AI·M³

中国人民大学多媒体计算实验室

文本生成中的部分词表学习：应用与启示

Partial Vocabulary Learning for Neural Text Generation

岳子豪 @中国人民大学

Outline

AI·M³
中国人民大学多媒体计算实验室

- Preliminary: Language Modeling
- Partial Vocabulary Learning
- Application
  - Descriptive Image Captioning
  - Multimodal Hallucination
- Discussion & Future Works

Preliminary: Language Modeling

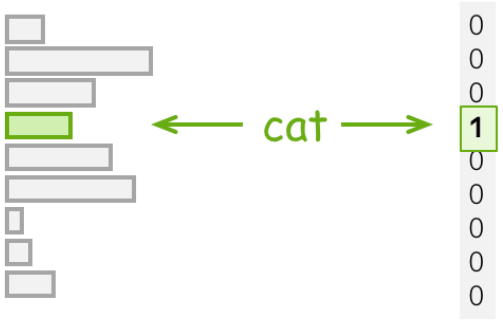
- 语言模型通常由下一词预测任务进行训练

we want the model to predict this

Training example: I saw a cat on a mat <eos>

给定前 N-1 个词，预测第 N 个词

Model prediction: $p(* \mid \text{I saw a})$ Target

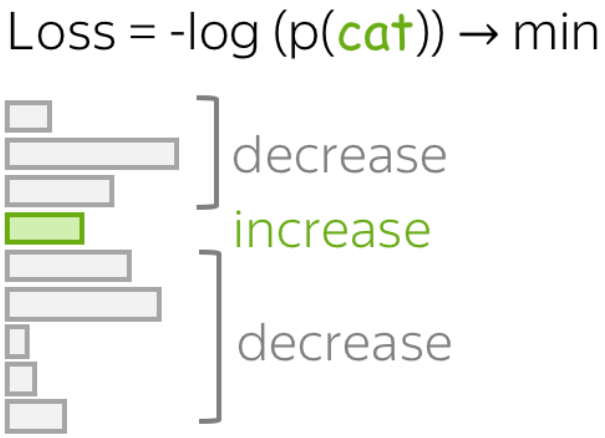


预测概率分布 vs. Ground truth 分布 (one-hot)

picture from https://lena-voita.github.io/nlp_course/language_modeling.html, blog by Lena Voita

Preliminary: Language Modeling

- 语言模型通常由下一词预测任务进行训练



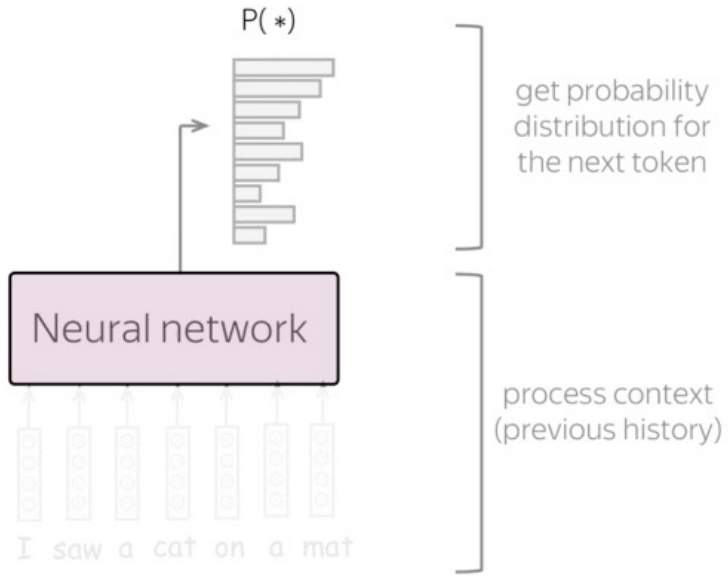
最大似然估计（交叉熵损失）：

- 最大化 label 的概率
- 最小化其它单词的概率

picture from https://lena-voita.github.io/nlp_course/language_modeling.html, blog by Lena Voita

Preliminary: Language Modeling

- 当模型学会下一词预测后，就能自回归完成句子



picture from https://lena-voita.github.io/nlp_course/language_modeling.html, blog by Lena Voita

Preliminary: Language Modeling

- 下一词预测是现代语言模型的基础



- 但并非没有问题……
  - Does the next token always worth learning?

Preliminary: Language Modeling

- 自然语言表达是多样的
  - I am happy to be here to give a talk
  - I'm glad to be here to give a talk
  - I feel good to be here to give a talk
  - I am pleased to be here to give a talk
- label 不一定是最好的/唯一正确的

Ground Truth

a pretty woman holding a cake ···

Model

a pretty woman holding a white ···

Is "cake" better than "white"?

- 强制模型学习 label 可能适得其反

Preliminary: Language Modeling

- 当最大化 label 概率时，其它单词的概率向 label 转移

P(better) – –转移–-> P(label) (undisired)

P(worse) – –转移–-> P(label) (desired)

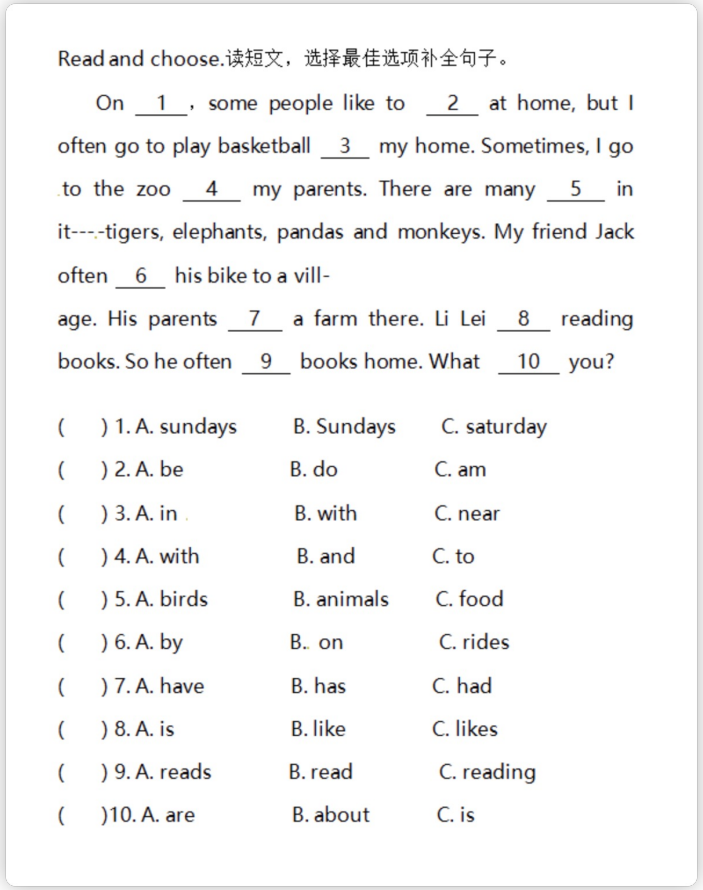
- 当词表中可能存在更优选项时，如何避免更优选项被惩罚，导致概率降低?

~~P(better) – –转移–-> P(label) (undisired)~~

P(worse) – –转移–-> P(label) (desired)

Partial Vocabulary Learning

➢ 从完形填空中获得启发



Read and choose.读短文，选择最佳选项补全句子。

On __1__, some people like to __2__ at home, but I often go to play basketball __3__ my home. Sometimes, I go to the zoo __4__ my parents. There are many __5__ in it----tigers, elephants, pandas and monkeys. My friend Jack often __6__ his bike to a vill-age. His parents __7__ a farm there. Li Lei __8__ reading books. So he often __9__ books home. What __10__ you?

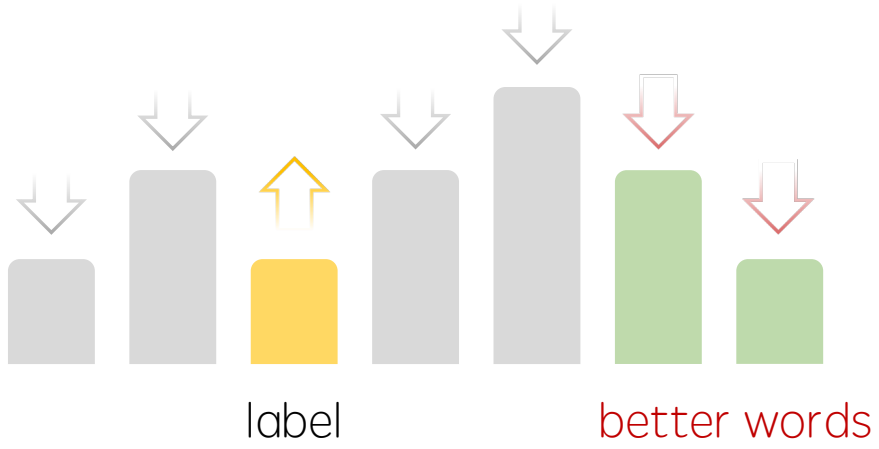
( ) 1. A. sundays B. Sundays C. saturday
( ) 2. A. be B. do C. am
( ) 3. A. in B. with C. near
( ) 4. A. with B. and C. to
( ) 5. A. birds B. animals C. food
( ) 6. A. by B. on C. rides
( ) 7. A. have B. has C. had
( ) 8. A. is B. like C. likes
( ) 9. A. reads B. read C. reading
( )10. A. are B. about C. is

