# When an Image Tells a Story: The Role of Visual and Semantic Information for Generating Paragraph Descriptions

Nikolai Ilinykh , Simon Dobnik

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

$17^{th}$ December 2020

# 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.

[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).

# Properties of Image Paragraphs (IP)

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.

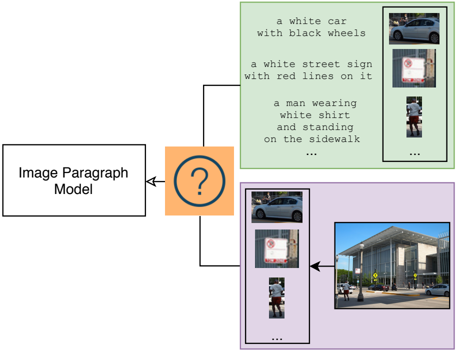
# Two Sources of Important Information for IP

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

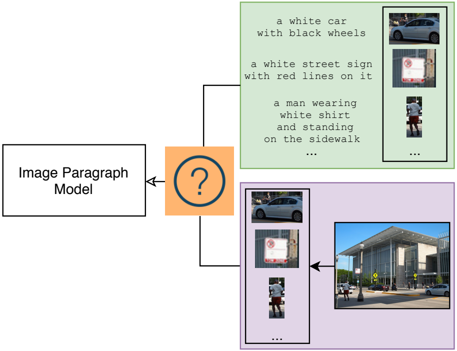
How to improve both *accuracy* and *diversity* of generated image paragraphs?




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

# Our paper

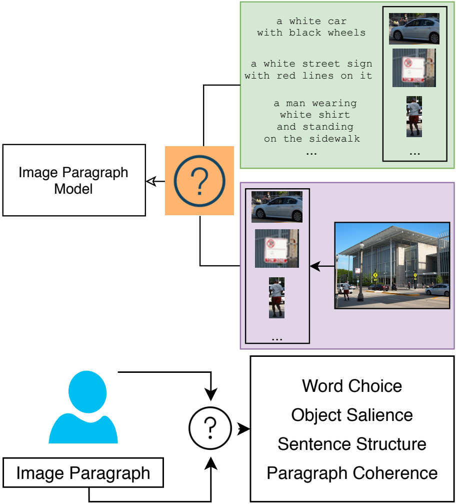
How to improve both *accuracy* and *diversity* of generated image paragraphs?




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

# Our paper

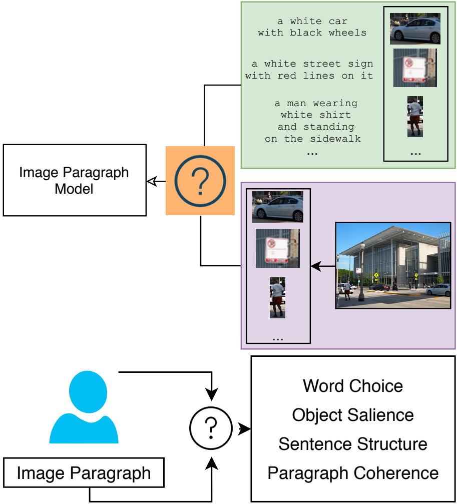
How to improve both *accuracy* and *diversity* of generated image paragraphs?




- **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?

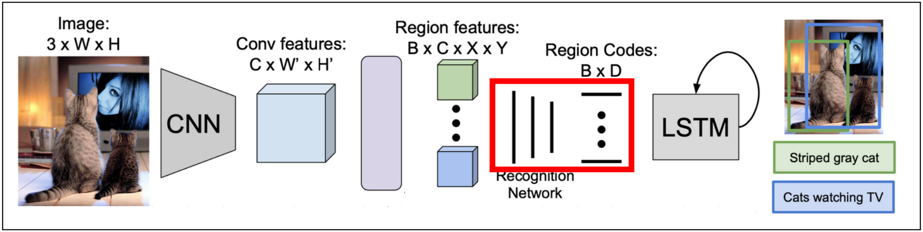



- **model input**: unimodal (visual / language) vs. multimodal
- **information fusion**: max-pooling vs. attention
- **paragraph evaluation**: automatic vs. human
- **human evaluation**: accuracy and diversity of generated paragraphs

