrightarrow{W}$ living_room[49]		
14	(00:37)	You can go [/n]orth [/e]ast [/s]outh [/w]est		

42	(04:11)		A: So, this kitchen, did it have red brick walls?	
43	(04:35)		B: Yes, does yours have white cabinets and a wood-topped center table thing?	
44	(04:48)			$\xleftarrow{W}$ kitchen[39]
45	(04:48)			You can go [/n]orth [/e]ast [/s]outh [/w]est
46	(04:52)		A: Yes. There are red cabinets attached to the wood table?	
47	(05:07)		B: Looks like it. Some sort of steel appliance?	
48	(05:09)		A: Above the oven, is there a small blue-framed picture?	
49	(05:16)		B: Yes.	
50	(05:18)		A: Not oven, my mistake.	
51	(05:25)		A: I think we're in the same space.	
52	(05:35)		B: I agree. Done?	
53	(05:38)		A: Yes.	
54	(05:40)	/done		
55	(05:44)			/done

What have we learned?

a.	Game Master: You have to meet in a room of type *utility room*.
b.	A: Hi. I'm in a bedroom with pink walls.
c.	B: I seem to be in a kitchen.
d.	A: I'll go look for a utility room.
e.	A (privately): *north*
f.	A (privately): *west*
g.	B (privately): *east*
h.	A: Found a room with a washing machine. Is that a utility room?
i.	B: Was wondering as well. Probably that's what it is.
j.	B: I'm in the pink bedroom now. I'll come to you.
k.	B (privately): *north*
l.	B (privately): *west*
m.	B: Poster above washing machine?
n.	A: Mine has a mirror on the wall.
o.	B: yeah, could be mirror. Plastic chair?
p.	A: And laundry basket.
q.	A: *done*
r.	B: Same
s.	B: *done*

What have we learned?

a. Game Master: You have to meet in a room of type *utility room*.

- setting up the classification task

b. A: Hi. I'm in a bedroom with pink walls.
c. B: I seem to be in a kitchen.
d. A: I'll go look for a utility room.
e. A (privately): *north*
f. A (privately): *west*
g. B (privately): *east*
h. A: Found a room with a washing machine. Is that a utility room?
i. B: Was wondering as well. Probably that's what it is.
j. B: I'm in the pink bedroom now. I'll come to you.
k. B (privately): *north*
l. B (privately): *west*
m. B: Poster above washing machine?
n. A: Mine has a mirror on the wall.
o. B: yeah, could be mirror. Plastic chair?
p. A: And laundry basket.
q. A: *done*
r. B: Same
s. B: *done*

What have we learned?

a.	Game Master: You have to meet in a room of type *utility room*.	- setting up the classification task
b.	A: Hi. I'm in a bedroom with pink walls.	- synchronize mutual state representations
c.	B: I seem to be in a kitchen.	
d.	A: I'll go look for a utility room.	
e.	A (privately): *north*	
f.	A (privately): *west*	
g.	B (privately): *east*	
h.	A: Found a room with a washing machine. Is that a utility room?	
i.	B: Was wondering as well. Probably that's what it is.	
j.	B: I'm in the pink bedroom now. I'll come to you.	
k.	B (privately): *north*	
l.	B (privately): *west*	
m.	B: Poster above washing machine?	
n.	A: Mine has a mirror on the wall.	
o.	B: yeah, could be mirror. Plastic chair?	
p.	A: And laundry basket.	
q.	A: *done*	
r.	B: Same	
s.	B: *done*	

What have we learned?

a.	Game Master: You have to meet in a room of type *utility room*.	- setting up the classification task
b.	A: Hi. I'm in a bedroom with pink walls.	- synchronize mutual state representations
c.	B: I seem to be in a kitchen.	
d.	A: I'll go look for a utility room.	
e.	A (privately): *north*	
f.	A (privately): *west*	
g.	B (privately): *east*	
h.	A: Found a room with a washing machine. Is that a utility room?	- coordination of strategy
i.	B: Was wondering as well. Probably that's what it is.	
j.	B: I'm in the pink bedroom now. I'll come to you.	
k.	B (privately): *north*	
l.	B (privately): *west*	
m.	B: Poster above washing machine?	
n.	A: Mine has a mirror on the wall.	
o.	B: yeah, could be mirror. Plastic chair?	
p.	A: And laundry basket.	
q.	A: *done*	
r.	B: Same	
s.	B: *done*	