	完形填空	语言建模
任务形式	单词预测	单词预测
学习方式	Maximize P(label)	Maximize P(label)
选项空间	4 选 1	50000 选 1

启发：选择词表子集构成选项空间

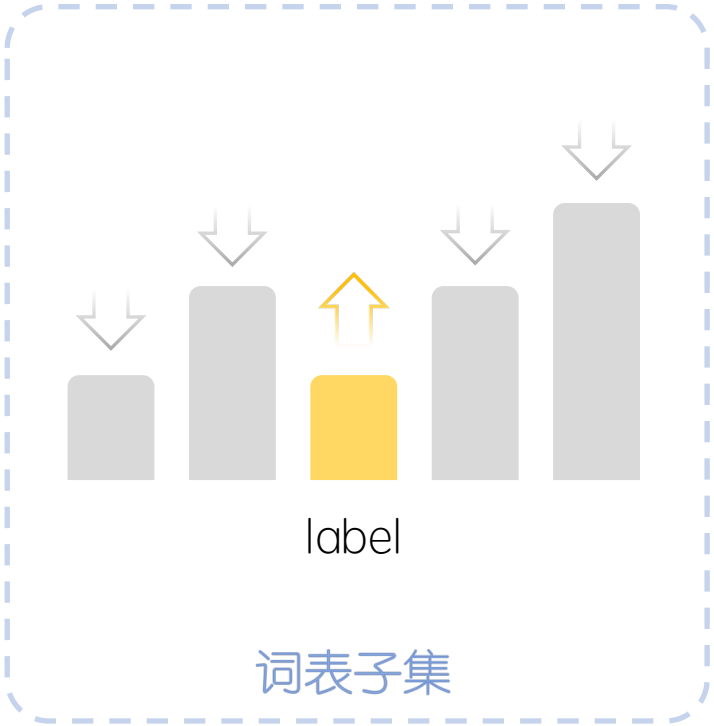
Partial Vocabulary Learning

- 常规学习目标：最大似然估计



Partial Vocabulary Learning

- 部分词表学习：选择词表子集构成选项空间，仅在子集中计算交叉熵损失



子集之外的词不会被抑制

Partial Vocabulary Learning

- 如何选择子集?

  对于不同任务，发现问题，寻找规律

- 应用示例

详细图像描述

仅从简单描述中学习，模型能否生成详细描述?

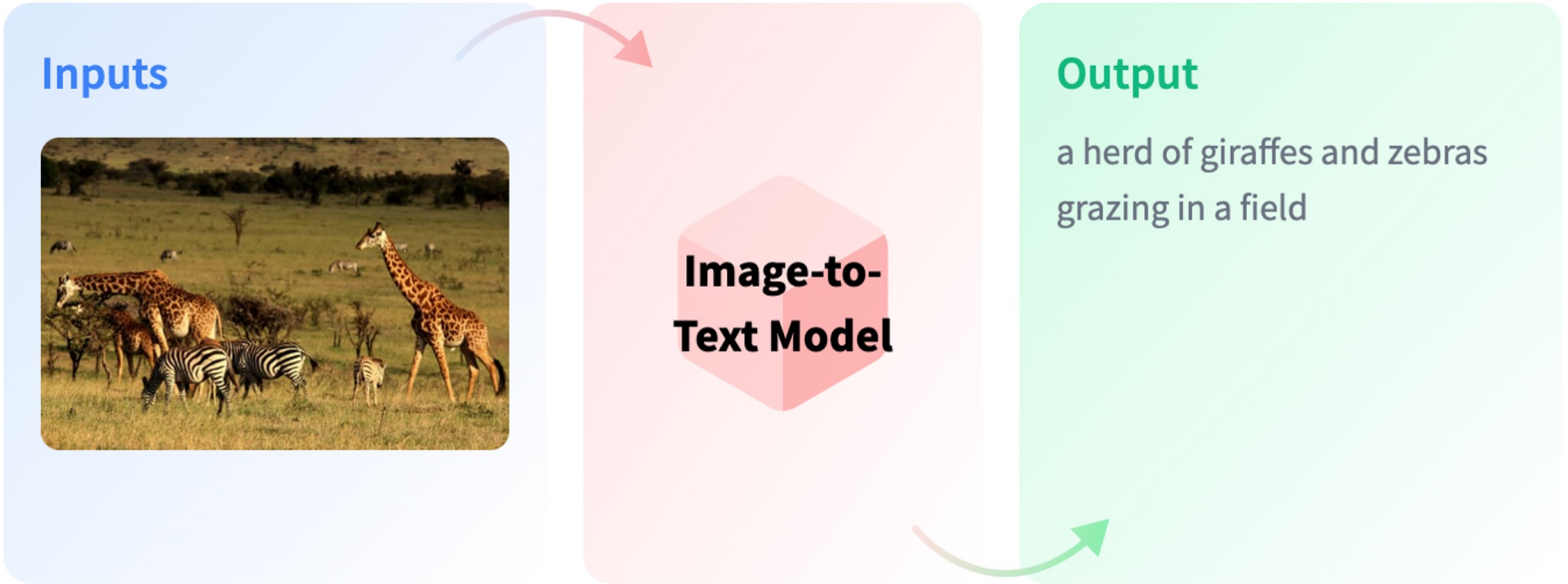
多模态大模型幻觉

模型倾向于输出自己不知道的内容，如何让模型适可而止?

Application: Descriptive Image Captioning

Application: Descriptive Image Captioning

- 图像描述 (Image Captioning): 为给定图像生成自然语言描述



picture from https://huggingface.co/tasks/image-to-text

Application: Descriptive Image Captioning

- MSCOCO, Flickr30K 等主流图像描述数据通常很短 (e.g. 平均 10 个单词)

A baby in plaid shirt eating a frosted cake.

A baby in plaid shirt eating a frosted cake.

A toddler eats cake with his hands in his high chair.

A toddler is getting messy while eating his cake.

Baby boy at the table eating cake frosting off his hand.

data examples from MSCOCO

Application: Descriptive Image Captioning

- 基于这些数据训练的模型，也会生成很短的描述，信息量少 & 过度重复

VinVL 模型为 MSCOCO 测试集中的 43 张冲浪图像生成完全相同的描述

picture from CapEnrich: Enriching Caption Semantics for Web Images via Cross-modal Pre-trained Knowledge (WWW'23)

Application: Descriptive Image Captioning

- 仅从简单描述中学习，模型能否生成详细描述?

training data

a woman holding a birthday cake with lit candles.

generated

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.

Application: Descriptive Image Captioning

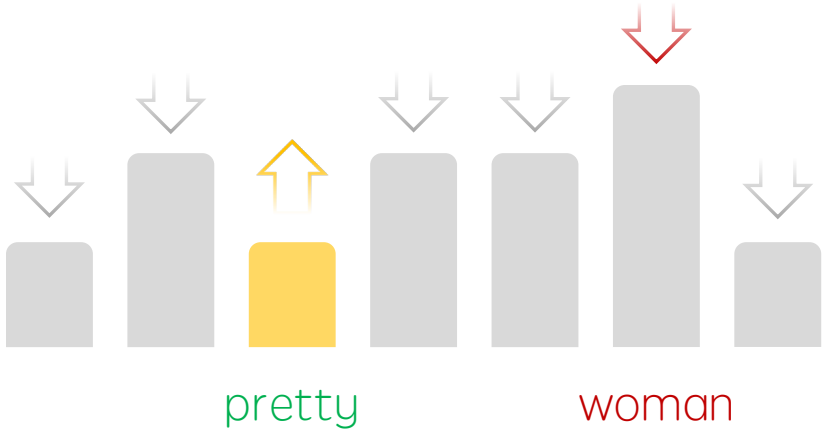
- 常规训练方法：最大似然估计

training data

a ___ (pretty) woman holding a cake

label: pretty

model: woman



增加 pretty 概率，降低 woman 概率，模型变得更倾向于预测 detail

Richness optimization

Application: Descriptive Image Captioning

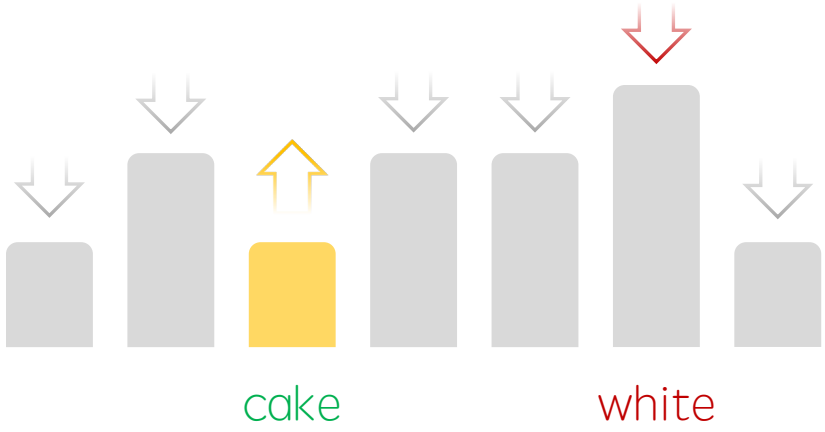
- 常规训练方法：最大似然估计

training data

a pretty woman holding a ___ (cake)

