

Нейронные сети

Семинар 06. Аугментация данных

Иванов И.Е.

MaTIS

22 ноября 2024 г.



- 1 Выдача домашнего задания
- 2 Аугментация данных

Трудности разметки больших датасетов

- Собрать представительный датасет картинок даже без разметки в данном предметном домене — не очень простое дело (например, из-за правовых вопросов)



Трудности разметки больших датасетов

- Собрать представительный датасет картинок даже без разметки в данном предметном домене — не очень простое дело (например, из-за правовых вопросов)
- Для разметки одного bounding box-а требуется 2 клика, для разметки семантической маски — уже гораздо больше



Трудности разметки больших датасетов

- Собрать представительный датасет картинок даже без разметки в данном предметном домене — не очень простое дело (например, из-за правовых вопросов)
- Для разметки одного bounding box-а требуется 2 клика, для разметки семантической маски — уже гораздо больше
- Важно не просто найти людей, которые будут размечать данные, но и следить за качеством разметки



Трудности разметки больших датасетов

- Собрать представительный датасет картинок даже без разметки в данном предметном домене — не очень простое дело (например, из-за правовых вопросов)
- Для разметки одного bounding box-а требуется 2 клика, для разметки семантической маски — уже гораздо больше
- Важно не просто найти людей, которые будут размечать данные, но и следить за качеством разметки
- Не раз сталкивался с плохо размеченными данными в своей практике, что обычно приводит к проблемам при обучении модели и/или тестировании



Трудности разметки больших датасетов

- Собрать представительный датасет картинок даже без разметки в данном предметном домене — не очень простое дело (например, из-за правовых вопросов)
- Для разметки одного bounding box-а требуется 2 клика, для разметки семантической маски — уже гораздо больше
- Важно не просто найти людей, которые будут размечать данные, но и следить за качеством разметки
- Не раз сталкивался с плохо размеченными данными в своей практике, что обычно приводит к проблемам при обучении модели и/или тестировании
- Иногда для сбора данных нужно дорогое оборудование (например, снимки МРТ), а для разметки квалифицированные эксперты



Данные — это дорого

Трудности разметки больших датасетов

- Собрать представительный датасет картинок даже без разметки в данном предметном домене — не очень простое дело (например, из-за правовых вопросов)
- Для разметки одного bounding box-а требуется 2 клика, для разметки семантической маски — уже гораздо больше
- Важно не просто найти людей, которые будут размечать данные, но и следить за качеством разметки
- Не раз сталкивался с плохо размеченными данными в своей практике, что обычно приводит к проблемам при обучении модели и/или тестировании
- Иногда для сбора данных нужно дорогое оборудование (например, снимки МРТ), а для разметки квалифицированные эксперты

Вывод

Собрать большой датасет могут позволить себе только большие компании

Зачем компании собирают большие датасеты

Данные побеждают алгоритмы

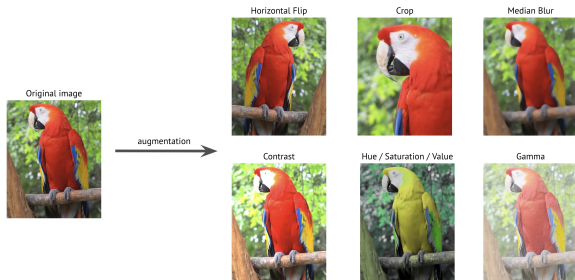
- Для достижения лучшего качества современные датасеты могут содержать миллионы изображений. Например, датасет JFT содержит 303 миллионов изображений и разметку на 18 тысяч классов
- Чем более разнообразнее и представительнее датасет, тем лучше будет работать модель компьютерного зрения
- Как правило легче улучшить качество текущего решения добавив данные, чем используя более продвинутые модели и способы обучения (на практике используют оба подхода)



Аугментация данных (data augmentation)

Определение

Аугментация данных — это процесс создания новых экземпляров данных из уже имеющихся. При этом разметка новых данных получается из уже имеющейся разметки.



Влияние аугментации данных на качество модели

- Все SOTA модели используют аугментации
- Было замечено, что правильно подобранная аугментация данных, может существенно улучшить итоговое качество модели
- Аугментация данных — один из способов борьбы с переобучением
- Использование аугментации данных относится к лучшим практикам компьютерного зрения

Model	Base augmentations	AutoAugment augmentations
ResNet-50	76.3	77.6
ResNet-200	78.5	80.0
AmoebaNet-B (6,190)	82.2	82.8
AmoebaNet-C (6,228)	83.1	83.5







Виды аугментаций данных: стандартные трансформации

- Вырезание кropa (случайный крoп, центральный крoп и т.д.)
- Отражение и повороты на 90
- Поворот на случайный угол
- Добавление смаза (blur)
- Добавление шума
- Изменение яркости / контрастности
- ...



Виды агментаций для задачи классификации¹

	ResNet-50	Mixup	Cutout	CutMix
Image				
Label	Dog 1.0	Dog 0.5 Cat 0.5	Dog 1.0	Dog 0.6 Cat 0.4
ImageNet Cls (%)	76.3 (+0.0)	77.4 (+1.1)	77.1 (+0.8)	78.4 (+2.1)
ImageNet Loc (%)	46.3 (+0.0)	45.8 (-0.5)	46.7 (+0.4)	47.3 (+1.0)
Pascal VOC Det (mAP)	75.6 (+0.0)	73.9 (-1.7)	75.1 (-0.5)	76.7 (+1.1)

¹<https://github.com/clovaai/CutMix-PyTorch>

Аугментации для задач обнаружения объектов и сегментации ²

Основная идея

Объекты можно вырезать из одних изображений и вставлять в другие



²<https://arxiv.org/pdf/2012.07177.pdf>

Идея

Иногда данные можно сгенерировать автоматически

Идея

Иногда данные можно сгенерировать автоматически

Пример

Автомобильные номера

Идея

Иногда данные можно сгенерировать автоматически

Пример

Автомобильные номера

Другой пример

Существует отдельный класс моделей, которые могут генерировать изображения, но есть нюансы.



Аугментации при тестировании (Test Time Augmentation)

Идея

Аугментирование при тестировании может улучшить итоговое качество модели

albumentations

- Удобный инструмент для реализации аугментации данных
- Поддерживает различные задачи в различных доменах
- Интеграция с Keras и PyTorch
- Open source решение

³<https://albumentations.ai/>



Спасибо за внимание!

