Аннотация

Аугментация данных — важный инструмент современных исследователей в области детекции, позволяющий увеличить объём обучающей выборки. Однако существующие методы ограничены, поскольку не обеспечивают существенного семантического расширения данных. Это может привести к снижению способности моделей обобщать информацию при работе с аугментированными данными. В этой работе предложен новый метод аугментации, основанный на семантической замене объектов на изображениях. Такой подход обеспечивает расширение обучающих выборок и повышает точность детекционных моделей.

Содержание

1 Введение

В современном машинном обучении одна из ключевых проблем — это нехватка доступных данных. Объём и разнообразие выборки напрямую влияют на обобщающую способность моделей, однако сбор и разметка новых образцов требуют значительных временных и финансовых затрат. В таких условиях аугментация данных становится эффективным инструментом расширения тренировочного набора, позволяющим получать синтетические примеры, близкие к исходным.

Аугментация широко применяется во множестве прикладных задач машинного обучения. В области обработки естественного языка к классическим приёмам относятся переформулирование предложений, синонимичная замена слов, случайное удаление токенов или перестановка фраз. В задаче обработки аудио используются изменения скорости воспроизведения, сдвиг тона, добавление фонового шума и эхо.

В компьютерном зрении распространены геометрические преобразования — повороты, отражения, масштабирование и обрезка, а также приёмы вроде добавления гауссовского шума или изменения яркости и контрастности изображения. В популярной сегментационной модели U-Net для расширения датасета используют деформации с помощью случайных смещений. Тем не менее подобные методы не вносят семантических изменений и ограничены в разнообразии синтезируемых примеров: при поворотах и отражениях сохраняется исходная структура изображения, а при корректировке яркости и контраста меняется лишь цветовая палитра.

С появлением генеративных моделей на базе GAN (Generative Adversarial Networks) и Stable Diffusion научные исследования в области аугментации изображений получили новый импульс. GAN впервые позволили синтезиро-

вать реалистичные изображения, однако их обучение часто сопровождается нестабильностью, проблемами сходимости и артефактами. Диффузионные модели демонстрируют на сегодняшний день ведущие результаты в задаче синтеза высококачественных изображений (state-of-the-art), хотя и требуют больших вычислительных ресурсов.

В этой работе мы предлагаем архитектуру для аугментации изображений в задаче детекции объектов. Наш вклад заключается в следующем:

1. Разработка автоматизированной архитектуры для аугментации.
2. Проведение экспериментов, демонстрирующих влияние аугментаций на целевую метрику качества модели детекции.
3. Анализ вклада отдельных компонентов архитектуры в итоговое качество.

Предложенный метод позволяет существенно разнообразить выборку, что значительно снижает затраты на ручной сбор данных и повышает устойчивость моделей к вариативности входных изображений. Таким образом, наш подход становится важным инструментом для исследователей в области компьютерного зрения, которые стремятся быстро и эффективно получить данные без потери качества.

2 Обзор литературы

Одним из наиболее ранних методов аугментации данных в компьютерном зрении считаются техники, основанные на пространственных и цветовых преобразованиях, такие как повороты, смещения, изменение яркости и контрастности. Например, в статье [2] подробно описаны методы и их реализация для различных задач аугментации. Подобные подходы активно применялись в работах, посвященных классическим архитектурам, таким как ImageNet[12] и U-Net[11], которые стали основой для многих современных моделей в области компьютерного зрения.

С появлением генеративных моделей, таких как Stable Diffusion[10] и GAN[5], исследования в области аугментации данных для задач детекции вышли на качественно новый уровень. Среди последних работ выделяются следующие:

Метод, предложенный в статье[7], использует диффузионные модели для генерации реалистичных фонов, что улучшает детекцию объектов. В работе[4] представлен подход для генерации объектов заданного класса с сохранением семантической согласованности. Модель DiffusionEngine[15] объединяет генерацию изображений с автоматическим созданием разметки, что упрощает подготовку данных для детекции. Исследование AeroGen[13] демонстрирует эффективность диффузионных моделей для аугментации данных в задачах анализа спутниковых изображений. Подход Erase, then Redraw[8] предлагает инновационный метод замены фрагментов изображения с использованием диффузионных моделей. Обзорная статья[1] систематизирует современные методы аугментации на основе диффузионных моделей. Среди других значимых работ можно отметить:

TTIDA[14], комбинирующую текстовые и визуальные модели для аугмен-

тации данных в задачах классификации. Метод[6], обеспечивает сохранение меток при генерации новых данных.

