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

Выпускная квалификационная работа бакалавра

Степанов Илья Дмитриевич
Научный руководитель: к.ф.-м.н. А. В. Грабовой
Научный констультант: А. В. Филатов

Кафедра интеллектуальных систем ФПМИ МФТИ
Специализация: Интеллектуальный анализ данных
Направление: 01.03.02 Прикладная математика и информатика

2024

# Цель исследования

**Задача**

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

**Проблема**

В машинном обучении одной из ключевых проблем является нехватка доступных данных. Объём и разнообразие выборки влияют на обобщающую способность моделей, однако сбор и разметка новых образцов требуют временных и финансовых затрат.

**Цель**

Разработать автоматизированную модель, способную генерировать качественные аугментации и тем самым повышать обобщающую способность моделей детекции. Провести сравнительный анализ влияния аугментаций на показатели качества и изучить вклад отдельных компонентов предложенного метода.

# Постановка задачи

Рассмотрим датасет для задачи детекции:

$$\mathcal{D} = \left\{(x_i,\ t_i),\ i = 1, \ldots, n\right\},$$

где $X$ — пространство изображений, $x_i \in X$ — исходное изображение, $T$ — пространство аннотаций отдельных объектов, $\mathcal{F}(T) \subseteq 2^T$ — пространство аннотаций изображений множества $X$, $t_i \in \mathcal{F}(T)$ — множество аннотаций, соответствующих объектам на изображении $x_i$.

# Постановка задачи

Рассмотрим произвольную модель детекции как отображение:

$$D : X \to \mathcal{F}(\hat{T}),$$

где $\hat{T}$ — пространство аннотаций для отдельных объектов, содержащих координаты ограничивающих прямоугольников, классы объектов и уверенность, предсказанных моделью. $\mathcal{F}(\hat{T}) \subseteq 2^{\hat{T}}$ — пространство предсказанных аннотаций изображений множества $X$.

# Постановка задачи

Определим функцию потерь для модели детекции YOLO $f_\theta$:

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

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

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

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

# Постановка задачи

где $S \times S$ — размер сетки, на которую разбивается изображение, $K$ — количество предсказанных ограничивающих прямоугольников в каждой ячейке сетки, $\lambda_{\text{coord}}, \lambda_{\text{noobj}}$ — коэффициенты, регулирующие вклад в функцию потерь, $\mathbf{I}_{ij}^{\text{obj}}$ — индикатор наличия объекта в $j$-ом прямоугольнике $i$-й ячейки, $\mathbf{I}_{ij}^{\text{noobj}}$ — индикатор отсутствия объекта в $j$-ом прямоугольнике $i$-й ячейки, $(x_i^{gt}, y_i^{gt}, w_i^{gt}, h_i^{gt})$ — координаты центра, ширина и высота истинного ограничивающего прямоугольника для $i$-й ячейки, $(\hat{x}_i, \hat{y}_i, \hat{w}_i, \hat{h}_i)$ — предсказанные координаты ограничивающего прямоугольника для $i$-й ячейки, $C_i$ и $\hat{C}_i$ — истинная и предсказанная вероятность наличия объекта в $i$-й ячейке, $\mathcal{C}$ — множество классов объектов, $p_i(c)$ и $\hat{p}_i(c)$ — истинная и предсказанная вероятность принадлежности объекта классу $c$ для $i$-й ячейки.

Решается следующая оптимизационная задача:

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

# Постановка задачи

Определим функцию потерь для модели детекции DETR $g_\phi$:

$$\mathcal{L}_{DETR}(\phi) = \sum_{i=1}^{N} \Big[ - \log \hat{p}_{\hat{\sigma}_\phi(i)}(c_i) \; + \; \mathbf{I}_{\{\, c_i \neq \varnothing \,\}} \Big( \lambda_{L1} \left\| b_i - \hat{b}_{\hat{\sigma}_\phi(i)} \right\|_1 + \lambda_{\text{giou}} \, \text{GIoU}\big(a_i, \; \hat{a}_{\hat{\sigma}_\phi(i)}\big) \Big) \Big],$$

Для вычисления данной функции потерь необходимо определить оптимальное соответствие с помощью алгоритма назначений.