What have we learned?

a. Game Master: You have to meet in a room of type *utility room*.
b. A: Hi. I'm in a bedroom with pink walls.
c. B: I seem to be in a kitchen.
d. A: I'll go look for a utility room.
e. A (privately): *north*
f. A (privately): *west*
g. B (privately): *east*
h. A: Found a room with a washing machine. Is that a utility room?
i. B: Was wondering as well. Probably that's what it is.
j. B: I'm in the pink bedroom now. I'll come to you.
k. B (privately): *north*
l. B (privately): *west*
m. B: Poster above washing machine?
n. A: Mine has a mirror on the wall.
o. B: yeah, could be mirror. Plastic chair?
p. A: And laundry basket.
q. A: *done*
r. B: Same
s. B: *done*

- setting up the classification task
- synchronize mutual state representations
- private actions (epistemic vs. pragmatic)
- coordination of strategy

What have we learned?

a. Game Master: You have to meet in a room of type *utility room*.

- setting up the classification task

b. A: Hi. I'm in a bedroom with pink walls.
c. B: I seem to be in a kitchen.

- synchronize mutual state representations

d. A: I'll go look for a utility room.
e. A (privately): *north*
f. A (privately): *west*
g. B (privately): *east*

- private actions (epistemic vs. pragmatic)

h. A: Found a room with a washing machine. Is that a utility room?
i. B: Was wondering as well. Probably that's what it is.
j. B: I'm in the pink bedroom now. I'll come to you.

- coordination of strategy

k. B (privately): *north*
l. B (privately): *west*
m. B: Poster above washing machine?
n. A: Mine has a mirror on the wall.
o. B: yeah, could be mirror. Plastic chair?

- meta-semantic interaction

p. A: And laundry basket.
q. A: *done*
r. B: Same
s. B: *done*

What have we learned?

a.	Game Master: You have to meet in a room of type *utility room*.	- setting up the classification task
b.	A: Hi. I'm in a bedroom with pink walls.	- synchronize mutual state representations
c.	B: I seem to be in a kitchen.	
d.	A: I'll go look for a utility room.	
e.	A (privately): *north*	- private actions (epistemic vs. pragmatic)
f.	A (privately): *west*	
g.	B (privately): *east*	
h.	A: Found a room with a washing machine. Is that a utility room?	- coordination of strategy
i.	B: Was wondering as well. Probably that's what it is.	
j.	B: I'm in the pink bedroom now. I'll come to you.	- discourse memory
k.	B (privately): *north*	
l.	B (privately): *west*	
m.	B: Poster above washing machine?	
n.	A: Mine has a mirror on the wall.	- meta-semantic interaction
o.	B: yeah, could be mirror. Plastic chair?	- perfoming dialogue acts indirectly
p.	A: And laundry basket.	
q.	A: *done*	
r.	B: Same	
s.	B: *done*	

Let's simplify the task!

Moving on with image description sequences...[4]

- **image description sequences (IDS)** are longer natural language texts (paragraphs) with single images they are meant to describe

Let's simplify the task!

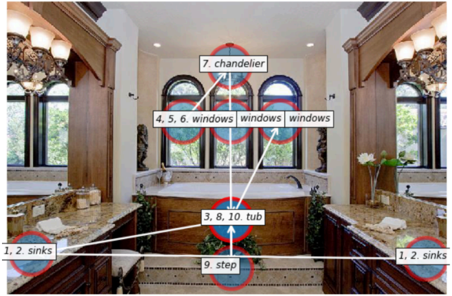
Moving on with image description sequences...[5]

- **image description sequences (IDS)** are longer natural language texts (paragraphs) with single images they are meant to describe
- this setting is a challenging tested for state-of-the-art models in NLG, where language and vision tasks need to be connected to core aspects of text generation, e.g. content selection, text structuring, or aggregation.

Let's simplify the task!

