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

Степанов И.Д., Дорин Д.Д., Игнашин И.Н., Изместьева У.А.

Московский физико-технический институт

Научный руководитель: к.ф-м.н. Грабовой Андрей Валерьевич

Консультант: Филатов Андрей Викторович

2025

Цель исследования

Задача

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

Проблема

Существующие методы генеративной аугментации обладают рядом недостатков, таких как: невозможность генерировать новые классы объектов, качество аугментаций низкое.

Цель

Создание автоматизированной модели, способной качественно генерировать аугментации и устранять недостатки существующих подходов. Проведение сравнительного анализа аугментаций на датасетах COCO и Pascal VOC с использованием модели детекции, а также анализ влияния компонентов предложенного метода на конечный результат.

Постановка задачи

Рассмотрим модель детекции f_θ как отображение:

$$f_\theta : X \to \hat{T},$$

где X — пространство изображений, $\hat{T}$ — пространство соответствующих разметок, содержащих координаты ограничивающих рамок и классы объектов, предсказанных моделью. Также определим T — истинное пространство разметок изображений из X.

Постановка задачи

Определим функцию потерь:

$$\mathcal{L}(\theta) = \lambda_{\text{coord}} \sum_{i=1}^{S^2} \sum_{j=1}^{B} \mathbb{I}_{ij}^{\text{obj}} \left[(x_i^{gt} - \hat{x}_i)^2 + (y_i^{gt} - \hat{y}_i)^2 \right]$$

$$+ \lambda_{\text{coord}} \sum_{i=1}^{S^2} \sum_{j=1}^{B} \mathbb{I}_{ij}^{\text{obj}} \left[(\sqrt{w_i^{gt}} - \sqrt{\hat{w}_i})^2 + (\sqrt{h_i^{gt}} - \sqrt{\hat{h}_i})^2 \right]$$

$$+ \sum_{i=1}^{S^2} \sum_{j=1}^{B} \mathbb{I}_{ij}^{\text{obj}} (\hat{C}_i - C_i)^2 + \lambda_{\text{noobj}} \sum_{i=1}^{S^2} \sum_{j=1}^{B} \mathbb{I}_{ij}^{\text{noobj}} (\hat{C}_i - C_i)^2$$

$$+ \sum_{i=1}^{S^2} \mathbb{I}_{i}^{\text{obj}} \sum_{c \in \mathcal{C}} (\hat{p}_i(c) - p_i(c))^2,$$

где $S \times S$ — размер сетки, на которую разбивается изображение,

Постановка задачи

B — количество предсказанных ограничивающих рамок (bounding boxes) в каждой ячейке сетки, $\lambda_{\text{coord}}, \lambda_{\text{noobj}}$ — коэффициенты, регулирующие вклад в функцию потерь, $\mathbb{I}_{ij}^{\text{obj}}$ — индикатор наличия объекта в j-й рамке i-й ячейки, $\mathbb{I}_{ij}^{\text{noobj}}$ — индикатор отсутствия объекта в j-й рамке i-й ячейки, $(x_i^{gt}, y_i^{gt}, w_i^{gt}, h_i^{gt})$ — координаты центра, ширина и высота истинного ограничивающего прямоугольника, $(\hat{x}_i, \hat{y}_i, \hat{w}_i, \hat{h}_i)$ — предсказанные координаты ограничивающего прямоугольника, C_i и $\hat{C}_i$ — истинная и предсказанная вероятность наличия объекта в ячейке, $\mathcal{C}$ — множество классов объектов, $p_i(c)$ и $\hat{p}_i(c)$ — истинная и предсказанная вероятность принадлежности объекта классу c.