$$\hat{\sigma} = \arg \min_{\sigma \in S_N} \sum_{i=1}^{N} \Big[ - \mathbf{I}_{\{c_i \neq \varnothing\}} \; \hat{p}_{\sigma(i)}(c_i) + \; \mathbf{I}_{\{c_i \neq \varnothing\}} \; \Big( \lambda_{L1} \left\| b_i - \hat{b}_{\sigma(i)} \right\|_1 + \; \lambda_{\text{giou}} \, \text{GIoU}\big(a_i, \; \hat{a}_{\sigma(i)}\big) \Big) \Big],$$

# Постановка задачи

где $\hat{\sigma}$ — оптимальное соответствие между истинными аннотациями и предсказанными, $S_N$ — множество инъективных отображений из $\{1, \ldots, M\}$ в $\{1, \ldots, N\}$, $M$ — число истинных аннотаций объектов на изображении, $N > M$ — число предсказаных аннотаций объектов на изображении, $\mathbf{I}_{\{c_i \neq \varnothing\}}$ — индикатор наличия объекта в истинном наборе, $c_i$ — истинная метка класса объекта $i$, $\hat{p}_j(c)$ — предсказанная моделью вероятность класса $c$ для аннотации $j$, $a_i$ — истинная аннотация объекта $i$, $b_i$ — истинный ограничивающий прямоугольник объекта $i$, $\hat{a}_j$ — предсказанная аннотация объекта $j$, $\hat{b}_j$ — предсказанный ограничивающий прямоугольник объекта $j$, $\lambda_{L_1}$ и $\lambda_{\text{giou}}$ — регуляризационные коэффициенты для задачи поиска оптимального соответствия, GIoU — функция качества, оценивающая совпадение предсказанной и истинной аннотации.

Решается следующая оптимизационная задача:

$$\phi^* = \arg\min_{\phi} \mathcal{L}_{DETR}(\phi).$$

# Функции качества

Определим функции качества для задачи детекции.
Рассмотрим функцию mAP (mean Average Precision).

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

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

$$\text{AP}(c, \tau, t, \hat{t}) = \int_0^1 P_c(r, \tau, t, \hat{t})\, dr,$$

где $P_c(r, \tau, t, \hat{t})$ — функция, задающая кривую Precision–Recall для класса $c$ при пороге $\tau$, $t \subseteq T$ — множество истинных разметок для класса $c$, $\hat{t} \subseteq \hat{T}$ — множество предсказанных разметок для класса $c$.

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

# Функции качества

Рассмотрим функцию $\text{mAP}_{50:95}$:

$$\text{mAP}_{50:95} : \{\hat{T}\} \times \{T\} \to [0, 1],$$

Определим промежуточную функцию $\text{AP}_{50:95}$ для каждого класса $c$ как усреднение $\text{AP}(c, \tau, t, \hat{t})$ по десяти порогам:

$$\text{AP}_{50:95}(c,\ t,\ \hat{t}) \ = \ \frac{1}{10} \sum_{\tau \in \{0.50,\, 0.55,\, \ldots,\, 0.95\}} \text{AP}(c,\ \tau,\ t,\ \hat{t}).$$

$$\text{mAP}_{50:95}(t,\ \hat{t}) \ = \ \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \text{AP}_{50:95}(c,\ t,\ \hat{t}).$$

# Генеративная аугментация

Рассмотрим модель генеративной аугментации как отображение:

$$F_{\psi,\alpha,\beta,\gamma} : X \times [0,1] \longrightarrow (X_{\text{aug}} \times T_{\text{aug}}) \cup \{\varnothing\},$$

$$f_\psi : X \to M \times L \times T_{\text{aug}}$$

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

$$h_\beta : X \times M \times P \to X_{\text{aug}}$$

$$r_\gamma : Y \times M \times L \times [0,1] \to \{0,1\}$$

где $X$ — пространство исходных изображений, $X_{\text{aug}}$ — пространство аугментированных изображений, $T_{\text{aug}}$ — пространство разметок аугментированных объектов на изображениях, отображение $f_\psi$ — модель детекции объекта, который будет аугментирован, отображение $g_\alpha$ — модель генерации текстового запроса для аугментации нового объекта, отображение $h_\beta$ — модель генерации нового объекта,

# Генеративная аугментация

отображение $r_\gamma$ — модель фильтрации некачественных генераций, $M$ — пространство бинарных масок объектов исходных изображений, $P$ — пространство текстовых запросов для аугментации объекта, $L \subset P$ — пространство классов объектов изображений, число из отрезка $[0, 1]$ отвечает за порог для модели фильтрации.

$$F_{\psi,\alpha,\beta,\gamma}(x, \tau) = \begin{cases} (x_{\text{aug}}, a_{\text{aug}}), & \text{если } r_\gamma\big(x_{\text{aug}},\ m,\ \ell,\ \tau\big) = 1, \\ \varnothing, & \text{если } r_\gamma\big(x_{\text{aug}},\ m,\ \ell,\ \tau\big) = 0, \end{cases}$$

где $(m,\ \ell,\ a_{\text{aug}}) = f_\psi(x),\ x_{\text{aug}} = h_\beta\big(x,\ m,\ g_\alpha(x, \ell)\big)$.