Moving on with image description sequences...[6]

- **image description sequences (IDS)** are longer natural language texts (paragraphs) with single images they are meant to describe
- this setting is a challenging tested for state-of-the-art models in NLG, where language and vision tasks need to be connected to core aspects of text generation, e.g. content selection, text structuring, or aggregation.
- IDS are aimed at partially resembling dialogical interaction
  - interface-wise: separate text input fields rather than one block
  - instruction-wise: talk to the imaginary partner who keeps asking to tell him more



1: It is a very fancy bathroom.

2: There are twin *sinks*[1, 2] across from each other.

3: There is a deep soaking *tub*[3] in front of 3 domed *windows*[4, 5, 6].

4: There is a very fancy *chandelier*[7] over the *bathtub*[8] and everything is done in brown woods and granite.

5: There is a *step*[9] up to the *bathtub*[10].

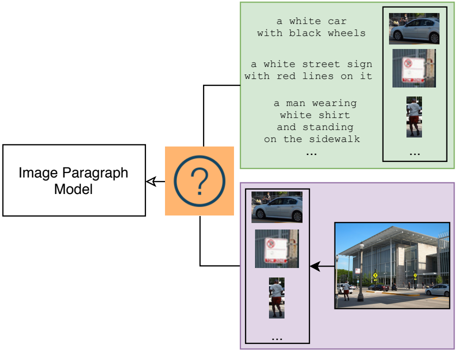
Two Sources of Important Information for IP

1. visual features of perceived objects (*what* to refer to)
2. background knowledge and communicative intent (*when* and *how* to refer)

People are standing on the **grass** behind **a concrete patch** that looks like it was just set. There are **two orange cones** in front of **the concrete and yellow tape** surrounding it. There are **three people in yellow vests and white hard hats**. There are **some people sitting on a bench** next to them.

Our paper

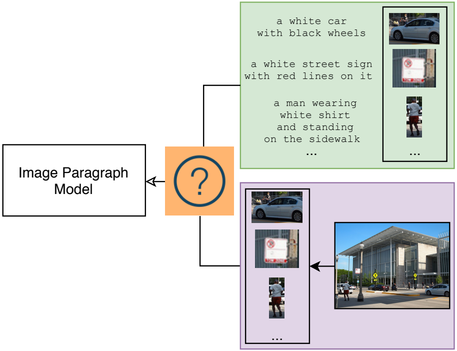
How to improve both *accuracy* and *diversity* of generated image paragraphs?



- **model input**:
  `unimodal` (visual / textual)
  vs. `multimodal`

Our paper

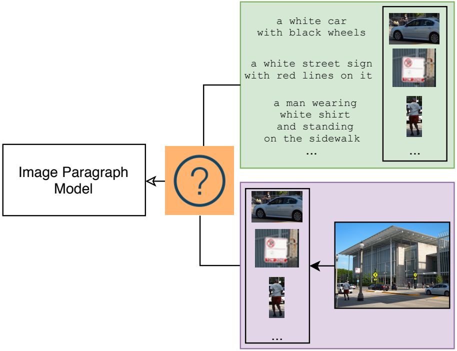
How to improve both *accuracy* and *diversity* of generated image paragraphs?

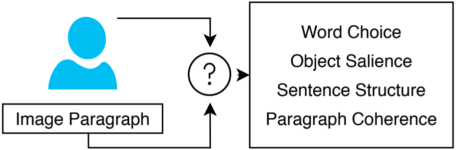


- **model input**:
  `unimodal` (visual / textual)
  vs. `multimodal`
- **information fusion**:
  `max-pooling` vs. `attention`

Our paper

How to improve both *accuracy* and *diversity* of generated image paragraphs?

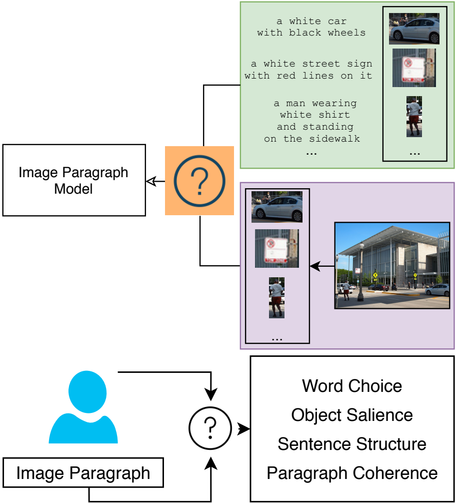




- **model input**:
  `unimodal` (visual / language)
  vs. `multimodal`
- **information fusion**:
  `max-pooling` vs. `attention`
- **paragraph evaluation**:
  `automatic` vs. `human`

Our paper

How to improve both *accuracy* and *diversity* of generated image paragraphs?