label: cake

model: white



增加 cake 概率，降低 white 概率，模型变得更不倾向于预测 detail

Conciseness optimization

Application: Descriptive Image Captioning

- 两种相互冲突的优化共同存在于训练过程中

  Richness optimization　　Conciseness optimization

- 如果希望生成更丰富的描述，应该：
  - 鼓励 Richness optimization
  - 抑制 Conciseness optimization

Application: Descriptive Image Captioning

- 部分词表学习：避免概率从 更详细的词 转移到 更简洁的词

e.g. white

e.g. cake

- 选择 ground truth caption 自身包含的单词作为子集

a
pretty
woman
holding
cake

Application: Descriptive Image Captioning

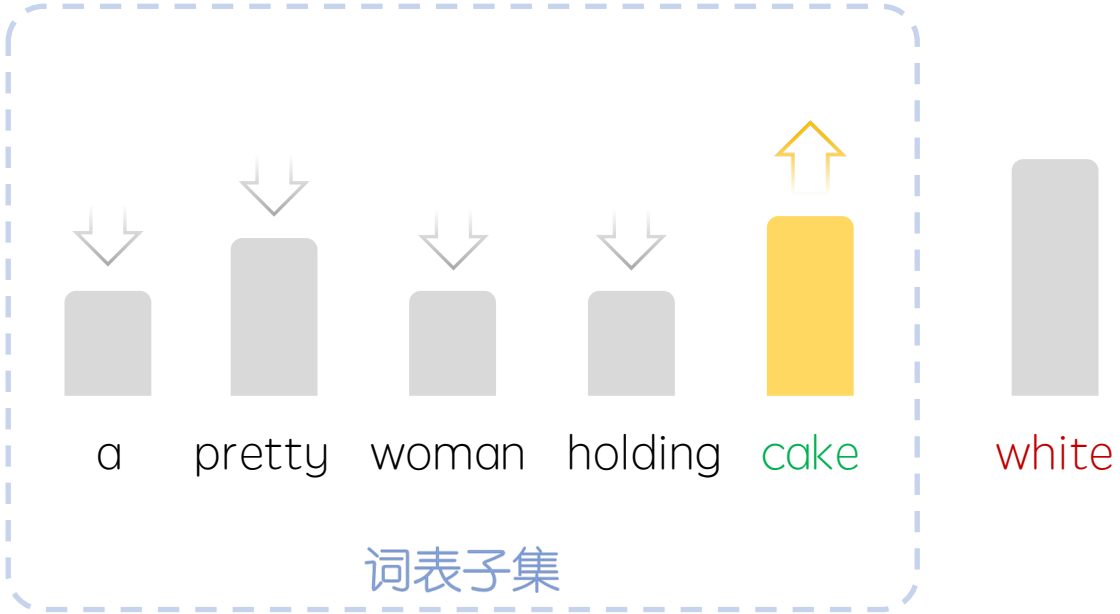
- 仅在子集中学习

training data

a pretty woman holding a ___ (cake)

label: cake

model: white



增加 cake 概率，不会降低 white 概率 (white 在子集之外)

Prevent conciseness optimization

Application: Descriptive Image Captioning

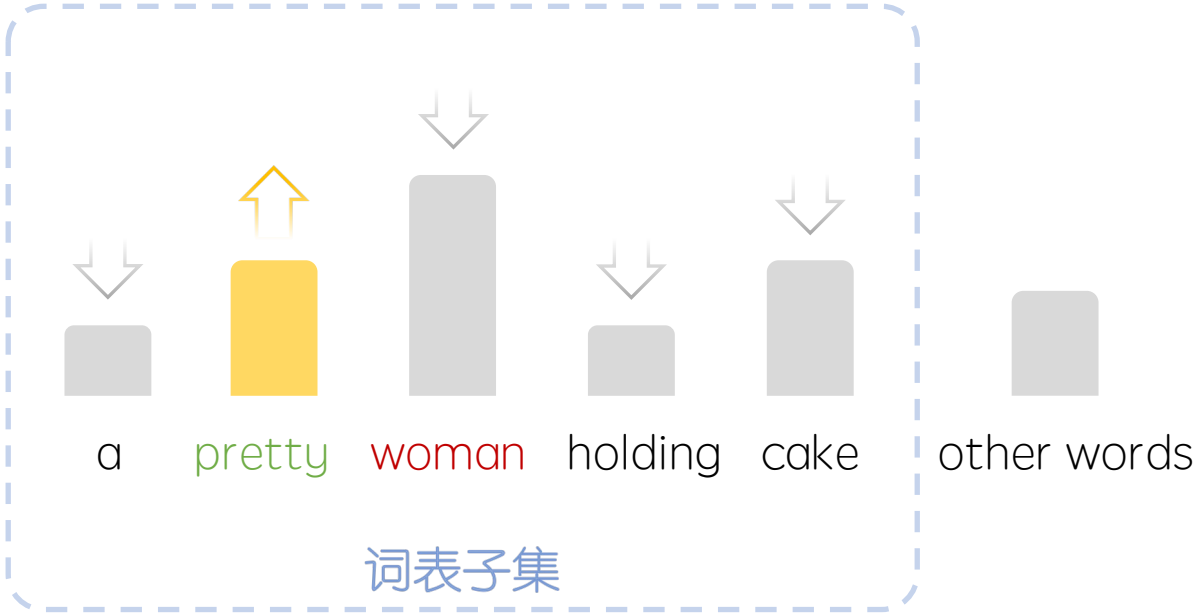
- 仅在子集中学习

training data

a ___ (pretty) woman holding a cake

label: pretty

model: woman



增加 pretty 概率，会降低 woman 概率 (woman 在子集之内)

Preserve richness optimization

Application: Descriptive Image Captioning

- 部分词表学习

Prevent conciseness optimization

Preserve richness optimization

- Why this works?
  - 更详细的词 (表达额外细节)，一般不在原句之内

    e.g. white (vs. cake)
  - 更简洁的词 (句子主干成分)，一般在原句之内

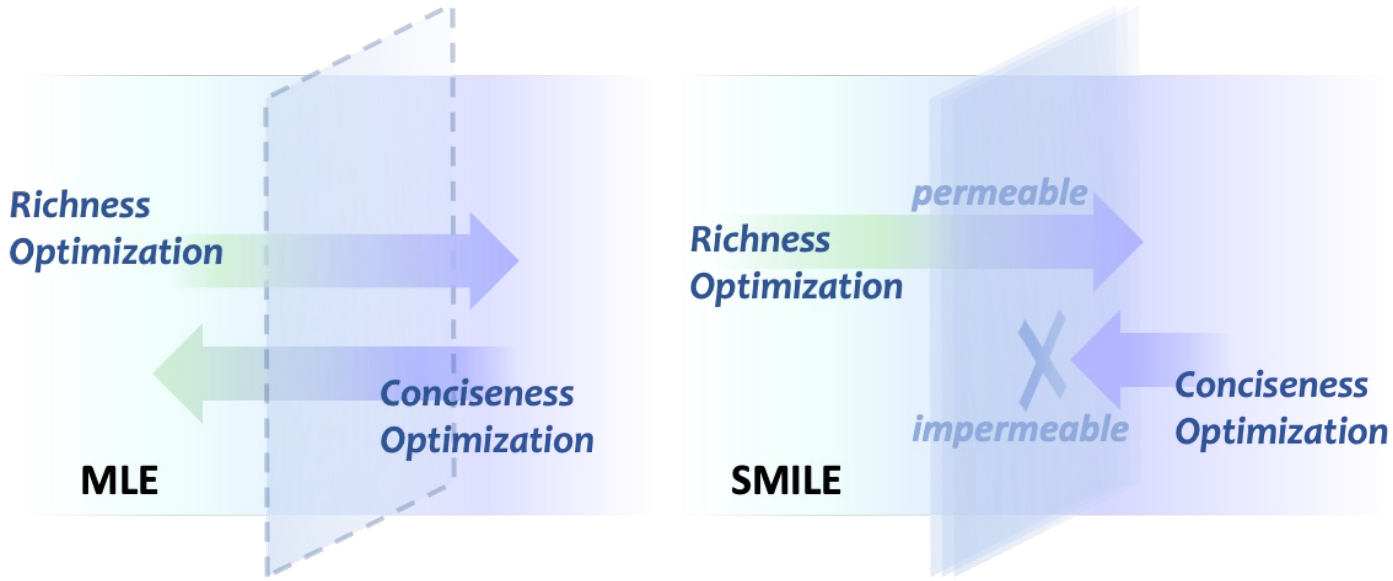
    e.g. woman (vs. pretty)

a
pretty
woman
holding
cake

white

Application: Descriptive Image Captioning

- Semipermeable (半渗透的) Maximum Likelihood Estimation (SMILE)
  - 屏蔽 conciseness optimization，只允许 richness optimization
  - 模型受到持续的 richness optimization，越来越倾向于生成更多细节