# Генеративная аугментация

Пусть $\mathcal{D} = \mathcal{D}_{\text{val}} \sqcup \mathcal{D}_{\text{train}}$. Рассмотрим аугментированный датасет для задачи детекции:

$$\mathcal{D}_{\text{aug}}(\tau) = \left\{ (x_i^{\text{aug}},\ t_i^{\text{aug}}),\ i = 1, \ldots, m \right\},$$

где $(x_i, t_i) \in \mathcal{D}_{\text{train}}$ — пара «изображение-разметка изображения» из обучающего датасета, $(x_i^{\text{aug}}, a_i^{\text{aug}}) = F_{\psi,\alpha,\beta,\gamma}(x_i, \tau)$ — пара «аугментированное изображение-разметка аугментированного объекта», $a_i^* \in t_i$ — аннотация объекта с наибольшей площадью ограничивающего прямоугольника, $t_i^{\text{aug}} = \left( t_i \setminus \{ a_i^* \} \right) \cup \{ a_i^{\text{aug}} \}$ — разметка аугментированного изображения, $\tau \in [0, 1]$ — пороговое значение для модели фильтрации.

# Генеративная аугментация

**Утверждение 1:**
Пусть $\mathcal{D}_{\text{val}} = \{(x_i,\ t_i),\ i = 1, \ldots, k\}$. Существует такое значение $\tau^* \in [0, 1]$, что модели детекции $f_{\theta_1}$ и $g_{\phi_1}$, обученные на объединённом датасете $\mathcal{D}_{\text{aug}}(\tau^*) \sqcup \mathcal{D}_{\text{train}}$, достигают не меньшего значения по функциям $\text{mAP}$ и $\text{mAP}_{50:95}$ на $\mathcal{D}_{\text{val}}$, чем модели $f_{\theta_2}$ и $g_{\phi_2}$, обученные на $\mathcal{D}_{\text{train}}$. То есть:

$$
\begin{aligned}
\text{mAP}\left(\{f_{\theta_1}(x_i)\}_{i=1}^k, \{t_i\}_{i=1}^k, 0.5\right) &\geq \text{mAP}\left(\{f_{\theta_2}(x_i)\}_{i=1}^k, \{t_i\}_{i=1}^k, 0.5\right), \\
\text{mAP}_{50:95}\left(\{f_{\theta_1}(x_i)\}_{i=1}^k, \{t_i\}_{i=1}^k\right) &\geq \text{mAP}_{50:95}\left(\{f_{\theta_2}(x_i)\}_{i=1}^k, \{t_i\}_{i=1}^k\right), \\
\text{mAP}\left(\{g_{\phi_1}(x_i)\}_{i=1}^k, \{t_i\}_{i=1}^k, 0.5\right) &\geq \text{mAP}\left(\{g_{\phi_2}(x_i)\}_{i=1}^k, \{t_i\}_{i=1}^k, 0.5\right), \\
\text{mAP}_{50:95}\left(\{g_{\phi_1}(x_i)\}_{i=1}^k, \{t_i\}_{i=1}^k\right) &\geq \text{mAP}_{50:95}\left(\{g_{\phi_2}(x_i)\}_{i=1}^k, \{t_i\}_{i=1}^k\right).
\end{aligned}
$$

# Исследование влияния компонент

Рассмотрим модель генеративной аугментации следующего вида:

$$F'_{\psi,\beta,\gamma}(x,\tau) = \begin{cases} (x_{\text{aug}}, a_{\text{aug}}), & \text{если } r_\gamma\left(x_{\text{aug}},\, m,\, \ell,\, \tau\right) = 1, \\ \varnothing, & \text{если } r_\gamma\left(x_{\text{aug}},\, m,\, \ell,\, \tau\right) = 0. \end{cases}$$

где $(m,\, \ell,\, a_{\text{aug}}) = f_\psi(x),\ x_{\text{aug}} = h_\beta\left(x,\, m,\, \ell\right)$.