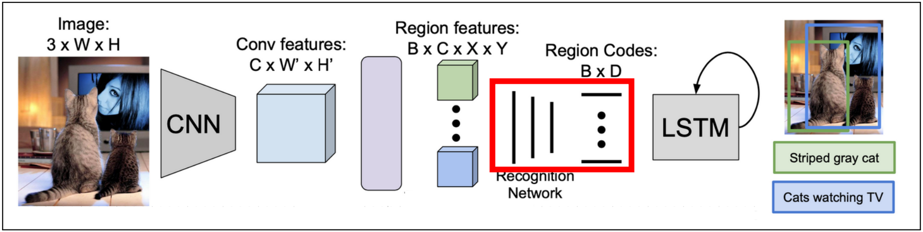


- **model input**:
  unimodal (visual / language) vs. multimodal
- **information fusion**:
  max-pooling vs. attention
- **paragraph evaluation**:
  automatic vs. human
- **human evaluation**:
  accuracy and diversity of generated paragraphs

Unimodal Features: Vision, Language

We use pre-trained **DenseCap**[7] model to extract both visual (V) and language (L) features for each image:

1. $V \in \mathbb{R}^{M \times D}$: the output of the recognition network (two fully connected layers, within the red box)

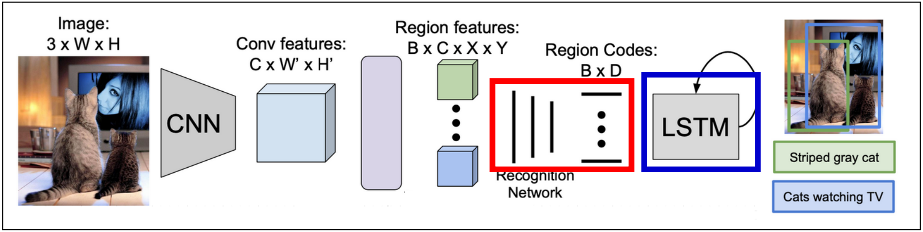


Notations: $M = 50, D = 4096, H = 512.$

Unimodal Features: Vision, Language

We use pre-trained **DenseCap**[7] model to extract both visual (V) and language (L) features for each image:

1. $V \in \mathbb{R}^{M \times D}$: the output of the recognition network (two fully connected layers, within the red box)
2. $L \in \mathbb{R}^{M \times H}$: the sequence of *hidden states* used to generate the region descriptions (within the blue box)



Notations: $M = 50, D = 4096, H = 512.$

Multimodal Features: Vision **and** Language

Mapping Visual Features

Mapping Sentence LSTM
last hidden state

$$mult_t = [\boxed{W_m^V V_t} \oplus \boxed{W_m^L L_t} \oplus \boxed{W_h h_{t-1}^{\delta}}]$$

Mapping Language Features

Multimodal Features: Vision **and** Language

Mapping Visual Features

Mapping Sentence LSTM last hidden state

$$mult_t = [\boxed{W_m^V V_t} \oplus \boxed{W_m^L L_t} \oplus \boxed{W_h h_{t-1}^{\delta}}]$$

Mapping Language Features

Note: passing multimodal features through a linear layer $FC(mult_t)$ did not affect the automatic metric scores.

