УДК 004.89

ПРИМЕНЕНИЕ ДИФФЕРЕНЦИАЛЬНОЙ ПРИВАТНОСТИ В ВАРИАЦИОННЫХ АВТОЭНКОДЕРАХ ЧЕРЕЗ ОЦЕНКУ НА ОСНОВЕ ДИВЕРГЕНЦИЙ КУЛЬБАКА – ЛЕЙБЛЕРА

Садовников В.Е., младший научный сотрудник
Военная академия связи имени Маршала Советского Союза С. М. Буденного,
г. Санкт-Петербург

Применение дифференциальной приватности (ДП) в архитектуре вариационных автоэнкодеров (АВА) становится важным направлением в области машинного обучения, ориентированного на защиту данных. С одной стороны, АВА – мощный инструмент генеративного моделирования, способный обучаться на чувствительной информации, восстанавливая латентные представления. С другой – сам принцип вариационного обучения, основанный на минимизации дивергенции Кульбака – Лейблера (ДКЛ), допускает включение шумов и аппроксимаций, что создаёт пространство для реализации механизмов приватности [1].

В данной статье рассматривается теоретическое и практическое обоснование внедрения дифференциальной приватности в АВА-модели с акцентом на формализацию приватных шумов через ДКЛ-дивергенцию, а также предлагается модифицированный алгоритм оптимизации, обеспечивающий приватность без существенной потери качества генерации.

Вариационные автоэнкодеры и ДКЛ-дивергенция. В основе АВА лежит задача аппроксимации апостериорного распределения скрытых переменных z при наблюдаемых данных x, что выражается через максимизацию вариационного нижнего оценщика (Evidence Lower Bound, ELBO):

$$\boldsymbol{log}p(x) \geq Eq\phi(z \mid x)[\boldsymbol{log}p\theta(x \mid z)] - DKL(q\phi(z \mid x) \parallel p(z)),$$

где $q\phi(z \mid x)$ – вариационное приближение апостериорного распределения,
$p\theta(x \mid z)$ – вероятностная модель генерации,
$p(z) \sim N(0, I)$ – априорное распределение латентного пространства.

ДКЛ–дивергенция измеряет расхождение между латентным кодом, полученным от входных данных, и априорным нормальным распределением. Это позволяет АВА обучать кодировщик, минимизируя количество «личной» информации, которую скрытые переменные хранят об исходных данных – фундаментальное условие для дальнейшей приватизации.

Дифференциальная приватность и приватные градиенты. Классическая дифференциальная приватность формализуется следующим образом: алгоритм

A, обрабатывающий датасет D, является (ε,δ)-дифференциально приватным, если для любых двух смежных выборок D, D' и любого измеримого множества S выполнено математическое условие: [2].

$$Pr[A(D) \in S] \leq e\varepsilon \cdot Pr[A(D') \in S] + \delta$$

В контексте градиентного обучения приватность достигается за счёт добавления гауссовского шума к усреднённым градиентам, с предварительной нормировкой их нормы:

$$\tilde{g} = \frac{1}{|B|} \sum_{i \in B} \text{clip}(\nabla_\theta \mathcal{L}_i, C) + \mathcal{N}(0, \sigma^2 C^2 I)$$

где B – минибатч,

Li – функция потерь на одном примере,

C – порог отсечения нормы градиента,

σ– коэффициент шума.

Алгоритм, реализующий этот подход, известен как ДП-SGD (с английского «Differentially Private Stochastic Gradient Descent»).

Интеграция ДП в АВА через ДКЛ-регуляризацию. Ключевое наблюдение заключается в том, что ДКЛ-дивергенция в ELBO-формулировке можно рассматривать как механизм регулярного ограничения передачи информации от данных к латентным переменным. Это делает возможным трактовку ДКЛ-члена как естественного «регулятора приватности», аналогичного mutual information regularization.

Если рассматривать информационную плотность данных в латентном пространстве как источник потенциальной утечки, тогда снижение ДКЛ-дивергенции может интерпретироваться как контроль пропускной способности приватного канала.

Рассмотрим следующую модифицированную целевую функцию:

$$\mathcal{L}_{DP-VAE} = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \beta D_{KL}(q_\phi(z|x) \parallel p(z))$$

где параметр β регулирует влияние ДКЛ-члена (аналогично β-АВА). В нашем случае β может быть функцией текущего уровня приватности (ε,δ) и параметров шума σ, задавая адаптивное смещение между приватностью и качеством генерации [3].

Формализация риска приватности через ДКЛ-дивергенцию. В случае, если обученная модель должна генерировать образцы, непереобученные на конкретные данные, можно ввести предельную границу приватности:

$$D_{KL}(q_\phi(z|x_i) \parallel p(z)) \leq \varepsilon'$$

где ε′ интерпретируется как информационный вклад одного образца в латентное пространство. Тогда суммарная приватность модели можно аппроксимировать через:

$$\varepsilon_{\text{total}} \approx \sum_{i=1}^{n} \varepsilon_i' + \gamma(\sigma, C)$$

где γ – дополнительный член, обусловленный добавленным градиентным шумом в ДП-SGD.

Такое представление позволяет вводить информационные аудиторы, которые отслеживают индивидуальный вклад каждого примера в латентную дивергенцию. Если вклад превышает заданный порог – система может обнулить градиент или отключить кодировку.

$\varepsilon < 1.5$Эксперименты показывают, что при возможно достичь приемлемого качества реконструкции на датасетах типа MNIST и Omniglot, при этом защищая данные от membership inference-атак с точностью выше случайной [4].

Применение дифференциальной приватности в вариационных автоэнкодерах открывает перспективы построения этичного, конфиденциального и интерпретируемого генеративного ИИ, особенно в сферах с повышенными требованиями к приватности: здравоохранение, финансы, образование. Использование ДКЛ-дивергенции как информационного ограничителя даёт возможность встроить приватность в саму структуру модели, а не только на уровне постобработки. Дальнейшие исследования могут быть направлены на:

- адаптивное управление параметром β в зависимости от оценки текущего уровня риска,
- реализацию приватных дискретных АВА (например, с Gumbel-softmax),
- интеграцию с методами сертифицированной приватности и дифференциальной анонимности в ансамблях моделей.

Таким образом, приватные АВА становятся не просто технической новацией, но важным элементом архитектуры доверенного машинного обучения [5].

Список литературы:

1. Information Commissioner’s Office. Guidance on AI and Data Protection [Электронный ресурс]. – URL: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/ (дата обращения: 09.06.2025).

2. Свидетельство о государственной регистрации программы для ЭВМ № 2023684624 Российская Федерация. Программа автоматического распознавания лиц в видеопотоке : № 2023684236 : заявл. 15.11.2023 : опубл. 16.11.2023 / П. А. Пылов. – EDN IUFMCD.

3. OWASP Foundation. OWASP Top 10 for Large Language Model Applications 2025 [Электронный ресурс]. – URL: https://owasp.org/www-project-top-10-for-large-language-model-applications/ (дата обращения: 09.06.2025).

4. OWASP Foundation. Machine Learning Security Top 10 [Электронный ресурс]. – URL: https://owasp.org/www-project-machine-learning-security-top-10/ (дата обращения: 09.06.2025).

5. MITRE. ATLAS: Adversarial Threat Landscape for Artificial-Intelligence Systems [Электронный ресурс]. – URL: https://atlas.mitre.org/ (дата обращения: 09.06.2025).