Рассмотрим аугментированный датасет для задачи детекции:

$$\mathcal{D}'_{\text{aug}}(\tau) = \left\{(x_i^{\text{aug}},\, t_i^{\text{aug}}),\ i = 1, \ldots, n\right\},$$

где $(x_i, t_i) \in \mathcal{D}_{\text{train}}$ — пара «изображение-разметка изображения» из обучающего датасета, $(x_i^{\text{aug}}, a_i^{\text{aug}}) = F'_{\psi,\beta,\gamma}(x_i, \tau)$ — пара «аугментированное изображение-разметка аугментированного объекта», $a_i^* \in t_i$ — аннотация объекта с наибольшей площадью ограничивающего прямоугольника, $t_i^{\text{aug}} = \left(t_i \setminus \{\, a_i^* \,\}\right) \cup \{\, a_i^{\text{aug}} \,\}$ — разметка аугментированного изображения, $\tau \in [0, 1]$ — пороговое значение для модели фильтрации.

# Исследование влияния компонент

**Утверждение 2:**
Пусть $\mathcal{D}_{\text{val}} = \{(x_i,\ t_i),\ i = 1, \ldots, k\}$. Существует такое значение $\tau^* \in [0, 1]$, что модели детекции $f_{\theta_1}$ и $g_{\phi_1}$, обученные на объединённом датасете $\mathcal{D}_{\text{aug}}(\tau^*) \sqcup \mathcal{D}_{\text{train}}$, достигают не меньшего значения по функциям $\text{mAP}$ и $\text{mAP}_{50:95}$ на $\mathcal{D}_{\text{val}}$, чем модели $f_{\theta_2}$ и $g_{\phi_2}$, обученные на $\mathcal{D}'_{\text{aug}}(\tau^*) \sqcup \mathcal{D}_{\text{train}}$. То есть:

$$
\begin{aligned}
\text{mAP}\left(\{f_{\theta_1}(x_i)\}_{i=1}^k, \{t_i\}_{i=1}^k, 0.5\right) &\geq \text{mAP}\left(\{f_{\theta_2}(x_i)\}_{i=1}^k, \{t_i\}_{i=1}^k, 0.5\right), \\
\text{mAP}_{50:95}\left(\{f_{\theta_1}(x_i)\}_{i=1}^k, \{t_i\}_{i=1}^k\right) &\geq \text{mAP}_{50:95}\left(\{f_{\theta_2}(x_i)\}_{i=1}^k, \{t_i\}_{i=1}^k\right), \\
\text{mAP}\left(\{g_{\phi_1}(x_i)\}_{i=1}^k, \{t_i\}_{i=1}^k, 0.5\right) &\geq \text{mAP}\left(\{g_{\phi_2}(x_i)\}_{i=1}^k, \{t_i\}_{i=1}^k, 0.5\right), \\
\text{mAP}_{50:95}\left(\{g_{\phi_1}(x_i)\}_{i=1}^k, \{t_i\}_{i=1}^k\right) &\geq \text{mAP}_{50:95}\left(\{g_{\phi_2}(x_i)\}_{i=1}^k, \{t_i\}_{i=1}^k\right).
\end{aligned}
$$

# Исследование влияния компонент

Аналогично рассмотрим модель генеративной аугментации следующего вида:

$$F''_{\psi,\alpha,\beta}(x,\tau) \;=\; (x_{\text{aug}}, a_{\text{aug}})$$

где $(m,\, \ell,\, a_{\text{aug}}) = f_\psi(x),\; x_{\text{aug}} = h_\beta\big(x,\, m,\, g_\alpha(x,\ell)\big)$.

Рассмотрим аугментированный датасет для задачи детекции:

$$\mathcal{D}''_{\text{aug}}(\tau) = \{(x_i^{\text{aug}},\, t_i^{\text{aug}}),\; i = 1,\ldots,n\}\,,$$

где $(x_i, t_i) \in \mathcal{D}_{\text{train}}$ — пара «изображение-разметка изображения» из обучающего датасета, $(x_i^{\text{aug}}, a_i^{\text{aug}}) = F''_{\psi,\alpha,\beta}(x_i,\tau)$ — пара «аугментированное изображение-разметка аугментированного объекта», $a_i^* \in t_i$ — аннотация объекта с наибольшей площадью ограничивающего прямоугольника, $t_i^{\text{aug}} = \big(t_i \setminus \{\, a_i^* \,\}\big) \cup \{\, t_{\text{aug}} \,\}$ — разметка аугментированного изображения, $\tau \in [0,1]$ — пороговое значение для модели фильтрации.