Application: Descriptive Image Captioning

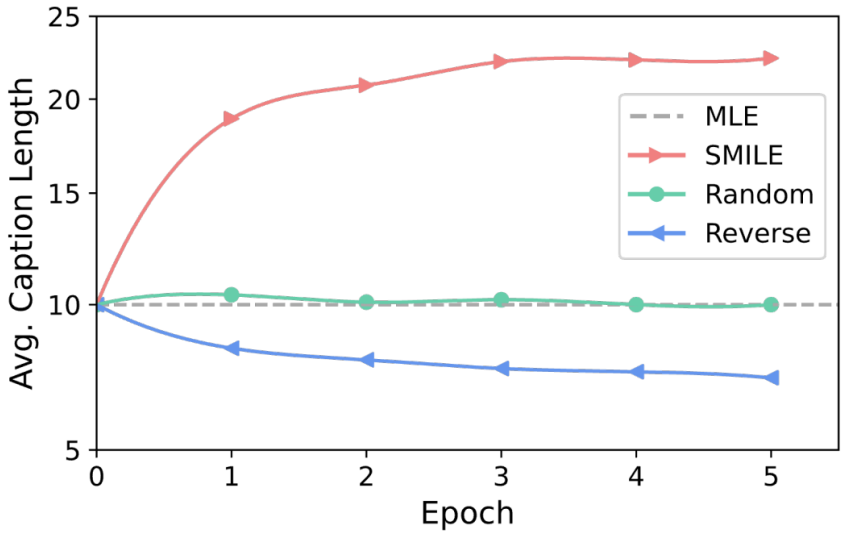
AI·M³ 中国人民大学多媒体计算实验室

- 使用 SMILE 对已有模型进行微调
  - 只需简单修改损失函数
  - 少量训练
  - 无需任何额外数据或知识

  大幅提升模型生成的丰富性

  0.2B BLIP 的 self-retrieval 性能超过 7B LLaVA/MiniGPT-4

- “半透膜”反过来，可以让输出更简洁



Application: Descriptive Image Captioning

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.

Application: Descriptive Image Captioning

AI·M³
中国人民大学多媒体计算实验室

FAQ

- 如果不惩罚子集之外的预测，如何保证准确性?

  子集之外的预测依赖 base model 自身性能

  SMILE 侧重于改变模型表达习惯——将模型输出更丰富内容的潜力释放出来

Application: Descriptive Image Captioning

- **部分词表学习**：详细图像描述

发现问题

训练数据中的一些 token 会抑制模型更详细的预测

总结规律

更详细词到更简洁词的“有害”概率转移

构建子集

选择句子自身包含的单词作为子集（更详细词往往出现在句子之外）

Application: Multimodal Hallucination

Application: Multimodal Hallucination

- 多模态大模型容易出现幻觉，输出自己无法正确识别的内容

Instruction: Please describe this image in detail.

Model: The image features a cute, red dog running across a grassy field ⋯ Around the dog, there are several other dogs visible in the background ⋯

case from Woodpecker: Hallucination Correction for Multimodal Large Language Models

Application: Multimodal Hallucination

- 幻觉重要来源：模型普遍倾向于过度表达，生成超出视觉识别极限的内容

Instruction: Please describe this image in detail.

LLaVA-1.5 (7B)

The image features a person wearing a wetsuit, standing in the water and holding a sailboard. The person is positioned in the middle of the scene, with the sailboard prominently visible in front of them. The water appears to be choppy, adding to the dynamic nature of the scene. In the background, there are several boats of varying sizes scattered across the water, creating a lively atmosphere. Some of the boats are closer to the person with the sailboard, while others are further away, providing a sense of depth to the image.

Application: Multimodal Hallucination

- 过度详细的训练数据破坏了模型正常的 expressiveness

In the image, a man is skillfully catching air while skateboarding in a public park in a city, seemingly flying through the air ⋯

Additionally, there are two benches visible in the park, one closer to the left and another further back ⋯

overly-detailed training data

Application: Multimodal Hallucination

AI·M³ 中国人民大学多媒体计算实验室

- 过度详细的训练数据破坏了模型正常的 expressiveness

In the image, a man is skillfully catching air while skateboarding in a public park in a city, seemingly flying through the air …

Additionally, there are two benches visible in the park, one closer to the left and another further back …

overly-detailed training data

exceeds model's visual perception capability

fit the detail level & length distribution

tends to overly generate details beyond visual perception limits

Application: Multimodal Hallucination

- 问题：Model Expressiveness > Model Visual Perception Capability

- Can we calibrate expressiveness to match visual perception capability?

  当模型生成达到视觉识别极限后，应及时停止生成（生成 EOS token）

  (1) 如何获取模型的视觉识别极限?

  (2) 如何校准模型的 expressiveness?

Application: Multimodal Hallucination

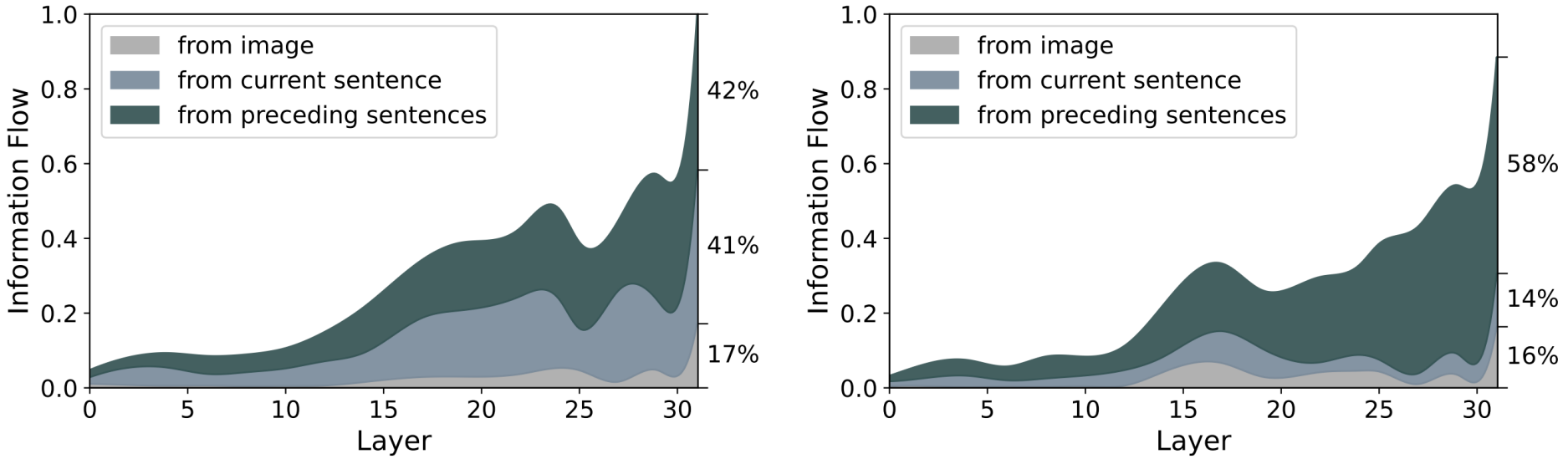
(1) 如何获取模型的视觉识别极限?