Решается следующая оптимизационная задача:

$$\theta^* = \arg\min_{\theta} \mathcal{L}(\theta)$$

Функция качества

Рассмотрим функцию качества для задачи детекции:

$$\text{mAP} : \hat{T} \times T \times [0,1] \to [0,1]$$

Для каждого класса $c \in \mathcal{C}$ вычисляется Average Precision (AP):

$$AP(c,\tau) = \int_0^1 P_c(r,\tau)\, dr,$$

где $P_c(r,\tau)$ — функция точности при полноте r для класса c, $\tau \in [0,1]$.

$$\text{mAP} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} AP(c,\tau).$$

Функция качества

Рассмотрим функцию IoU (Intersection over Union):

$$\text{IoU} : \hat{T} \times T \to [0, 1],$$

которая рассчитывается по формуле:

$$\text{IoU} = \frac{|B_p \cap B_{gt}|}{|B_p \cup B_{gt}|},$$

где B_p — предсказанный ограничивающий прямоугольник (bounding box), B_{gt} — истинный ограничивающий прямоугольник. Рассмотрим функцию следующего вида:

$$\text{mAP}^* : \hat{T} \times T \to [0, 1]$$

$$\text{mAP}^* = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \left(\frac{1}{|\mathcal{T}|} \sum_{\tau \in \mathcal{T}} AP(c, \tau) \right), \quad \text{где } \mathcal{T} = \{0.50, 0.55, \ldots, 0.95\}$$

Генеративная аугментация

Рассмотрим модель генеративной аугментации как композицию отображений:

$$r_\gamma \circ h_\beta \circ g_\alpha \circ f_\psi : X \times [0,1] \to Y \cup \varnothing$$

$$f_\psi : X \times [0,1] \to M \times L \times [0,1]$$

$$g_\alpha : X \times L \to P$$

$$h_\beta : X \times M \times P \to Y$$

$$r_\gamma : Y \times M \times L \times [0,1] \to Y \cup \varnothing,$$

X — пространство исходных изображений, Y — пространство аугментированных изображений,

Генеративная аугментация

f_ψ — модель детекции объекта, который будем аугментировать, g_α — модель расширения текстового запроса для аугментации нового объекта, h_β — модель генерации нового объекта, r_γ — модель фильтрации некачественных генераций, где M — пространство бинарных масок объектов исходных изображений, L — пространство классов объектов исходных изображений, P — пространство расширенных текстовых запросов для аугментации объекта.

Генеративная аугментация

Утверждение 1:
Пусть $\mathcal{D}_{\text{orig}} = \{(x_i, t_i)\}_{i=1}^{N}$ — исходный датасет, где $x_i \in X$ — исходные изображения, $t_i \in T$ — соответствующие разметки, $\mathcal{D}_{\text{orig}} = \mathcal{D}_{\text{val}} \cup \mathcal{D}_{\text{train}}$. Пусть $\mathcal{D}_{\text{aug}} = \{(x_i^{\text{aug}}, t_i^{\text{aug}})\}_{i=1}^{M}$ — аугментированный датасет, где $x_i^{\text{aug}} \in Y$ — аугментированные изображения, $t_i^{\text{aug}} \in T$ — соответствующие разметки.
Рассмотрим дивергенцию Кульбака-Лейблера между распределениями аугментированных и исходных данных

$$D_{\text{KL}}(\{x \mid (x,t) \in \mathcal{D}_{\text{orig}}\} \parallel \{x \mid (x,t) \in \mathcal{D}_{\text{aug}}\}) \approx 0$$

рассмотрим модель детекции f_θ, обученную на $\mathcal{D}_{\text{train}}$, также рассмотрим f_ϕ, обученную на $\mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{aug}}$. Тогда:

$$\text{mAP}(f_\phi(\{x \mid (x,t) \in \mathcal{D}_{\text{val}}\}, T) \geq \text{mAP}(f_\theta(\{x \mid (x,t) \in \mathcal{D}_{\text{val}}\}, T).$$

Эксперимент

Dataset	Size	mAP@50	mAP* = mAP@50-95
COCO	6000	0.278	0.18
COCO + COCO augmentations	6000 + 4500	**0.307**	**0.2**
VOC	5717	0.644	0.461
VOC + VOC augmentations	5717 + 4150	**0.664**	**0.475**

Таблица: Сравнение показателей mAP для модели детекции YOLO, обученной в течение 500 эпох, с порогом фильтрации аугментаций 0.2.