# Исследование влияния компонент

**Утверждение 3:**
Пусть $\mathcal{D}_{\text{val}} = \{(x_i,\ t_i),\ i = 1, \ldots, k\}$. Существует такое значение $\tau^* \in [0, 1]$, что модели детекции $f_{\theta_1}$ и $g_{\phi_1}$, обученные на объединённом датасете $\mathcal{D}_{\text{aug}}(\tau^*) \sqcup \mathcal{D}_{\text{train}}$, достигают не меньшего значения по функциям $\text{mAP}$ и $\text{mAP}_{50:95}$ на $\mathcal{D}_{\text{val}}$, чем модели $f_{\theta_2}$ и $g_{\phi_2}$, обученные на $\mathcal{D}''_{\text{aug}}(\tau^*) \sqcup \mathcal{D}_{\text{train}}$. То есть:

$$
\begin{aligned}
\text{mAP}\left(\{f_{\theta_1}(x_i)\}_{i=1}^k, \{t_i\}_{i=1}^k, 0.5\right) &\geq \text{mAP}\left(\{f_{\theta_2}(x_i)\}_{i=1}^k, \{t_i\}_{i=1}^k, 0.5\right), \\
\text{mAP}_{50:95}\left(\{f_{\theta_1}(x_i)\}_{i=1}^k, \{t_i\}_{i=1}^k\right) &\geq \text{mAP}_{50:95}\left(\{f_{\theta_2}(x_i)\}_{i=1}^k, \{t_i\}_{i=1}^k\right), \\
\text{mAP}\left(\{g_{\phi_1}(x_i)\}_{i=1}^k, \{t_i\}_{i=1}^k, 0.5\right) &\geq \text{mAP}\left(\{g_{\phi_2}(x_i)\}_{i=1}^k, \{t_i\}_{i=1}^k, 0.5\right), \\
\text{mAP}_{50:95}\left(\{g_{\phi_1}(x_i)\}_{i=1}^k, \{t_i\}_{i=1}^k\right) &\geq \text{mAP}_{50:95}\left(\{g_{\phi_2}(x_i)\}_{i=1}^k, \{t_i\}_{i=1}^k\right).
\end{aligned}
$$

# Эксперимент

| Dataset | Model | Setting | Size | mAP@50 | mAP@50:95 |
|---|---|---|---|---|---|
| Pascal VOC | DETR | original | 4000 | 57.2 | 41.2 |
| | | w/o expanded prompt | 4000 + 4000 | 55.4 | 38.7 |
| | | w/o filter model | 4000 + 4000 | 57.4 | 40.9 |
| | | ours | 4000 + 4000 | **58.2** | **41.4** |
| | YOLO | original | 4000 | 59.6 | 41.5 |
| | | w/o expanded prompt | 4000 + 4000 | 59.4 | 41.2 |
| | | w/o filter model | 4000 + 4000 | 61.4 | **43.2** |
| | | ours | 4000 + 4000 | **61.5** | **43.2** |
| COCO | DETR | original | 5000 | 26.6 | 17.6 |
| | | w/o expanded prompt | 5000 + 5000 | 27.5 | **17.8** |
| | | w/o filter model | 5000 + 5000 | 26 | 16.5 |
| | | ours | 5000 + 5000 | **27.8** | **17.8** |
| | YOLO | original | 5000 | 26.7 | 17.4 |
| | | w/o expanded prompt | 5000 + 5000 | 27.5 | 17.9 |
| | | w/o filter model | 5000 + 5000 | 27.7 | 17.9 |
| | | ours | 5000 + 5000 | **28.2** | **18.3** |

Таблица: Проведение сравнительного анализа значений функций качества mAP@50 и mAP@50:95 моделей DETR и YOLO, обученных на датасетах Pascal VOC и COCO с применением аугментаций и без них, а также анализ влияния отдельных компонентов.

# Выносится на защиту

1. Предложен автоматизированный подход к созданию аугментированных изображений.
2. Проведены эксперименты, демонстрирующие влияние аугментаций на качество работы модели детекции.
3. Проведён анализ влияния отдельных компонентов метода на итоговое значение функций качества.