- 无需获取，模型自身具备根据视觉感知及时停止生成的潜力

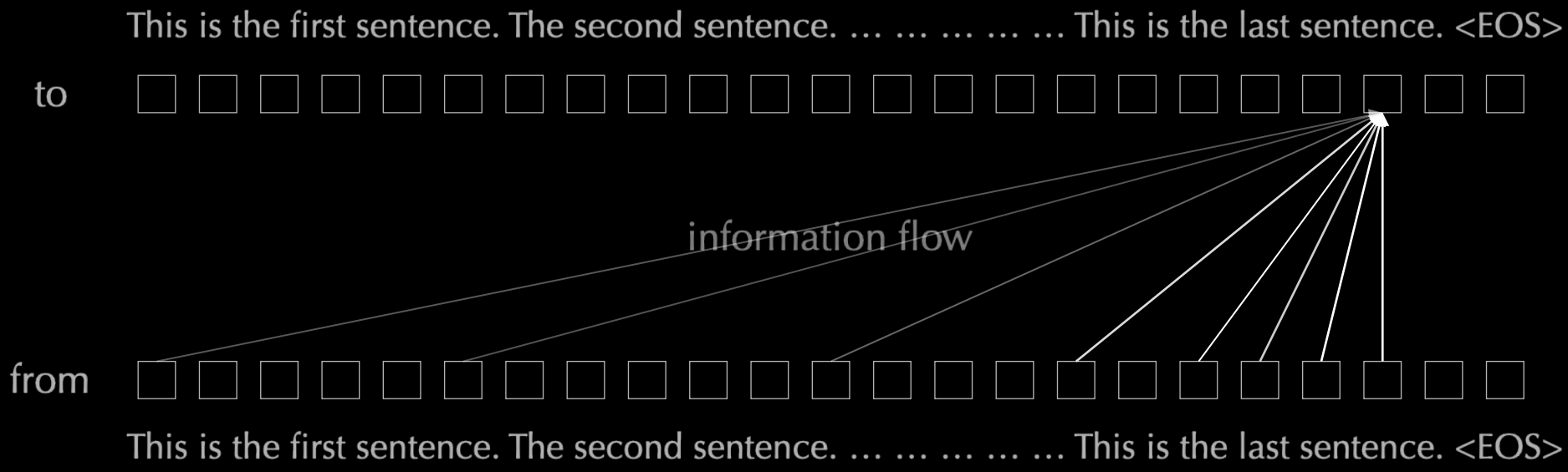
Application: Multimodal Hallucination

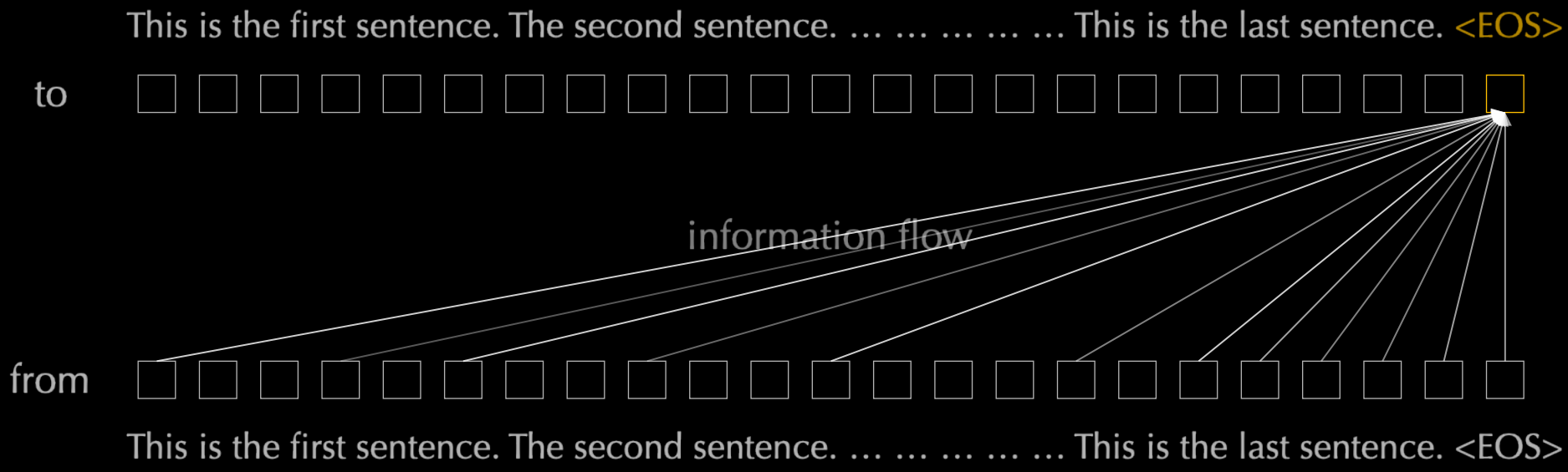
AI·M³ 中国人民大学多媒体计算实验室

(1) 如何获取模型的视觉识别极限?



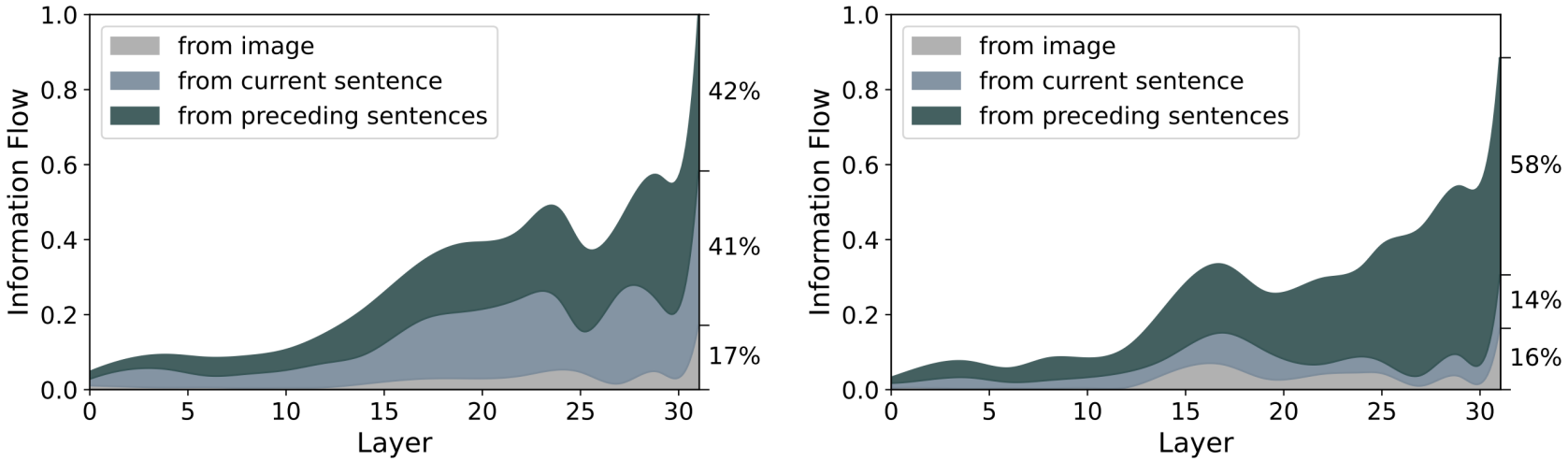
模型预测下一词时，上下文中流向预测位置的信息流：普通 token (左) 和 EOS token (右)


This is the first sentence. The second sentence. … … … … … This is the last sentence. <EOS>
to
information flow
from
This is the first sentence. The second sentence. … … … … … This is the last sentence. <EOS>


This is the first sentence. The second sentence. … … … … … This is the last sentence. <EOS>
to
information flow
from
This is the first sentence. The second sentence. … … … … … This is the last sentence. <EOS>

Application: Multimodal Hallucination

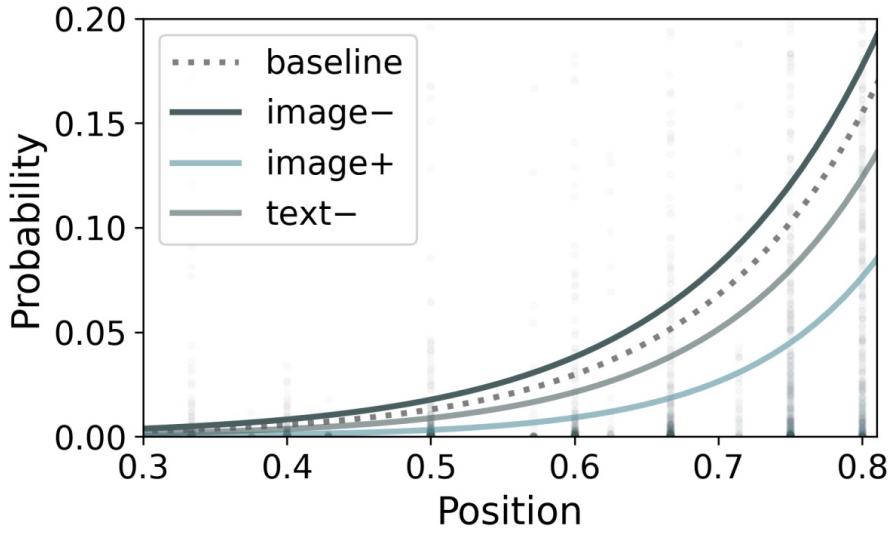
(1) 如何获取模型的视觉识别极限?



- 模型预测 EOS token 时，显著关注全部上下文，而非当前句子
- 猜想：模型可能在评估序列完整性，以决定是否生成 EOS

Application: Multimodal Hallucination

(1) 如何获取模型的视觉识别极限?



不同 condition 下，模型的 EOS 倾向

- 对图文上下文进行扰动 (e.g. 增加图像信息/遮挡文本信息)，模型 EOS 倾向随之改变
- 模型通过对比文本和图像来判断序列完整性，决定是否生成 EOS

Application: Multimodal Hallucination

(1) 如何获取模型的视觉识别极限?

- 模型自身具备根据视觉感知及时停止生成的潜力

  当模型倾向于停止，表明当前文本已充分描述模型能够感知的视觉内容

Application: Multimodal Hallucination

(1) 如何获取模型的视觉识别极限?

- 模型自身具备根据视觉感知及时停止生成的潜力

  当模型倾向于停止，表明当前文本已充分描述模型能够感知的视觉内容

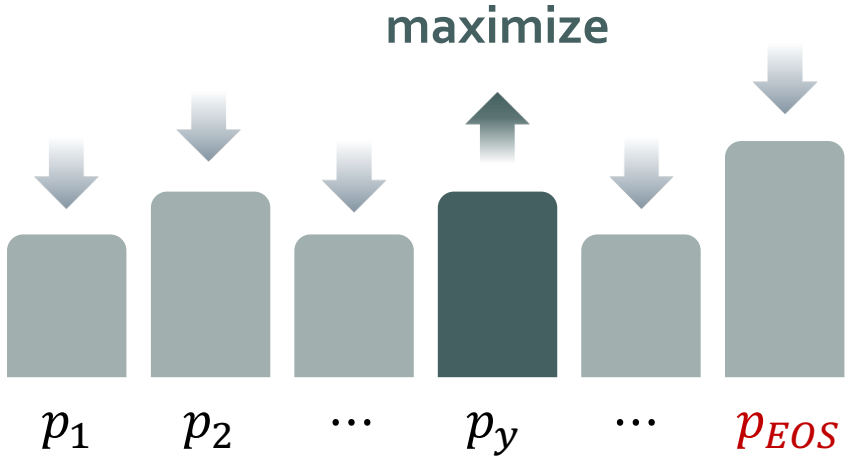
(2) 如何校准模型的 expressiveness?