Information Fusion: Max-Pooling

For uni-modal experiments, we use max-pooling on either mapped visual features $x = W_m^V V_t$ or mapped language features $x = W_m^L L_t$:

$$x_s^\varsigma = max_{i=1}^M (x) \tag{1}$$

Information Fusion: Max-Pooling

For uni-modal experiments, we use max-pooling on either mapped visual features $x = W_m^V V_t$ or mapped language features $x = W_m^L L_t$:

$$x_s^{\varsigma} = max_{i=1}^{M}(x) \tag{1}$$

For multimodal experiments, we concatenate max-pooled vectors of both modalities:

$$x_s^{\varsigma} = [max_{i=1}^{M}(W_m^L L_t) \oplus max_{i=1}^{M}(W_m^V V_t)] \tag{2}$$

Information Fusion: Late Attention

We apply **additive\concat** attention on either unimodal or multimodal features (F_t):

$$\alpha_t^{mult} = softmax(W_a^A tanh(F_t \oplus W_h h_{t-1}^{\delta}) \tag{3}$$

$$f_t = [\alpha_t^{mult} \odot F_t] \tag{4}$$

Information Fusion: Late Attention

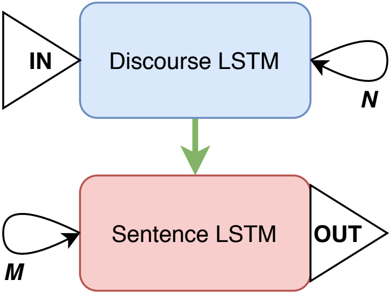
We apply **additive\concat** attention on either unimodal or multimodal features (F_t):

$$\alpha_t^{mult} = softmax(W_a^A tanh(F_t \oplus W_h h_{t-1}^{\delta})) \tag{5}$$

$$f_t = [\alpha_t^{mult} \odot F_t] \tag{6}$$

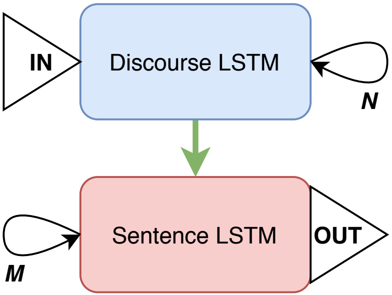
Note: Although some work on multimodal machine translation has shown that early attention improves quality of text generations [8,9] , using **modality-dependent / early** attention (unique W_a^A and, therefore, unique α_t^{mult} for each modality) provided us with worse automatic metric scores.

Image Paragraph Model



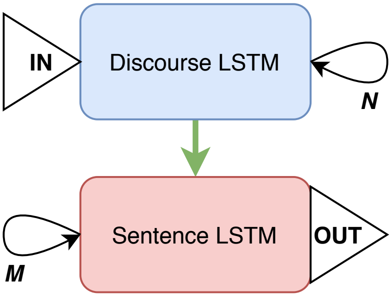
- **IN**: visual / language / multimodal features

Image Paragraph Model



- **IN**: visual / language / multimodal features
- **Discourse LSTM** produces topics for each sentence $n_t \in N$
- **Sentence LSTM** uses each topic to generate the corresponding sentence

Image Paragraph Model



- **IN**: visual / language / multimodal features
- **Discourse LSTM** produces topics for each sentence $n_t \in N$
- **Sentence LSTM** uses each topic to generate the corresponding sentence
- The model is trained on pairs of images and paragraphs from the Stanford Image Paragraph Dataset

Results: automatic metrics, accuracy

Model Input	Type	WMD	CIDEr	METEOR	BLEU-1	BLEU-2	BLEU-3	BLEU-4
IMG	+MAX	7.48	25.66	11.20	24.51	13.67	7.96	4.51
LNG	+MAX	7.19	22.27	10.81	23.20	12.69	7.34	4.19
IMG+LNG	+MAX	**7.61**	**26.38**	**11.30**	**25.10**	**13.88**	**8.11**	**4.61**
IMG	+ATT	7.47	26.01	11.26	24.88	**13.99**	**8.13**	**4.67**
LNG	+ATT	7.20	22.11	10.82	23.20	12.55	7.16	3.97
IMG+LNG	+ATT	**7.54**	**26.04**	**11.28**	**24.96**	13.82	8.04	4.60

1. using multimodal features seems to improve the quality of generated paragraphs