Альтернативой аугментации, направленной на улучшение семантики данных, является подход Open Vocabulary Detection[16], где модель обучается обнаруживать объекты с использованием языковых моделей. Однако этот метод ограничен предопределенным словарем объектов и не позволяет динамически расширять разнообразие данных через уточнение свойств объектов.

3 Постановка задачи

Рассмотрим датасет для задачи детекции:

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

$x_i \in X$ — входное изображение, X — пространство входных изображений $t_i \in T$ — соответствующая аннотация, T — пространство соответствующих аннотаций, содержащих координаты ограничивающих рамок и классы объектов

Рассмотрим модели детекции YOLO[9], DETR[3] как отображения f_θ, g_ϕ :

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

$\hat{T}$ — пространство соответствующих аннотаций, содержащих координаты ограничивающих рамок и классы объектов, предсказанных моделью.

Определим функцию потерь для модели детекции f_θ:

$$\begin{aligned}
\mathcal{L}_{YOLO}(\theta) = \lambda_{\text{coord}} &\sum_{i=1}^{S^2}\sum_{j=1}^{B} \mathbf{I}_{ij}^{\text{obj}} \left[(x_i^{gt} - \hat{x}_i)^2 + (y_i^{gt} - \hat{y}_i)^2\right] \\
&+ \lambda_{\text{coord}} \sum_{i=1}^{S^2}\sum_{j=1}^{B} \mathbf{I}_{ij}^{\text{obj}} \left[(\sqrt{w_i^{gt}} - \sqrt{\hat{w}_i})^2 + (\sqrt{h_i^{gt}} - \sqrt{\hat{h}_i})^2\right] \\
&+ \sum_{i=1}^{S^2}\sum_{j=1}^{B} \mathbf{I}_{ij}^{\text{obj}} (\hat{C}_i - C_i)^2 + \lambda_{\text{noobj}} \sum_{i=1}^{S^2}\sum_{j=1}^{B} \mathbf{I}_{ij}^{\text{noobj}} (\hat{C}_i - C_i)^2 \\
&+ \sum_{i=1}^{S^2} \mathbf{I}_{i}^{\text{obj}} \sum_{c \in \mathcal{C}} (\hat{p}_i(c) - p_i(c))^2,
\end{aligned}$$

где $S \times S$ — размер сетки, на которую разбивается изображение, B — количество предсказанных ограничивающих рамок (bounding boxes) 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-й рамке i-й ячейки, $\mathbf{I}_{ij}^{\text{noobj}}$ —

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.

Определим функцию потерь для модели детекции g_ϕ:

$$\mathcal{L}_{DETR}(\phi) = \sum_{i=1}^{M} \Big[-\log \hat{p}_{\sigma(i)}(c_i) + \lambda_{\ell_1} \|b_i - \hat{b}_{\sigma(i)}\|_1 + \lambda_{\text{giou}} \left(1 - \text{GIoU}(b_i, \hat{b}_{\sigma(i)})\right) \Big] + \sum_{j \in [N] \setminus \sigma([M])} \Big[-\log \hat{p}_j(c_\varnothing) \Big],$$

где M — количество объектов в разметке (ground truth), N — фиксированное количество предсказаний, выдаваемых моделью, $\hat{p}_j(c)$ — вероятность, с которой j-е предсказание относится к классу c, b_i — ограничивающая рамка (bounding box) i-го объекта в разметке, $\hat{b}_j$ — предсказанная моделью рамка для j-го объекта, c_i — истинная метка класса i-го объекта, $c_\varnothing$ — специальный "пустой"класс (фон, no-object), обозначающий отсутствие объекта, $\mathfrak{S}_N$ — множество всех перестановок $\{1, 2, \ldots, N\}$, используемое для поиска наилучшего сопоставления предсказаний с размеченными объектами, $\sigma \in \mathfrak{S}_N$ — конкретная перестановка, сопоставляющая предсказания $\hat{b}_{\sigma(i)}$ объектам b_i. λ_{ℓ_1} — вес для L1-компоненты ошибки по координатам рамки, λ_{giou} — вес для компоненты ошибки на основе обобщённого IoU, $\text{GIoU}(b_i, \hat{b}_{\sigma(i)})$ — метрика Generalized Intersection-over-Union, измеряющая степень совпадения рамок b_i и $\hat{b}_{\sigma(i)}$.

