ClipCap: CLIP Prefix for Image Captioning

Подготовили Горкунов Николай, Соколов Илья, Иванченко Максим
Группа ML-21

Анализ данных

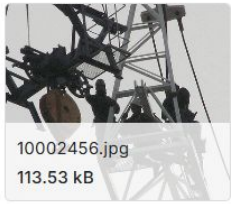
flickr30k_images (31.8k files)

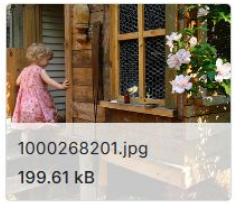
About this directory

Add Suggestion

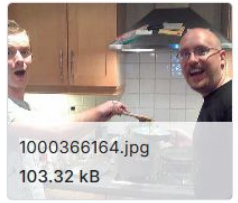
Image directory

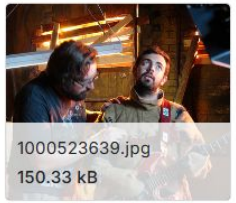

1000092795.jpg
218.14 kB


10002456.jpg
113.53 kB


1000268201.jpg
199.61 kB


1000344755.jpg
154.01 kB


1000366164.jpg
103.32 kB


1000523639.jpg
150.33 kB


1000919630.jpg
117.18 kB


10010052.jpg
44.51 kB


1001465944.jpg
141.08 kB

captions.txt (13.16 MB)

About this file

Add Suggestion

5 captions provided by human annotators for each image

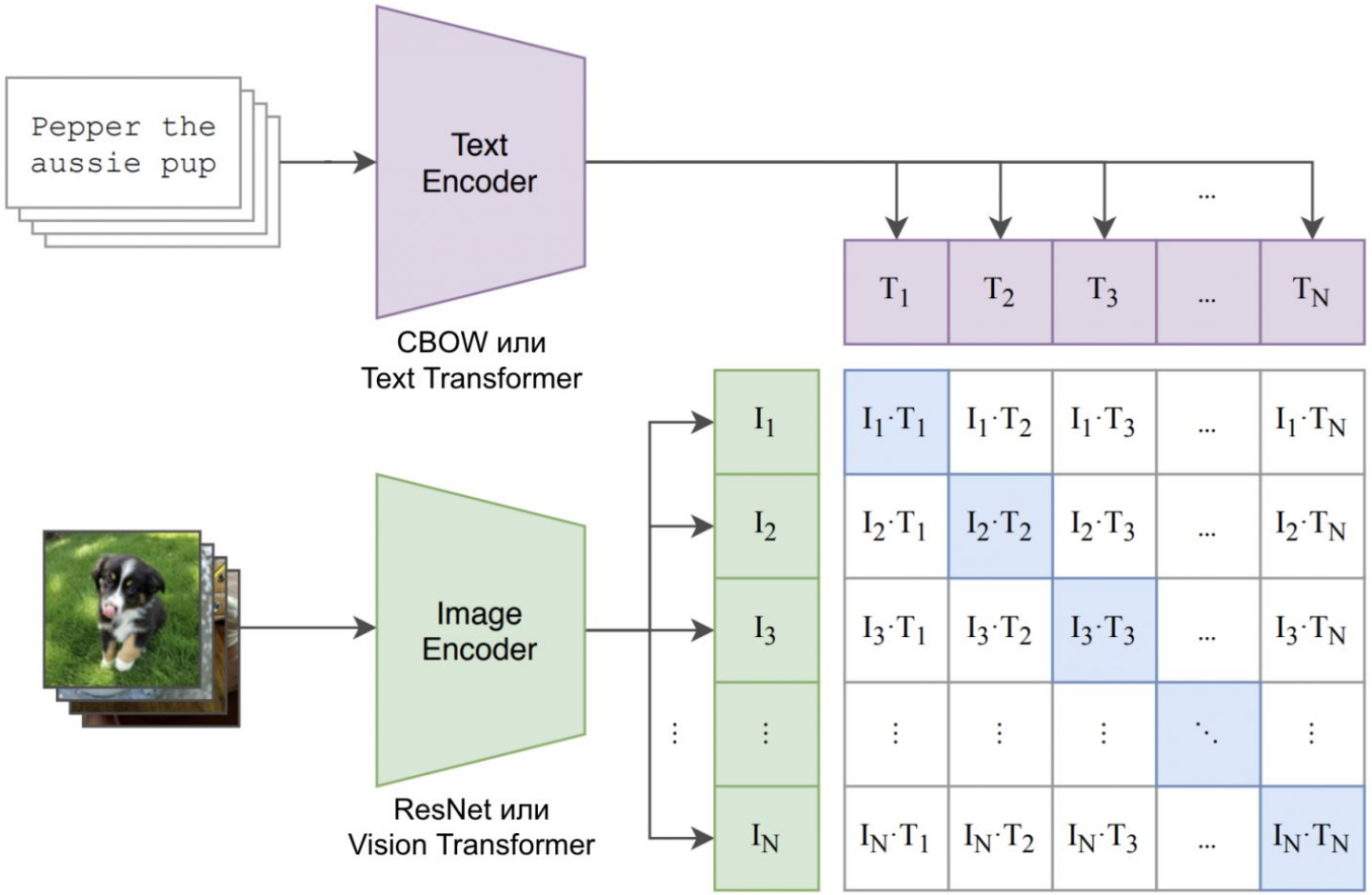
This preview is truncated due to the large file size. Create a Notebook or download this file to see the full content.