Исследование влияния компонент

Утверждение 2:
В условиях предыдущего утверждения рассмотрим $\mathcal{D}'_{\text{aug}} = \{(x_i^{\text{aug}'}, t_i^{\text{aug}'})\}_{i=1}^{M}$ — аугментированный датасет, где $x_i^{\text{aug}'} \in Y'$ — аугментированные изображения, $t_i^{\text{aug}'} \in T$ — соответствующие разметки. $Y' \subseteq Y$ — пространство аугментированных изображений, сгенерированных моделью следующего вида:

$$r_\gamma \circ h_\beta|_L \circ f_\psi : X \times [0,1] \to Y' \cup \varnothing$$

Рассмотрим модель детекции f_η, обученную на $\mathcal{D}'_{\text{aug}} \cup \mathcal{D}_{\text{train}}$, также рассмотрим f_ϕ, обученную на $\mathcal{D}_{\text{aug}} \cup \mathcal{D}_{\text{train}}$. Тогда:

$$\text{mAP}(f_\phi(\{x \mid (x,t) \in \mathcal{D}_{\text{val}}\}, T) \geq \text{mAP}(f_\eta(\{x \mid (x,t) \in \mathcal{D}_{\text{val}}\}, T).$$

Эксперимент

Dataset	Size	mAP@50	mAP@50-95
COCO + COCO augmentations without expanded prompt	6000 + 4500	0.288	0.186
COCO + COCO augmentations	6000 + 4500	**0.307**	**0.2**
VOC + VOC augmentations without expanded prompt	5717 + 4150	0.663	0.474
VOC + VOC augmentations	5717 + 4150	**0.664**	**0.475**

Таблица: Сравнение показателей mAP для модели детекции YOLO, обученной в течение 500 эпох, с порогом фильтрации аугментаций 0.2.

Исследование влияния компонент

Утверждение 3:
В условиях утверждения 1 рассмотрим $\mathcal{D}''_{\text{aug}} = \{(x_i^{\text{aug}''}, t_i^{\text{aug}''})\}_{i=1}^{M}$ — аугментированный датасет, где $x_i^{\text{aug}''} \in Y''$ — аугментированные изображения пространства $Y'' \supseteq Y$, $t_i^{\text{aug}''} \in T$ — соответствующие разметки. Изображения получены генеративной моделью следующего вида:

$$h_\beta \circ g_\alpha \circ f_\psi : X \to Y''$$

Рассмотрим модель детекции f_ω, обученную на $\mathcal{D}''_{\text{aug}} \cup \mathcal{D}_{\text{train}}$, также рассмотрим f_ϕ, обученную на $\mathcal{D}_{\text{aug}} \cup \mathcal{D}_{\text{train}}$. Тогда:

$$\text{mAP}(f_\phi(\{x \mid (x,t) \in \mathcal{D}_{\text{val}}\}, T) \geq \text{mAP}(f_\omega(\{x \mid (x,t) \in \mathcal{D}_{\text{val}}\}, T).$$

Эксперимент

Dataset	Size	mAP@50	mAP@50-95
COCO + COCO augmentations without filtration model	6000 + 4500	0.287	0.185
COCO + COCO augmentations	6000 + 4500	**0.307**	**0.2**
VOC + VOC augmentations without filtration model	5717 + 4150	0.644	0.46
VOC + VOC augmentations	5717 + 4150	**0.664**	**0.475**

Таблица: Сравнение показателей mAP для модели детекции YOLO, обученной в течение 500 эпох, с порогом фильтрации аугментаций 0.2.

Model	Dataset	Size	mAP@50	mAP@50-95
PowerPaint	COCO + COCO augmentations	6500 + 1500	0.287	0.182
Our			**0.295**	**0.188**
PowerPaint	VOC + VOC augmentations	6800 + 1800	0.673	0.48
Our			**0.675**	**0.49**

Таблица: Сравнение показателей mAP для модели детекции YOLO, обученной в течение 500 эпох, с порогом фильтрации аугментаций 0.23.

Выносится на защиту

1. Предложен автоматизированный подход к созданию аугментированных изображений.
2. Проведены эксперименты, демонстрирующие влияние предложенного метода на качество работы модели детекции, а также выполнено сравнение с существующим подходом.
3. Также проведён анализ влияния отдельных компонентов нашего метода на итоговое значение функции качества.