- 部分词表学习

Application: Multimodal Hallucination

➢ 分析

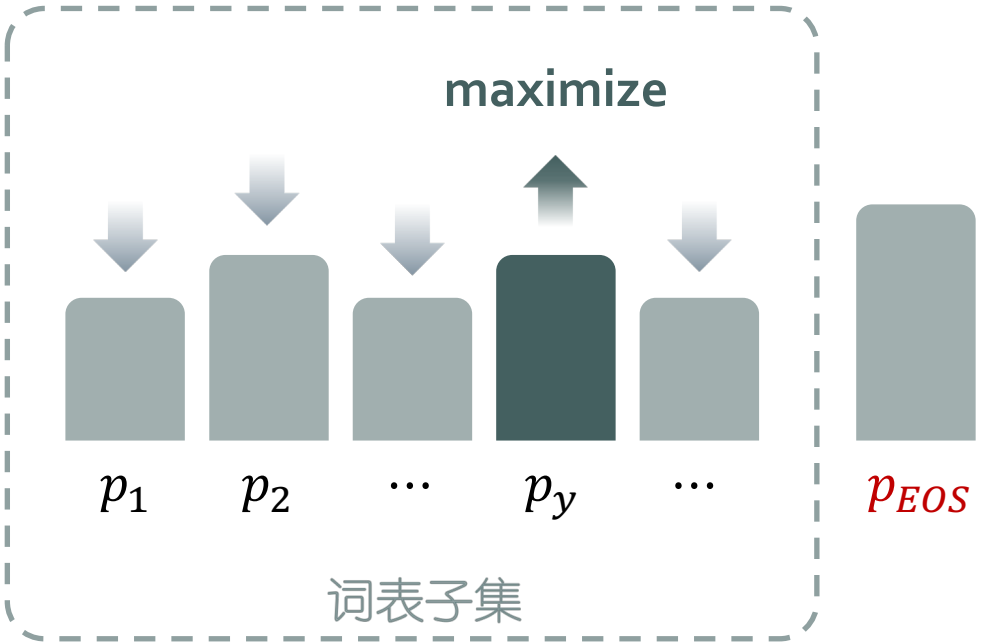
- 当模型倾向于预测 EOS，表明当前文本已完整描述视觉感知内容，此时应该停止生成
- 然而，若 label 不是 EOS，MLE 将导致 EOS 概率转移至 label
  - 破坏模型自身的 EOS 预测倾向，鼓励模型继续生成



Application: Multimodal Hallucination

➢ 分析

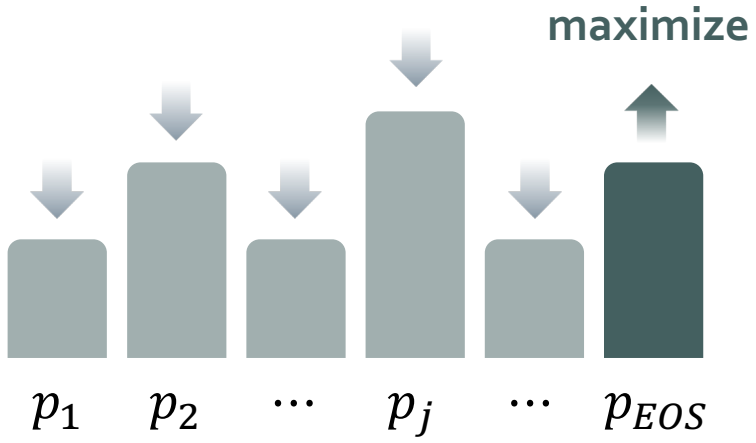
- 将 EOS 移出词表 (构建一个不包含 EOS 的子集)



模型预测 EOS 的倾向不会被惩罚

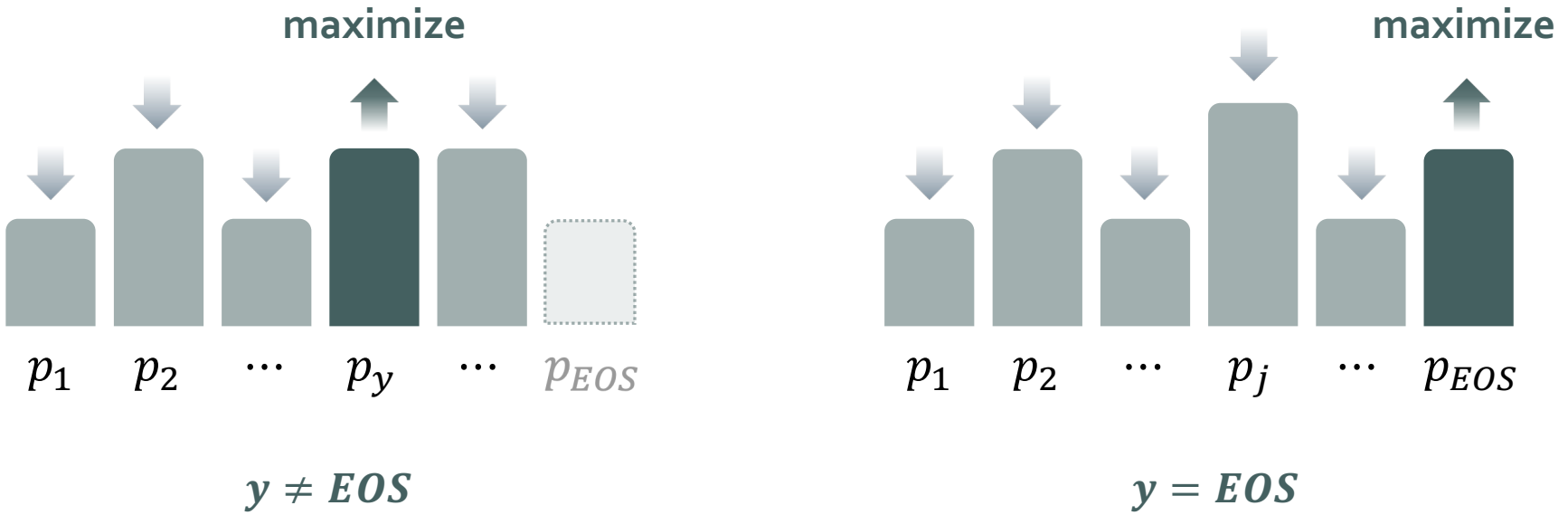
Application: Multimodal Hallucination

- 分析
  - 若 label 是 EOS (训练数据结束时)，模型倾向于继续生成，则应优化模型及时停止
  - 此时，应保持原始 MLE



Application: Multimodal Hallucination

- Selective EOS Supervision



Application: Multimodal Hallucination

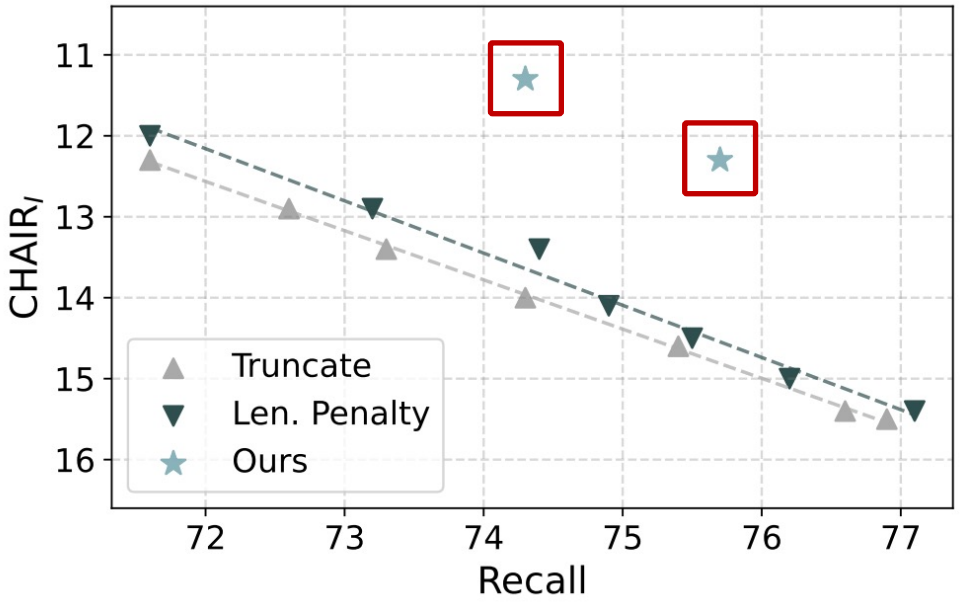
- 在普通 SFT 数据上对模型进行短时间微调
  - 显著减少 hallucination，同时维持 informativeness