Решаются следующие оптимизационные задачи:

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

4 Методология

Рассмотрим модель 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, f_ψ — модель детекции объекта, который будет аугментирован, g_α — модель генерации 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. Число из отрезка [0,1] отвечает за порог для модели фильтрации.

4.1 Модель для детекции исходного объекта

Наша архитектура полностью автоматизирована. Для выбора объекта, который будет аугментироваться, используется полностью предобученная модель детекции YOLO: она находит на изображении объекты и возвращает тот, который имеет наибольшую площадь ограничивающего прямоугольника

(bbox). Если же требуется аугментация конкретного объекта, пользователь может передать модели маску в формате ограничивающего прямоугольника (bbox), обозначающего нужный объект. По результатам детекции модель возвращает bbox выбранного для аугментации объекта и название объекта. Для корректной математической постановки задачи модель принимает на вход вещественное число из отрезка [0,1] и возвращает его без изменений.

4.2 Модель для генерации текстового запроса

Далее архитектура в автоматическом режиме генерирует текстовый запрос, данный процесс разбит на несколько важных частей: 1)

4.3 Модель генерации нового объекта

4.4 Модель фильтрации сгенерированного изображения

5 Вычислительные эксперименты

6 Заключение

Список литературы

[1] Alimisis, P., I. Mademlis, P. Radoglou-Grammatikis, P. Sarigiannidis, and G. T. Papadopoulos (2025). Advances in diffusion models for image data augmentation: A review of methods, models, evaluation metrics and future research directions.

[2] Buslaev, A. V., A. Parinov, E. Khvedchenya, V. I. Iglovikov, and A. A. Kalinin (2018). Albumentations: fast and flexible image augmentations. *CoRR abs/1809.06839*.

[3] Carion, N., F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko (2020). End-to-end object detection with transformers. *CoRR abs/2005.12872*.

[4] Fang, H., B. Han, S. Zhang, S. Zhou, C. Hu, and W.-M. Ye (2024). Data augmentation for object detection via controllable diffusion models. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1246–1255.

[5] Goodfellow, I. J., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio (2014). Generative adversarial networks.

[6] Kupyn, O. and C. Rupprecht (2024). Dataset enhancement with instance-level augmentations.

[7] Li, Y., X. Dong, C. Chen, W. Zhuang, and L. Lyu (2024). A simple background augmentation method for object detection with diffusion model.

[8] Ma, F., W. Qi, G. Zhao, M. Liu, and J. Ma (2025). Erase, then redraw: A novel data augmentation approach for free space detection using diffusion model.

[9] Redmon, J., S. K. Divvala, R. B. Girshick, and A. Farhadi (2015). You only look once: Unified, real-time object detection. *CoRR abs/1506.02640*.

[10] Rombach, R., A. Blattmann, D. Lorenz, P. Esser, and B. Ommer (2021). High-resolution image synthesis with latent diffusion models. *CoRR abs/2112.10752*.

[11] Ronneberger, O., P. Fischer, and T. Brox (2015). U-net: Convolutional networks for biomedical image segmentation. *CoRR abs/1505.04597*.

[12] Russakovsky, O., J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and L. Fei-Fei (2014). Imagenet large scale visual recognition challenge. *CoRR abs/1409.0575*.

[13] Tang, D., X. Cao, X. Wu, J. Li, J. Yao, X. Bai, D. Jiang, Y. Li, and D. Meng (2025). Aerogen: Enhancing remote sensing object detection with diffusion-driven data generation.

[14] Yin, Y., J. Kaddour, X. Zhang, Y. Nie, Z. Liu, L. Kong, and Q. Liu (2023). Ttida: Controllable generative data augmentation via text-to-text and text-to-image models.

[15] Zhang, M., J. Wu, Y. Ren, M. Li, J. Qin, X. Xiao, W. Liu, R. Wang, M. Zheng, and A. J. Ma (2023). Diffusionengine: Diffusion model is scalable data engine for object detection.

[16] Zhu, C. and L. Chen (2024). A survey on open-vocabulary detection and segmentation: Past, present, and future.