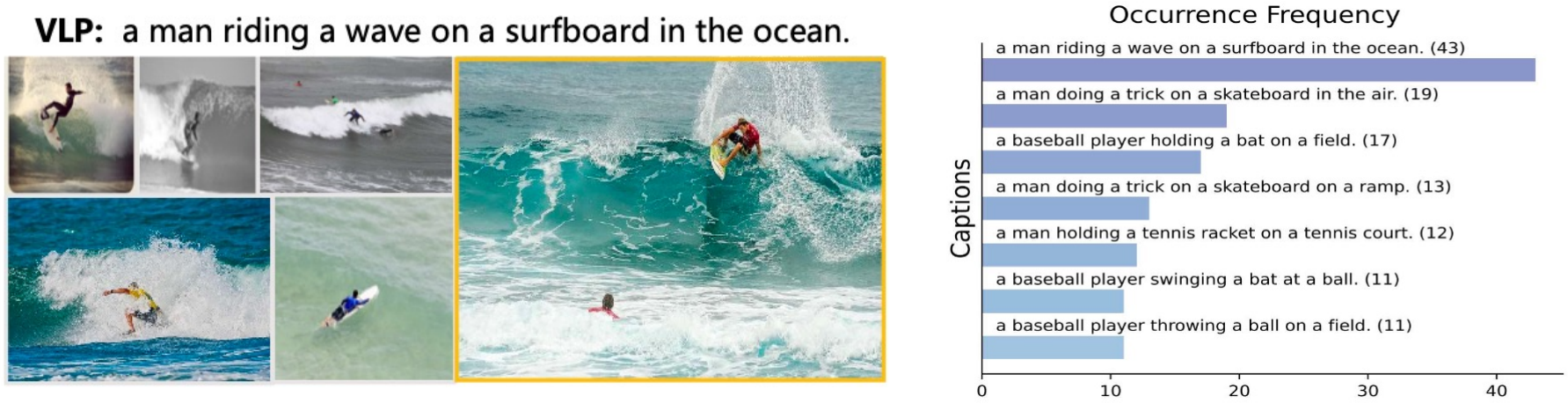
Learning Descriptive Image Captioning via Semipermeable Maximum Likelihood Estimation

Zihao Yue, Anwen Hu, Liang Zhang, Qin Jin


AIM³ Lab, Renmin University of China

Background

- **Image captioning suffers from overly-generic outputs**
  - Generate with simple concepts and minimal details
  - Repetitive captions for different images



Repetitive captions generated by VinVL[A] for different images in the MSCOCO test set.[B]

[A] Vinvl: revisiting visual representations in vision-language models, in *CVPR 2021*
[B] CapEnrich: enriching caption semantics for web images via cross-modal pre-trained knowledge, in *WWW 2023*

Foundation

- **Image Captioning Optimization: Maximum Likelihood Estimation (MLE)**
  - Proxied by Next-token Prediction
    - Predictive distribution of the next token
    - Cross Entropy Loss

$$\mathcal{L}_{\text{MLE}} = -\sum_{j}^{|\mathcal{V}|} y_j \log \hat{P}^{\mathcal{V}}(w|w_<, v; \theta),$$

  - ‘A picture is worth a thousand words’
    - Various correct descriptions for an image
    - Multiple reasonable answers for token prediction
    - MLE’s strict supervision is not perfectly suitable for captioning optimization!

SMILE: Semipermeable MLE

- **What happens when optimizing with MLE**
  - Penalize model whenever its prediction mismatches label, leading to:
    - **Richness Optimization**, which makes prediction richer
    - **Conciseness Optimization**, which makes prediction more concise

Case 1:

a ___ woman holding a cake

GT: pretty Prediction: woman (*less rich than GT*)

→ ***Richness Optimization***

Case 2:

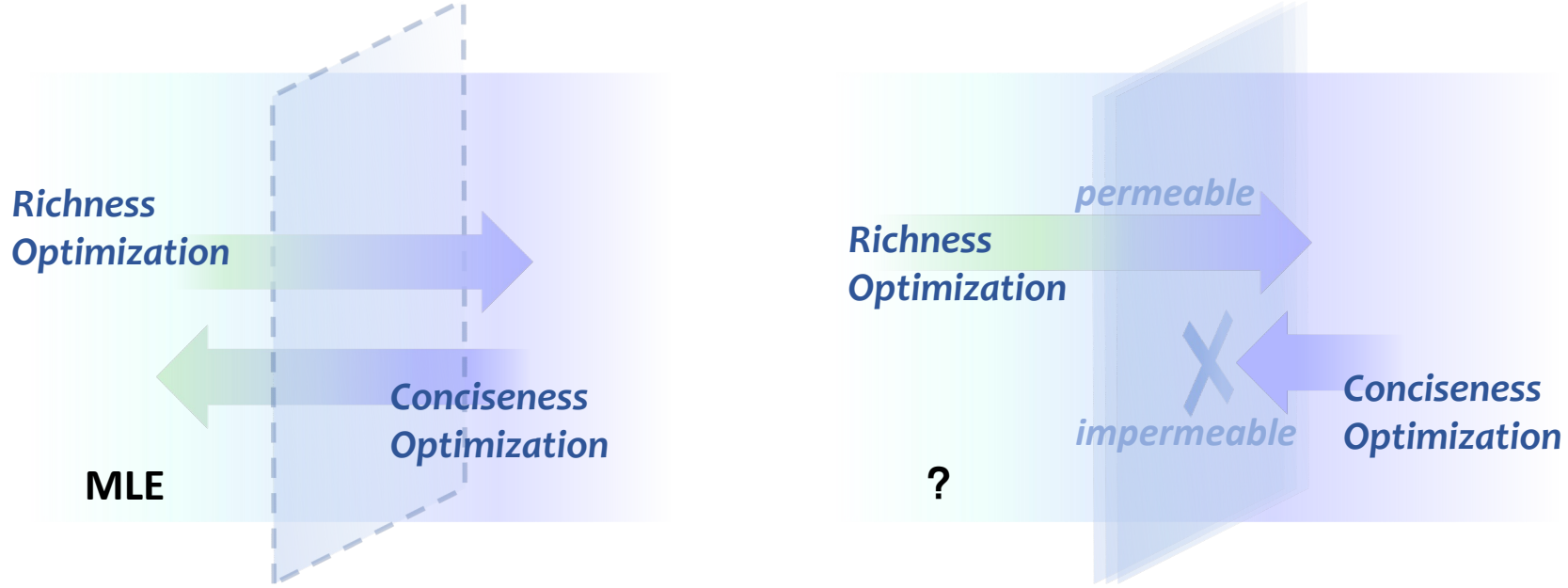
a pretty woman holding a ___

GT: cake Model: white (*less concise than GT*)

→ ***Conciseness Optimization***

SMILE: Semipermeable MLE

- Can we allow rich. optimization while preventing concise. optimization?



- **S**emipermeable **M**ax**I**mum **L**ikelihood **E**stimation (SMILE)

SMILE: Semipermeable MLE

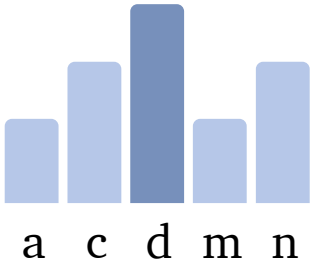
- **A simple modification to MLE loss** *CrossEntropy(P^V, P^y)*
  - Calculating cross entropy loss over a small vocabulary subset V_D
  - The subset contains only words in the ground truth sentence

$$\mathcal{L}_{\text{SMILE}} = -\sum_{j}^{|\mathcal{V}_D|} y_j \log \hat{P}^{\mathcal{V}_D}(w|w_<, v; \theta). \qquad \hat{p}_j = \text{softmax}(\mathbf{z}_j) = \frac{\exp(\mathbf{z}_j)}{\sum_{k \in \mathcal{V}_D} \exp(\mathbf{z}_k)}.$$

Ground truth sentence: " a c d m n "

Next-token prediction: " a c _ (d) "