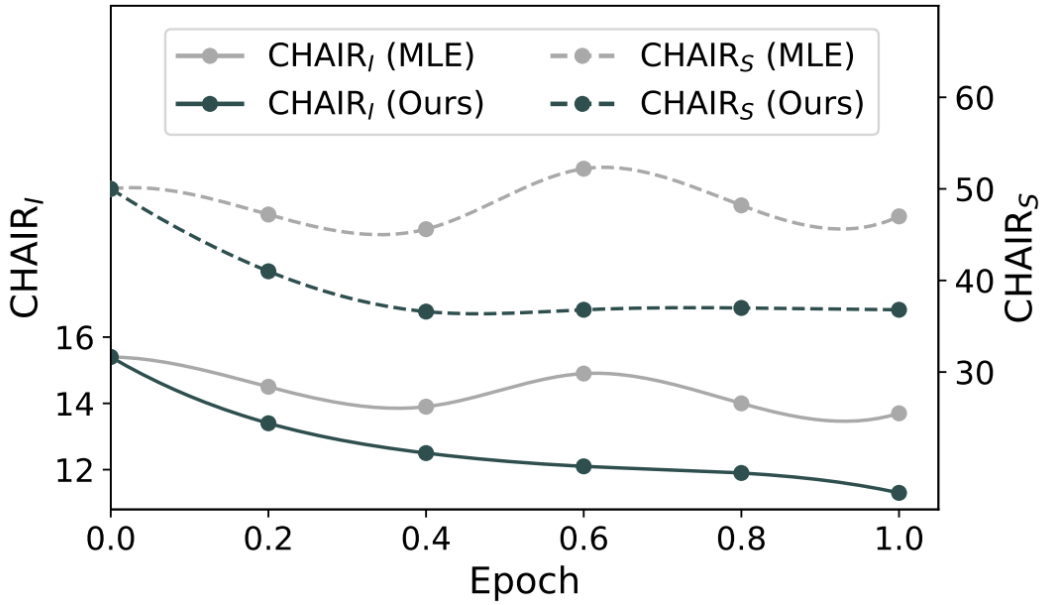
Row	Model	Method	Length	$CHAIR_S \downarrow$	$CHAIR_I \downarrow$	Recall ↑	Faith ↑	$Faith_S$ ↑
1	LLaVA-1.5 (7b)	-	100.6	50.0	15.4	77.1	87.0	68.8
2		VCD	100.4	48.6	14.9	**77.3**	87.1	70.2
3		OPERA	98.6	47.8	14.6	76.8	88.0	72.6
4		OPERA (fast)	85.3	48.6	14.5	76.7	87.7	71.3
5		**Ours (w/ Inst.)**	76.2	**36.8**	**11.3**	74.3	88.4	**73.0**
6		**Ours (w/ Cap.)**	79.7	40.2	12.3	75.7	**89.3**	72.3
7	LLaVA-1.5 (13b)	-	100.9	47.2	13.0	**77.3**	87.6	**73.1**
8		**Ours (w/ Cap.)**	85.1	**36.8**	**11.4**	75.3	**88.8**	72.8
9	LLaVA (7b)	-	57.8	35.4	13.8	**64.8**	86.9	67.4
10		**Ours (w/ Cap.)**	39.9	**27.0**	**13.2**	57.1	**88.9**	**71.6**
11	MiniGPTv2 (7b)	-	87.2	38.0	11.1	66.3	85.6	67.8
12		**Ours (w/ Cap.)**	62.2	**27.0**	**9.8**	**66.6**	**89.9**	**76.0**

hallucination　informativeness

Application: Multimodal Hallucination

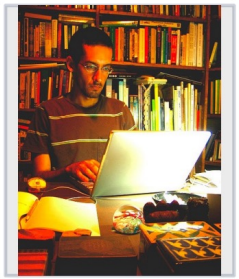
- 并不是简单地减少生成长度
  - 明显优于 truncating / length penalty
- 而是对模型 expressiveness 的校准
  - 模型幻觉率持续下降直到稳定





Application: Multimodal Hallucination

AI·M³
中国人民大学多媒体计算实验室

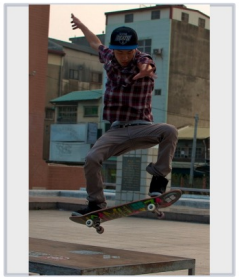


Baseline:
The image features a man sitting at a desk, working on a laptop computer. He is surrounded by numerous books, which are scattered around the desk and on the floor. Some books are placed close to the laptop, while others are located further away.
In addition to the laptop and books, there are a few other items on the desk, such as a mouse and a cell phone. A clock can also be seen on the wall, indicating the time. The scene suggests a focused and studious environment, with the man likely engaged in work or research.

Ours:
The image features a man sitting at a desk, working on a laptop computer. He is surrounded by numerous books, which are scattered across the desk and on the shelves behind him. The books vary in size and are placed in different orientations, creating a busy and studious atmosphere.

Baseline:
The image features a person wearing a wetsuit, standing in the water and holding a sailboard. The person is positioned in the middle of the scene, with the sailboard prominently visible in front of them. The water appears to be choppy, adding to the dynamic nature of the scene.
In the background, there are several boats of varying sizes scattered across the water, creating a lively atmosphere. Some of the boats are closer to the person with the sailboard, while others are further away, providing a sense of depth to the image.

Ours:
The image features a person wearing a wetsuit, standing in the water and holding a sail. The person appears to be preparing to windsurf or engage in a water sport that involves a sail. The sail is large and positioned in front of the person, with the water surrounding them.



Baseline:
The image captures a young man skillfully performing a trick on a skateboard, jumping into the air while riding it. He is wearing a plaid shirt and appears to be in a park-like setting. The skateboard is positioned underneath him, with the wheels visible as he jumps.
In the background, there are a few other people present, possibly watching the skateboarder or engaging in their own activities. The scene also features a bench, a chair, and a potted plant, adding to the park-like atmosphere.

Ours:
The image captures a young man skillfully performing a trick on his skateboard, jumping into the air with the skateboard beneath his feet. He is wearing a plaid shirt and appears to be in a park-like setting. The skateboarder is the main focus of the scene, with his impressive stunt drawing attention.

Baseline:
The image features a train station with a train on the tracks. A man is standing next to the train, holding a green trash can, likely preparing to clean the train. Another man is standing in the train, possibly a passenger or a worker.
There are several other people in the scene, some of them standing near the train, while others are scattered around the station. A handbag can be seen placed on the ground, possibly belonging to one of the people in the scene.

Ours:
The image features a train station with a train on the tracks. A man in a blue shirt is standing next to the train, holding a green trash can. He appears to be in the process of unloading the trash can from the train. Another man is standing in the open door of the train, possibly waiting to get off or assisting with the unloading process.

Baseline:
The image features a wooden dining table with a tray holding three delicious donuts. The donuts are arranged in a visually appealing manner, with one donut placed in the center and the other two positioned on either side. The table appears to be in a restaurant setting, as there is a bench situated nearby.
In addition to the donuts, there is a cup placed on the table, possibly containing a beverage to accompany the sweet treats. The overall scene conveys a pleasant and inviting atmosphere, perfect for enjoying a snack or dessert.

Ours:
The image features a wooden dining table with three donuts placed on a napkin. The donuts are of different sizes and are arranged in a way that they are stacked on top of each other. The table is positioned near a window, allowing natural light to illuminate the scene.

Baseline:
The image captures a man skillfully riding a wave on a surfboard in the ocean. He is wearing a wetsuit and appears to be enjoying the thrill of the sport. The surfer is positioned in the center of the scene, with the surfboard beneath him.
In the background, there are several cars parked along the shore, likely belonging to beachgoers or spectators. The cars are scattered around the area, with some closer to the water and others further back. The scene is a lively representation of a day at the beach, with people enjoying various activities.

Ours:
The image captures a man in a black wetsuit skillfully riding a wave on a surfboard. He is positioned in the center of the scene, with the surfboard beneath him as he expertly navigates the water. The wave he is riding is quite large, providing an exciting challenge for the surfer.

