Сравнение архитектур для генерации описаний изображений

Кочетков Александр
Хорошенко Дмитрий

Цель работы

Целью работы является разработка, реализация и сравнительный анализ архитектур нейронных сетей для автоматической генерации описаний изображений.

Датасет

Для нашей задачи был выбран корпус Flickr8k.

Количество изображений: 8 092

Подписей на изображение: 5 (всего ~40 460 описаний)

Тематика: повседневные сцены, люди, животные, действия, природа

Стандартные разбиения:

- train ≈ 6 000 изображений
- dev / val ≈ 1 000
- test ≈ 1 000

- A black dog and a spotted dog are fighting
- A black dog and a tri-colored dog playing with each other on the road .
- A black dog and a white dog with brown spots are staring at each other in the street .
- Two dogs of different breeds looking at each other on the road .
- Two dogs on pavement moving toward each other

Обзор архитектур

- Классика (2015–2018): CNN (ResNet/VGG) + RNN/LSTM
- Современный стандарт (2018–2025): Transformer-архитектуры (ViT / CLIP + Transformer Decoder)
- Тренд 2025: MLLM (LLaVA, Qwen-VL, Molmo и др.) — end-to-end, zero-shot

Наш выбор – первые два пункта, так как дообучение больших LLM требует много ресурсов.

Выбранные архитектуры

ResNet50 + Transformer

- CNN → Adaptive Pool → Linear Proj → PosEnc → Transformer Encoder
- Decoder: Transformer Decoder

ResNet50 + LSTM + Attention

- CNN → Adaptive Pool → Conv Proj → PosEnc → Transformer Encoder
- Decoder: LSTM + Attention

CLIP ViT-B/32 + Transformer

- CLIP Vision → Linear Proj → PosEnc → Transformer Encoder
- Decoder: Transformer Decoder

Метрики оценки

- BLEU-4 — точность n-грамм (n=4), строго к порядку
- METEOR — учитывает синонимы и стемминг, лучше корреляция с человеком
- ROUGE-L — longest common subsequence, хорошо для структуры
- CIDEr — consensus-based, **лучше всего коррелирует с человеческими оценками** в image captioning

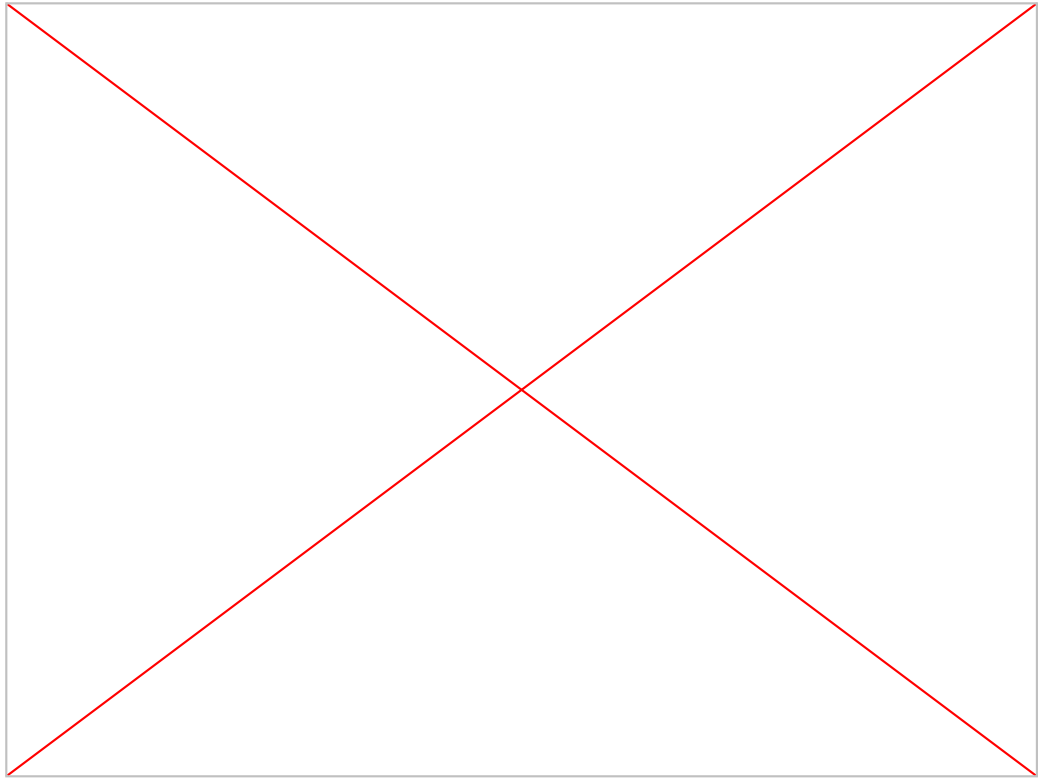
Графики обучения

- resnet-lstm
- resnet-transformer
- clip-transformer

Результаты

Модель	BLEU-4	METEOR	ROUGE-L	CIDEr
ResNet + LSTM	0.194	0.216	0.461	0.516
ResNet + Transformer	0.229	0.232	0.477	0.613
CLIP + Transformer	0.253	0.245	0.501	0.684

Демонстрация



Итоги и планы

CLIP + Transformer — оптимальный выбор для Flickr8k
Дальше:

- fine-tune CLIP
- попробовать современные MLLM (Qwen2-VL, LLaVA и др.)
- оценить на более сложных датасетах (COCO, nocaps)

Спасибо за внаимание!