Download

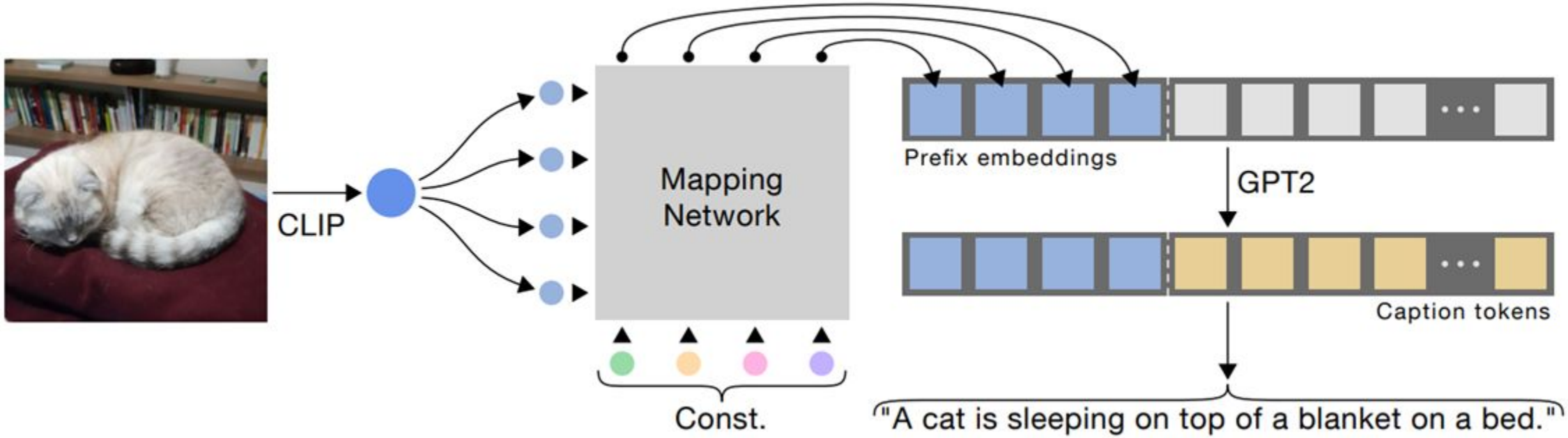
Create Notebook

```
image_name,comment_number,comment
1000092795.jpg,0,Two young guys with shaggy hair look at their hands while hanging out in the yard .
1000092795.jpg,1,Two young  White males are outside near many bushes .
1000092795.jpg,2,Two men in green shirts are standing in a yard .
1000092795.jpg,3,A man in a blue shirt standing in a garden .
1000092795.jpg,4,Two friends enjoy time spent together .
```