# Unimodal Features: Vision, Language

We use pre-trained **DenseCap**[2] 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)



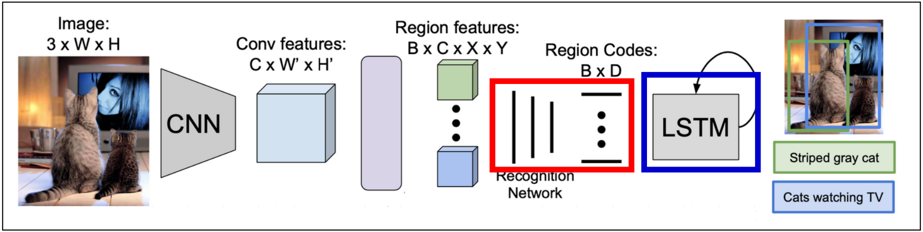

Notations: $M = 50, D = 4096, H = 512.$

[2] 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.

# Unimodal Features: Vision, Language

We use pre-trained **DenseCap**[3] 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: $M = 50, D = 4096, H = 512$.

[3] 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.

# Multimodal Features: Vision **and** Language

Mapping Visual Features

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

# Multimodal Features: Vision **and** Language

Mapping Visual Features

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

Mapping Language Features

# Multimodal Features: Vision **and** Language

Mapping Sentence LSTM
last hidden state

Mapping Visual Features

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

Mapping Language Features

# Multimodal Features: Vision **and** Language

$$mult_t = [\overbrace{\boxed{W_m^V V_t}}^{\text{Mapping Visual Features}} \oplus \underbrace{\boxed{W_m^L L_t}}_{\text{Mapping Language Features}} \oplus \overbrace{\boxed{W_h h_{t-1}^{\delta}}}^{\text{Mapping Sentence LSTM last hidden state}}]$$

**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 $x = W_m^L L_t$:

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

# 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 ($F_t$):

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

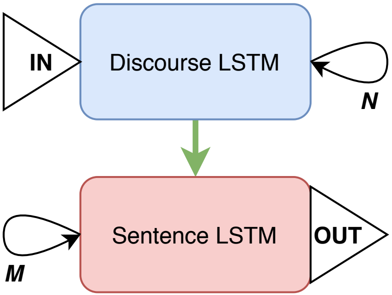
$$f_t = [\alpha_t^{mult} \odot F_t] \tag{6}$$

**Note**: Although some work on multimodal machine translation has shown that early attention improves quality of text generations [4,5] , using **modality-dependent / early** attention (unique $W_a^A$ and, therefore, unique $\alpha_t^{mult}$ for each modality) provided us with worse automatic metric scores.

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

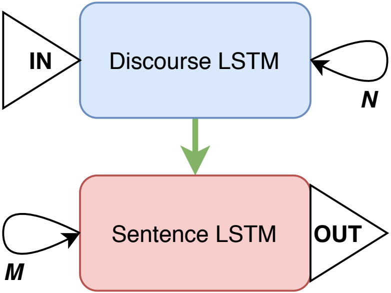
[5] Ozan Caglayan, Loïc Barrault, & Fethi Bougares. (2016). Multimodal Attention for Neural Machine Translation.

# Image Paragraph Model



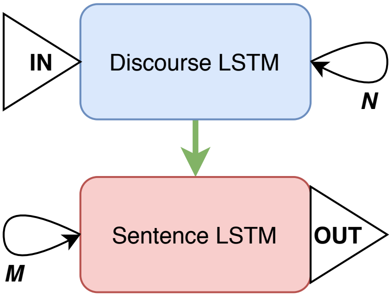

- **IN**: visual / language / multimodal features

# 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 (PP): 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 (PP): categories where semantic information would matter the most
4. attention seems to affect semantic information more than visual features

# 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)

**We thank you for your attention!**

All code is available at the [github link]