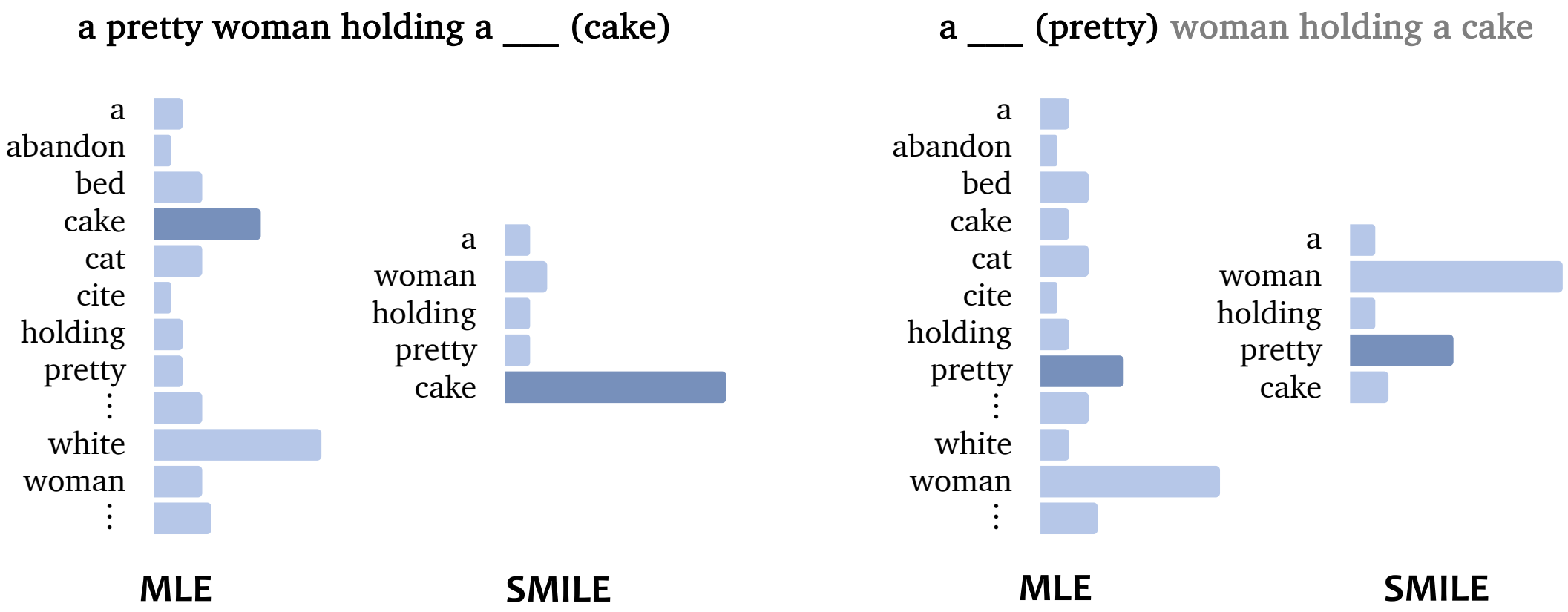
Predictive distribution over entire vocabulary **(MLE)**



Predictive distribution over the subset **(SMILE)**

SMILE: Semipermeable MLE

- **How SMILE responds to Concise. Optimization and Rich. Optimization?**



- Conciseness Optimization is removed; Richness Optimization is reserved

SMILE: Semipermeable MLE

- **Key idea behind the subsetting strategy:**
  - "Richer" words expressing additional details are typically outside GT caption
    - e.g., "white" for "a pretty woman holding a cake"
    - By subsetting we exclude "richer" words causing conciseness optimization
  - "More concise" words are likely elements in the ground truth caption
    - e.g., "woman" for "a pretty woman holding a cake"
    - Subsetting won't exclude "more concise" words causing richness optimization

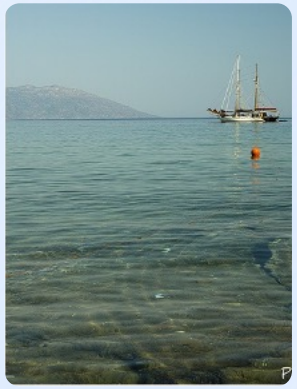
- Allowing richness optimization while preventing conciseness optimization

Cases

- **Descriptive captions generated by SMILE-optimized captioning model**

Human
a woman holding a birthday cake with lit candles.
MLE
a woman holding a cake with lit candles.
SMILE (ours)
a pretty young lady that has some kind of white frosted birthday cake with lots of lit candles on top of it, surrounded by several other people looking onwardly at something in the distance.

Human
a wire with a street light hanging from it.
MLE
a traffic light hanging from the side of a building.
SMILE (ours)
a close-up image of two red stoplights hanging from an electrical cable system outside a large brick church building, during the early morning hours.



Human
ocean showing a boat sailing on the waters.
MLE
a boat floating on top of a large body of water.
SMILE (ours)
a lone fishing vessel that appears to be floating peacefully across the vast expanse of crystal blue water near an island in the far distance.

Human
a man in black surfing in high and strong waves.
MLE
a man riding a wave on top of a surfboard.
SMILE (ours)
a young adult male leans forward as he stands atop an extremely wide red and white striped surfboard in the midst of crashing waves.

SMILE: Semipermeable MLE

- **Paper**
  - https://arxiv.org/abs/2306.13460
- **Code**
  - https://github.com/yuezih/SMILE
- **Demo**
  - https://huggingface.co/spaces/yuezih/BLIP-SMILE
- **Contact**
  - yzihao@ruc.edu.cn