Принцип работы CLIP

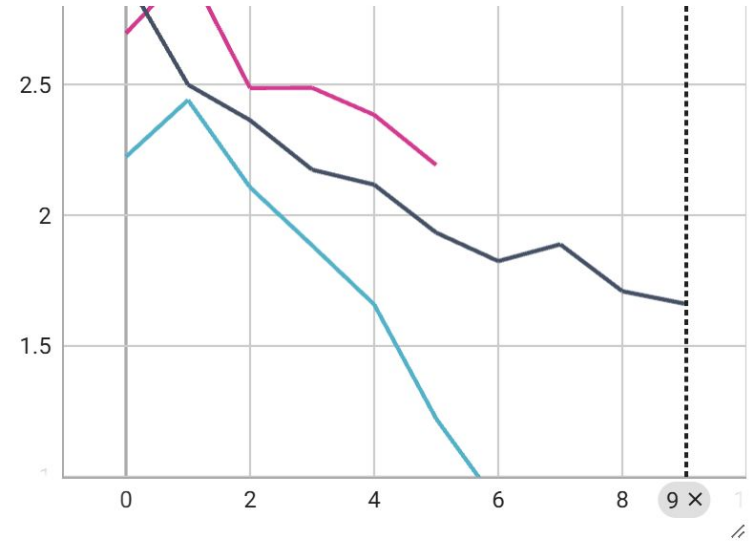


Архитектура ClipCap



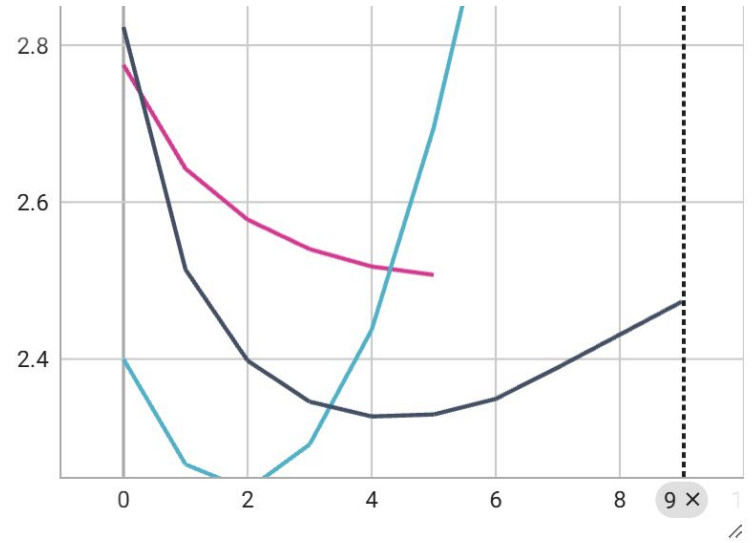
Логирование. Tensorboard

Loss/train



Run ↑	Value	Step	Relative
mapper-only	2,1921	5	3.31 hr
pref40batch15	0,7033	7	6.049 hr
pref5batch60	1,6623	9	4.24 hr

Loss/val



Run ↑	Value	Step	Relative
mapper-only	2,5074	5	3.31 hr
pref40batch15	3,3436	7	6.049 hr
pref5batch60	2,4733	9	4.24 hr

В чем были сложности

- Высокие требования к объему видеопамяти
- Недостаточный объем памяти HDD на Kaggle
- Переобучение

Демонстрация

[9]:
```python
# Смотрим картинку
img_url = 'https://www.brunetterunning.com/wp/wp-content/uploads/2018/09/nnrunningclub1-300x200.jpg'
images = Image.open(requests.get(img_url, stream=True).raw).convert("RGB")
display(images)
```

+ Code + Markdown

[10]:
```python
image = preprocess(images).unsqueeze(0).to(device)

with torch.no_grad():
    prefix = clip_model.encode_image(image).to(device, dtype=torch.float32)
    # prefix = prefix / prefix.norm(2, -1).item()
    prefix_embed = model.clip_project(prefix).reshape(1, prefix_length, -1)

generated_text_prefix = generate_beam(model, tokenizer, beam_size=7,embed=prefix_embed,stop_token='<|endoftext|>',entry_length=30)[0]

print(generated_text_prefix)
```

```
A group of people are jogging on a grassy area with trees in the background. A man in a black shirt is jogging with a woman
```

Результаты

	CIDEr	SPICE
ClipCap gpt2 + MLP batch 60 prefix 5	1.2224	0.2796
ClipCap gpt2 + MLP batch 10 prefix 40	**1.3029**	**0.3134**

Заключение

Идеи для улучшения:

- Увеличение объема входных данных и вычислительных мощностей.
- Использование более сложных моделей для предобработки визуальных данных.
- Использование LLM большего размера.

Вопросы