Application: Multimodal Hallucination

- 部分词表学习：多模态幻觉

发现问题

冗长的训练数据使得模型倾向于过度表达

总结规律

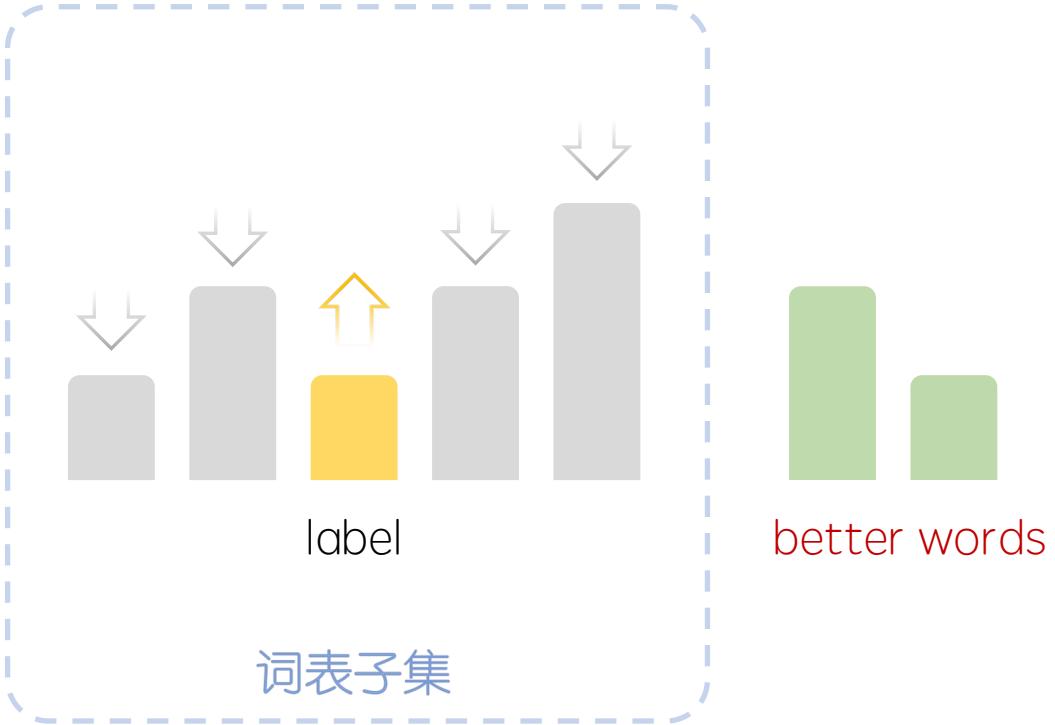
EOS token 到 其它单词 的“有害”概率转移

构建子集

排除 EOS token

Discussion

- 部分词表学习： Supervision Selection

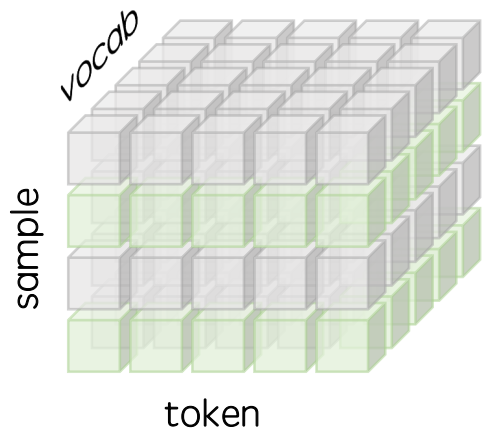


保留子集内单词到 label 的概率转移

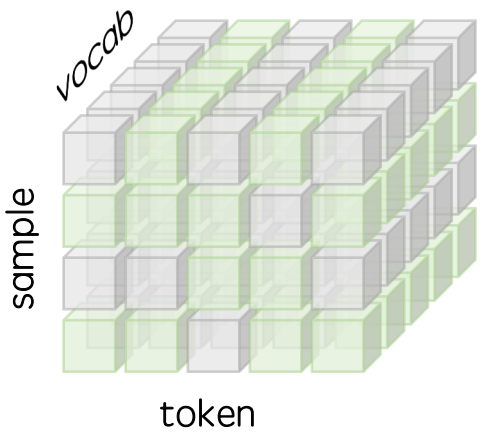
避免子集外单词到 label 的概率转移

Discussion

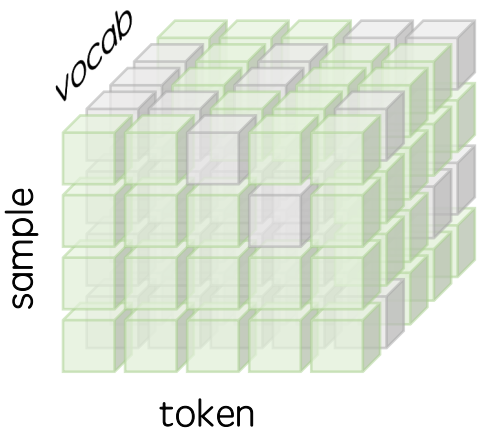
- data selection 视角
  - 选择性保留 vocabulary 内的概率转移



sample (sequence) –level



token–level

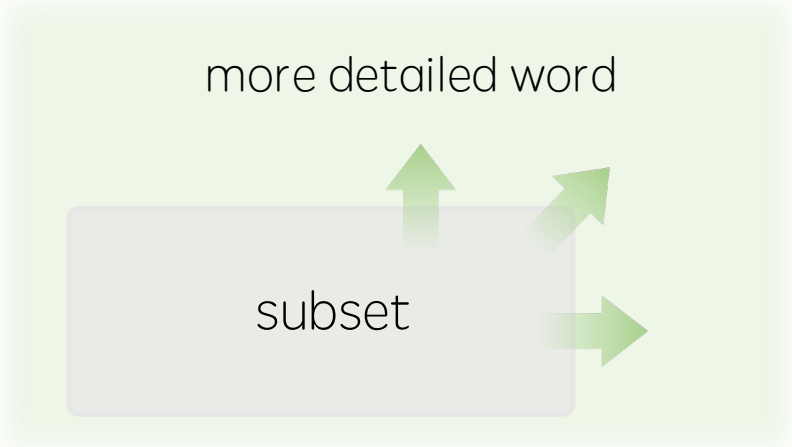


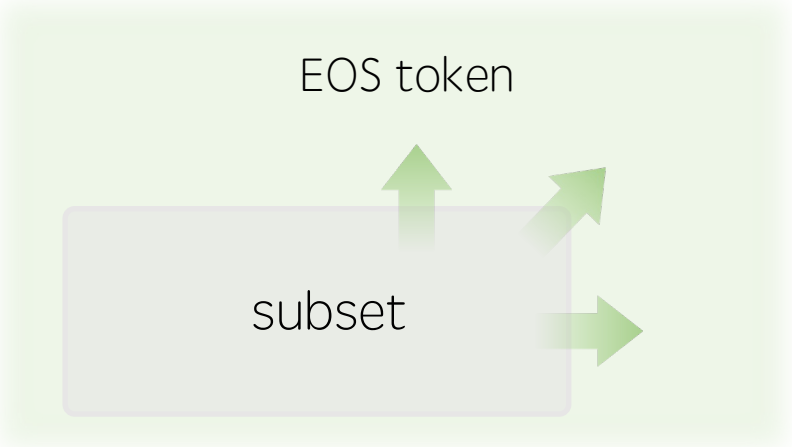
vocabulary–level

更细粒度的控制和更大的灵活性

Discussion

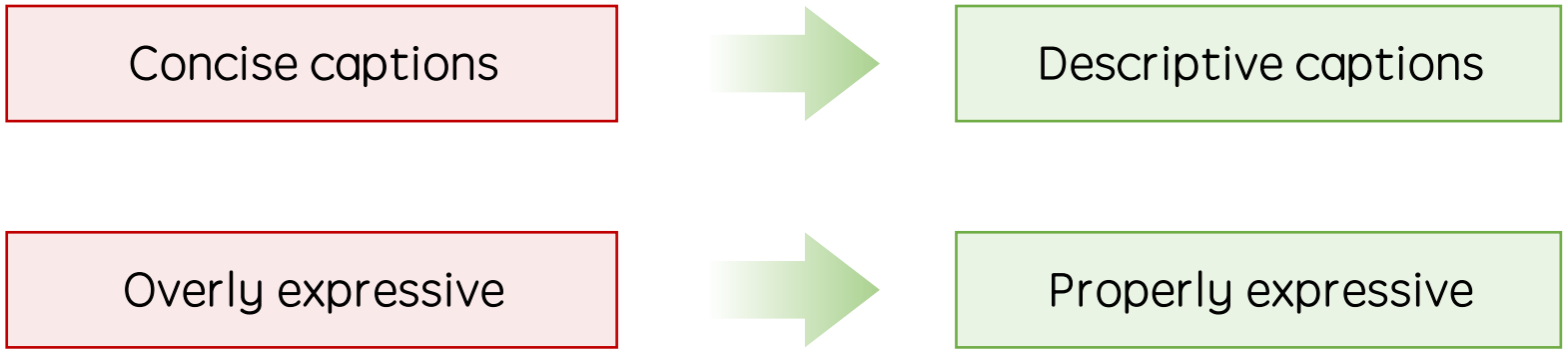
- model learning 视角
  - 子集之外，提供一个 supervision-free 的自由区域，鼓励模型保持目标行为
  - 单向性：允许模型目标行为“只增不减”





Discussion

- distribution modeling 视角
  - 基于已有训练数据分布，学习一个更符合预期的“更优分布”



Future Works

- From Specific to General-purpose Supervision Selection
  - e.g. LLM pretrain
- From Heuristical to Automatic Subsetting Strategy
  - e.g. learning to select subsets (by model)
- From Text to Multimodality Generation

References

- Learning Descriptive Image Captioning via Semipermeable Maximum Likelihood Estimation (NeurIPS 2023)
- Less is More: Mitigating Multimodal Hallucination from an EOS Decision Perspective (ACL 2024)
- More exploration in progress!