Results: automatic metrics, accuracy

Model Input	Type	WMD	CIDEr	METEOR	BLEU-1	BLEU-2	BLEU-3	BLEU-4
IMG	+MAX	7.48	25.66	11.20	24.51	13.67	7.96	4.51
LNG	+MAX	7.19	22.27	10.81	23.20	12.69	7.34	4.19
IMG+LNG	+MAX	**7.61**	**26.38**	**11.30**	**25.10**	**13.88**	**8.11**	**4.61**
IMG	+ATT	7.47	26.01	11.26	24.88	**13.99**	**8.13**	**4.67**
LNG	+ATT	7.20	22.11	10.82	23.20	12.55	7.16	3.97
IMG+LNG	+ATT	**7.54**	**26.04**	**11.28**	**24.96**	13.82	8.04	4.60

1. using multimodal features seems to improve the quality of generated paragraphs
2. max-pooling performs overall better for multimodal features

Results: automatic metrics, diversity

Model Input	Type	mBLEU	self-CIDEr
IMG	+MAX	**50.63**	76.43
LNG	+MAX	52.24	75.59
IMG+LNG	+MAX	52.09	**76.46**
IMG	+ATT	51.82	75.51
LNG	+ATT	50.93	76.41
IMG+LNG	+ATT	**47.42**	**78.39**
GT	-	18.84	96.51

1. multimodal features along with attention improve the overall diversity of generated paragraphs

Results: automatic metrics, diversity

Model Input	Type	mBLEU	self-CIDEr
IMG	+MAX	**50.63**	76.43
LNG	+MAX	52.24	75.59
IMG+LNG	+MAX	52.09	**76.46**
IMG	+ATT	51.82	75.51
LNG	+ATT	50.93	76.41
IMG+LNG	+ATT	**47.42**	**78.39**
GT	-	18.84	96.51

1. multimodal features along with attention improve the overall diversity of generated paragraphs
2. the best performing model is still quite far from the scores for ground-truth paragraphs

Results: human evaluation

Input	Type	WC	OS	SS	PC	Mean
IMG	+MAX	31.58	38.24	**59.57**	**37.87**	41.81
LNG	+MAX	29.64	36.43	56.43	36.95	39.86
IMG+LNG	+MAX	**34.20**	**38.72**	57.85	37.06	41.95
Mean	+MAX	31.80	37.79	57.95	37.29	-
IMG	+ATT	36.91	45.10	69.34	32.27	45.90
LNG	+ATT	**37.06**	**46.78**	**72.95**	**40.88**	49.41
IMG+LNG	+ATT	33.81	37.67	45.37	34.71	37.89
Mean	+ATT	35.92	43.18	62.55	35.95	-
GT	-	89.83	87.36	83.07	84.78	-

Results: human evaluation

Input	Type	WC	OS	SS	PC	Mean
IMG	+MAX	31.58	38.24	**59.57**	**37.87**	41.81
LNG	+MAX	29.64	36.43	56.43	36.95	39.86
IMG+LNG	+MAX	**34.20**	**38.72**	57.85	37.06	41.95
Mean	+MAX	31.80	37.79	57.95	37.29	-
IMG	+ATT	36.91	45.10	69.34	32.27	45.90
LNG	+ATT	**37.06**	**46.78**	**72.95**	**40.88**	49.41
IMG+LNG	+ATT	33.81	37.67	45.37	34.71	37.89
Mean	+ATT	35.92	43.18	62.55	35.95	-
GT	-	89.83	87.36	83.07	84.78	-

1. **IMG+LNG+MAX** might be a beneficial choice in terms of word choice (WC) and object salience (OS): categories which are directly connected to the accuracy and diversity of paragraphs

Results: human evaluation

Input	Type	WC	OS	SS	PC	Mean
IMG	+MAX	31.58	38.24	**59.57**	**37.87**	41.81
LNG	+MAX	29.64	36.43	56.43	36.95	39.86
IMG+LNG	+MAX	**34.20**	**38.72**	57.85	37.06	41.95
Mean	+MAX	31.80	37.79	57.95	37.29	-
IMG	+ATT	36.91	45.10	69.34	32.27	45.90
LNG	+ATT	**37.06**	**46.78**	**72.95**	**40.88**	49.41
IMG+LNG	+ATT	33.81	37.67	45.37	34.71	37.89
Mean	+ATT	35.92	43.18	62.55	35.95	-
GT	-	89.83	87.36	83.07	84.78	-

