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.

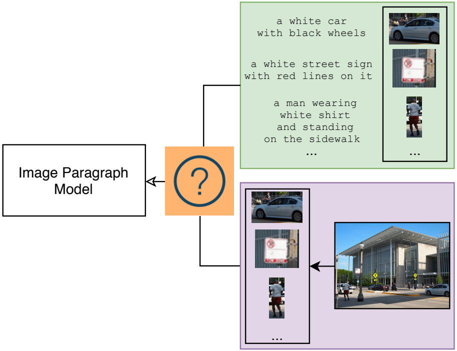
Two Sources of Important Information for IP

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

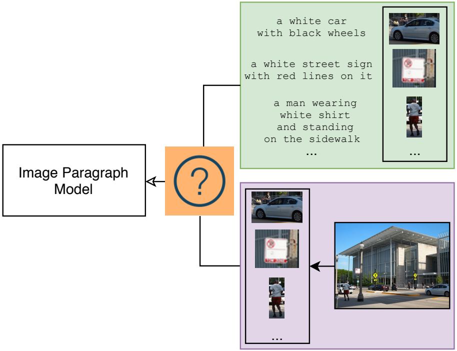
How to improve both *accuracy* and *diversity* of generated image paragraphs?



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

Our paper

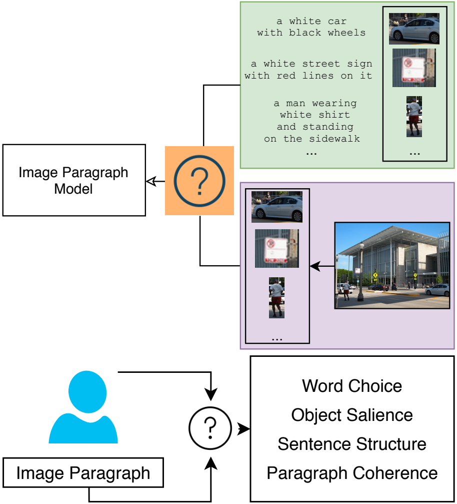
How to improve both *accuracy* and *diversity* of generated image paragraphs?



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

Our paper

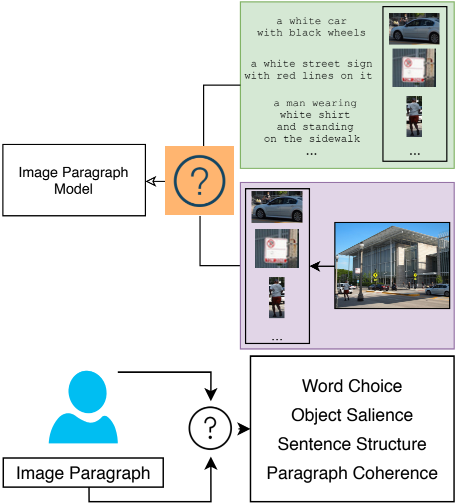
How to improve both *accuracy* and *diversity* of generated image paragraphs?



- **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?

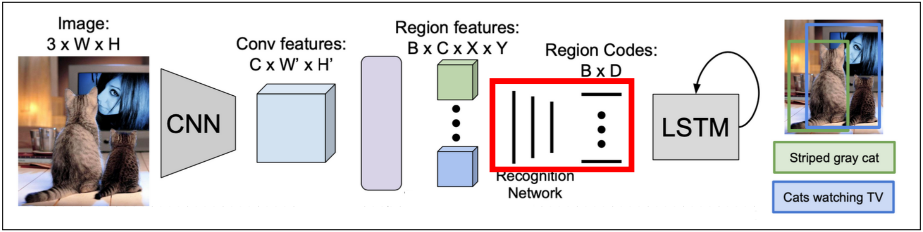


- **model input**:
  unimodal (visual / language) vs. multimodal
- **information fusion**:
  max-pooling vs. attention
- **paragraph evaluation**:
  automatic vs. human
- **human evaluation**:
  accuracy and diversity of generated paragraphs

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)



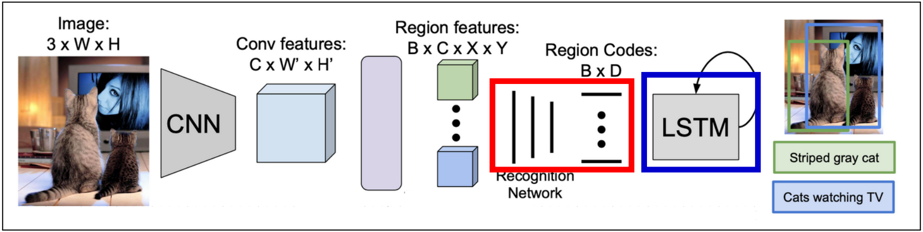
Notations: $M = 50, D = 4096, H = 512.$

[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 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 applied **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 applied **additive\concat** attention on either unimodal or multimodal features (F_t):

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

$$f_t = [\alpha_t^{mult} \odot F_t] \quad (6)$$

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 α_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