

Применение синтетических данных, полученных с помощью генеративной нейросети, для повышения качества моделей детекции

Степанов Илья

Московский физико-технический институт

Курс: НИР

Научный руководитель: Грабовой Андрей Валерьевич

Консультант: Филатов Андрей Викторович

2024

Цель исследования

Задача

Создание высококачественных аугментаций с использованием генеративной нейросети для повышения качества моделей детекции.

Проблема

Существующие методы аугментации с применением генеративных нейросетей обладают рядом недостатков, таких как: невозможность генерировать новые классы объектов; отсутствие физичности у синтетических изображений.

Цель

Создание автоматизированного pipeline, способного качественно генерировать аугментации, нивелируя проблемы предыдущих подходов. Проведение сравнений аугментаций на датасетах COCO и Pascal VOC с использованием моделей детекции, а также проведение ablation.

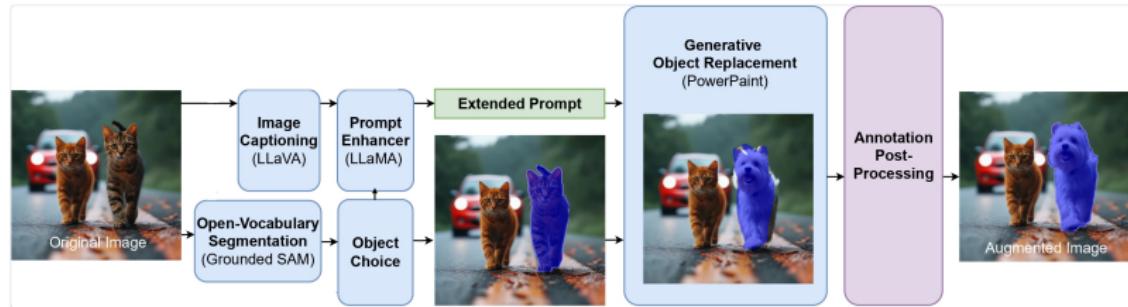
Постановка задачи

Определим датасет как $\mathfrak{D} = \{x_i : i = 1, \dots, n\}$, x_i — изображение.
Рассматривается диффузионная модель для аугментации ϵ_θ .
Определим функцию потерь:

$$\mathcal{L}(\epsilon, \epsilon_\theta) = \mathbb{E}_{\epsilon \sim N(0, I), m_i, \tau_i, \tilde{x}_i, t} \|\epsilon - \epsilon_\theta(x_i^t, m_i, \tau_i, \tilde{x}_i, t)\|^2,$$

где x_i^t — зашумленное изображение на шаге t , m_i — сегментационная маска данного изображения, τ_i — текстовая подсказка данного изображения, $\tilde{x}_i = (1 - m_i) \odot x_i$, $t \in [0, T]$ — шаг диффузионного процесса. Для получения масок и текстовых подсказок мы используем модели сегментации и "image-to-text".

Архитектура модели



На рисунке изображён автоматический pipeline. Модель принимает изображение, после чего процесс аугментации разделён на следующие этапы:

- ▶ Подготовка к аугментации начинается с генерации описания изображения с помощью модели LLaVA. Далее модель SAM генерирует маску для случайного объекта, после чего модель LLaMA использует эту маску и описание изображения для создания расширенной подсказки.
- ▶ Расширенная подсказка, маска и исходное изображение подаются в модель PowerPoint, которая генерирует аугментацию.
- ▶ На этапе постобработки проводится фильтрация с использованием AlphaCLIP, а также уточняется маска сгенерированного объекта при помощи SAM.

Эксперименты

Датасет	Модель	Данные	0%	25%	50%	75%	100%
Pascal VOC	DETR	оригинальные	0.0	57.5	61.5	65.4	69.1
		наши	5.3	55.6	62.2	66.3	70.3
	YOLOv10-N	оригинальные	0.0	50.9	54.8	56.9	60.4
	Faster RCNN	наши	31.5	51.3	53.6	57.9	61.3
		оригинальные	0.0	74.5	77.4	75.5	76.6
		наши	66.4	77.3	80.1	83.5	84.0

В таблице записаны результаты детекции объектов на наборе данных Pascal VOC. Обучение на наших данных значительно улучшает производительность моделей детекции, что подтверждается более высокими значениями AP в строках "наши" по сравнению с "оригинальные". Проценты представляют собой долю изображений с кошками, использованных из оригинального набора данных.

Эксперименты

AP on cat category	
w/o expanded prompts	expanded prompts
64.6	66.4



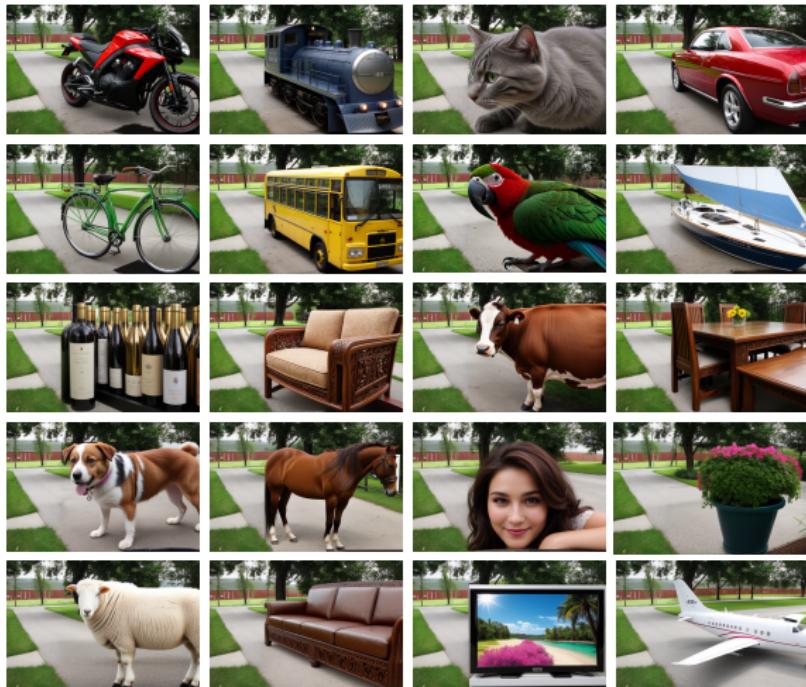
Prompt: Cat



Prompt: The gray tabby cat has distinctive stripes on its fur, large and expressive golden eyes.

В таблице записаны результаты сравнения качества обучения модели FasterRCNN в зависимости от расширения подсказки. Результаты показывают, что использование расширенных подсказок значительно повышает AP для категории "кошка" по сравнению с использованием обычных инструкций. На картинке продемонстрирована генерация с помощью разных подсказок.

Эксперименты



На картинке продемонстрированы результаты генераций с помощью нашего подхода.

План работы на следующий семестр

- ▶ Собрать автоматический pipeline
- ▶ Сделать сравнение с существующими методами на датасетах Coco и Pascal VOC
- ▶ Провести ablation модели