1. **IMG+LNG+MAX** might be a beneficial choice in terms of word choice (WC) and object salience (OS): categories which are directly connected to the accuracy and diversity of paragraphs
2. models with attention have higher mean scores across all criteria compared to the ones of models with max-pooling

Results: human evaluation

Input	Type	WC	OS	SS	PC	Mean
IMG	+MAX	31.58	38.24	**59.57**	**37.87**	41.81
LNG	+MAX	29.64	36.43	56.43	36.95	39.86
IMG+LNG	+MAX	**34.20**	**38.72**	57.85	37.06	41.95
Mean	+MAX	31.80	37.79	57.95	37.29	-
IMG	+ATT	36.91	45.10	69.34	32.27	45.90
LNG	+ATT	**37.06**	**46.78**	**72.95**	**40.88**	49.41
IMG+LNG	+ATT	33.81	37.67	45.37	34.71	37.89
Mean	+ATT	35.92	43.18	62.55	35.95	-
GT	-	89.83	87.36	83.07	84.78	-

1. **IMG+LNG+MAX** might be a beneficial choice in terms of word choice (WC) and object salience (OS): categories which are directly connected to the accuracy and diversity of paragraphs
2. models with attention have higher mean scores across all criteria compared to the ones of models with max-pooling
3. **LNG+ATT** performs much better than **IMG+ATT** for sentence structure (SS) and paragraph coherence (PC): categories where semantic information would matter the most

Results: human evaluation

Input	Type	WC	OS	SS	PC	Mean
IMG	+MAX	31.58	38.24	**59.57**	**37.87**	41.81
LNG	+MAX	29.64	36.43	56.43	36.95	39.86
IMG+LNG	+MAX	**34.20**	**38.72**	57.85	37.06	41.95
Mean	+MAX	31.80	37.79	57.95	37.29	-
IMG	+ATT	36.91	45.10	69.34	32.27	45.90
LNG	+ATT	**37.06**	**46.78**	**72.95**	**40.88**	49.41
IMG+LNG	+ATT	33.81	37.67	45.37	34.71	37.89
Mean	+ATT	35.92	43.18	62.55	35.95	-
GT	-	89.83	87.36	83.07	84.78	-

1. **IMG+LNG+MAX** might be a beneficial choice in terms of word choice (WC) and object salience (OS): categories which are directly connected to the accuracy and diversity of paragraphs
2. models with attention have higher mean scores across all criteria compared to the ones of models with max-pooling
3. **LNG+ATT** performs much better than **IMG+ATT** for sentence structure (SS) and paragraph coherence (PC): categories where semantic information would matter the most
4. attention seems to affect semantic information more than visual features

Results: paragraph examples

(a) **HUMAN**: *There are several cars parked along a street. There are many trees in a field in front of the street. There are small blue parking meters on the sidewalk next to the street.*
IMG+MAX : *There are several cars parked on the road. There are cars parked on the street. There are trees behind the street.*
LNG+MAX : *There are several cars on the street. There are trees on the street. There are trees on the street.*
IMG+LNG+MAX : *There are several cars on the street. There are two cars on the street. There are cars parked on the sidewalk.*
IMG+ATT : *There are several cars parked on the street. There are two cars parked on the road. There are two cars parked on the road.*
LNG+ATT : *There are several signs on the street. There are signs on the street. The pole is white.*
IMG+LNG+ATT : *There is a parking meter on a sidewalk. There are cars next to the street. There is a parking lot next to the street.*

(b) **HUMAN**: *A large splash is in front of a wave in the water. There is a large white and black surf board in the water. There is a black dog that is riding on top of the surf board.*
IMG+MAX : *A man is riding a wave. He is holding a surfboard. The man is wearing a black wet suit.*
LNG+MAX : *A person is surfing in the water. The surfboard is black and white. The surfboard is black and white.*
IMG+LNG+MAX : *A man is standing on a surfboard. The surfboard is black. The man is wearing black shorts.*
IMG+ATT : *A man is standing on a surfboard. The surfboard is black and white. The man has black hair.*
LNG+ATT : *A person is standing in the water. The person is wearing a black suit. The person is holding a black surfboard.*
IMG+LNG+ATT : *A person is surfing in the ocean. She is wearing a black wet suit. She is holding a white surfboard.*

Conclusion and Future Work

- Multimodal features **improve** the quality of paragraphs generated by image paragraph models in various ways as judged by both automatic and human evaluation

Conclusion and Future Work

- Multimodal features **improve** the quality of paragraphs generated by image paragraph models in various ways as judged by both automatic and human evaluation
- We need more control over human evaluation, more plausible automatic metrics for diversity

Conclusion and Future Work

- Multimodal features **improve** the quality of paragraphs generated by image paragraph models in various ways as judged by both automatic and human evaluation
- We need more control over human evaluation, more plausible automatic metrics for diversity
- We plan to investigate more the effects of **early** vs. **late** information fusion

Conclusion and Future Work

- Multimodal features **improve** the quality of paragraphs generated by image paragraph models in various ways as judged by both automatic and human evaluation
- We need more control over human evaluation, more plausible automatic metrics for diversity
- We plan to investigate more the effects of **early** vs. **late** information fusion
- How would using different decoding strategies (sampling, Nucleus sampling, etc.) affect the quality of paragraphs?

Conclusion and Future Work

- Multimodal features **improve** the quality of paragraphs generated by image paragraph models in various ways as judged by both automatic and human evaluation
- We need more control over human evaluation, more plausible automatic metrics for diversity
- We plan to investigate more the effects of **early** vs. **late** information fusion
- How would using different decoding strategies (sampling, Nucleus sampling, etc.) affect the quality of paragraphs?
- Our goal is to investigate the generation of task-dependent paragraphs (more structured and ordered)

Conclusion and Future Work

- Multimodal features **improve** the quality of paragraphs generated by image paragraph models in various ways as judged by both automatic and human evaluation
- We need more control over human evaluation, more plausible automatic metrics for diversity
- We plan to investigate more the effects of **early** vs. **late** information fusion
- How would using different decoding strategies (sampling, Nucleus sampling, etc.) affect the quality of paragraphs?
- Our goal is to investigate the generation of task-dependent paragraphs (more structured and ordered)
- Ultimately, we want to return to more interactive and dialogue settings as we initially thought about

Thank you for your attention!

[1]Krause, J., Johnson, J., Krishna, R., & Fei-Fei, L. (2017). A Hierarchical Approach for Generating Descriptive Image Paragraphs. In Computer Vision and Pattern Recognition (CVPR).

[2]Nikolai Ilinykh, Sina Zarrieß, and David Schlangen. 2019. MeetIp! A corpus of joint activity dialogues in a visual environment. In Proceedings of the 23rdWorkshop on the Semantics and Pragmatics of Dialogue (SemDial 2019 / LondonLogue), London, UK

[3]Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura and Devi Parikh, & Dhruv Batra (2017). Visual Dialog. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[4]Ilinykh, N., Zarriess, S., & Schlangen, D. (2019). Tell Me More: A Dataset of Visual Scene Description Sequences. In Proceedings of the 12th International Conference on Natural Language Generation (pp. 152–157). Association for Computational Linguistics.

[5]Ilinykh, N., Zarriess, S., & Schlangen, D. (2019). Tell Me More: A Dataset of Visual Scene Description Sequences. In Proceedings of the 12th International Conference on Natural Language Generation (pp. 152–157). Association for Computational Linguistics.

[6]Ilinykh, N., Zarriess, S., & Schlangen, D. (2019). Tell Me More: A Dataset of Visual Scene Description Sequences. In Proceedings of the 12th International Conference on Natural Language Generation (pp. 152–157). Association for Computational Linguistics.

[7]Johnson, J., Karpathy, A., & Fei-Fei, L. (2016). DenseCap: Fully Convolutional Localization Networks for Dense Captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

[8]Ozan Caglayan, Pranava Madhyastha, Lucia Specia, & Loïc Barrault. (2019). Probing the Need for Visual Context in Multimodal Machine Translation

[9]Ozan Caglayan, Loïc Barrault, & Fethi Bougares. (2016). Multimodal Attention for Neural Machine Translation.