УДК 004.912

Выравнивание представлений в подходе многозадачного обучения для детектирования машинно-сгенерированных текстов

Г.М. Грицай, А.В. Грабовой

[1]Московский физико-технический институт
(национальный исследовательский университет)

Большое количество алгоритмов глубокого обучения оптимизируют пространство параметров моделей в режиме решения одной задачи (single-task). Зачастую этого может быть недостаточно в попытке получить устойчивую модель с высокой обобщающей способностью. В целях достижения более качественных и высокоинформативных представлений для примеров в векторном пространстве модели распространено применение многозадачного (multi-task) обучения, за счет обработки информации, разделяемой между задачами. При совместном использовании заданного количества параметров сети модель вносит структурные особенности в компактное представление данных, что в свою очередь повышает скорость обучения и позволяет повысить качество модели на целевых метриках [1].

Многообразие и стремительное развитие языковых генеративных моделей привело к появлению большого количества искусственных текстов, едва отличимых от написанных человеком. Сгенерированные фрагменты могут содержать плагиат, логические ошибки и информацию, не соответствующую действительности. В данной работе рассмотрено применение метода многозадачного обучения для повышения точности детектирования текстов, сгенерированных различными языковыми моделями.

Пусть M задачам классификации соответствует множество датасетов D = $\{d_1, d_2, \dots, d_M\}$. Модель мультизадачного обучения (MTL) с сильным совместным использованием параметров (HPS) состоит из общей подсети h_{θ_s} с параметрами θ_s и T специфичных сетей под конкретную задачу $g_{\theta_1}, g_{\theta_2}, \dots, g_{\theta_T}$ с параметрами $\{\theta_i\}$. Все параметры MTL:

$$\theta = \theta_s \cup \bigcup_{i \in [T]} \theta_i$$

Обозначим $L_1, L_2, \dots, L_T$ функции потерь каждой задачи. В подходе с MTL будем оптимизировать:

$$\mathcal{L}(\theta) = \sum_{x_j \in \mathbb{D}} \sum_{t \in [T]} L_t\left(g_{\theta_t} \circ h_{\theta_s}(x_j), c_t\right)$$

Данная методология обучения применима как для одной целевой задачи классификации $T = 1$, так и для $T > 1$. В случае одной целевой задачи, дополнительные T-1 сети g_{θ_i} уточняют общее представление и решают задачу внутреннего анализа данных. В текущей работе в качестве общей подсетиb h_{θ_s} выбрана архитектура BERT-подобной модели с общим энкодером, в качестве специфичный сетей — добавление дополнительных «голов» (классификаторов). Выдвинута гипотеза, что данный метод обучения детектора позволяет улучшить представление текстов в векторном пространстве, в частности, сблизить эмбеддинги фрагментов от одного автора и отдалить от разных [2], а также внести в нем кластерную структуру [3].

В качестве обучающей выборки выбран набор данных для детекции сгенерированных текстов, содержащий в себе тексты, написанные человеком и полученные от разных генеративных моделей на разные тематики. Таким образом, создается классификатор для решения исходной целевой задачи — бинарной классификации, а также два дополнительных классификатора для уточнения для каждого примера авторства (генеративной модели) и тематики текста.

Для экспериментов была выбрана мультиязычная модель DeBERTa-v3-base, а метрика качества — F_1-score. Результаты работы модели в режиме single-task и multi-task с разным количеством уточняющих голов показаны в Таб. 1.

Подход/данные	**Development set**	**Test set**
Single-task режим	0.825	0.785
Multi-task с 1 доп. сетью (уточнение авторства)	**0.922**	0.827

Multi-task с 1 доп. сетью (уточнение тематики)	0.917	0.811
Multi-task с 2 доп. сетями	0.879	**0.831**

Таб. 1. Результаты работы стандартного и многозадачного подходов дообучения архитектуры DeBERTa на валидационной и тестовой выборках

Результат PCA разложения представлений текстов в векторном пространстве модели в single-task и multi-task режимах изображен на Рис. 1.

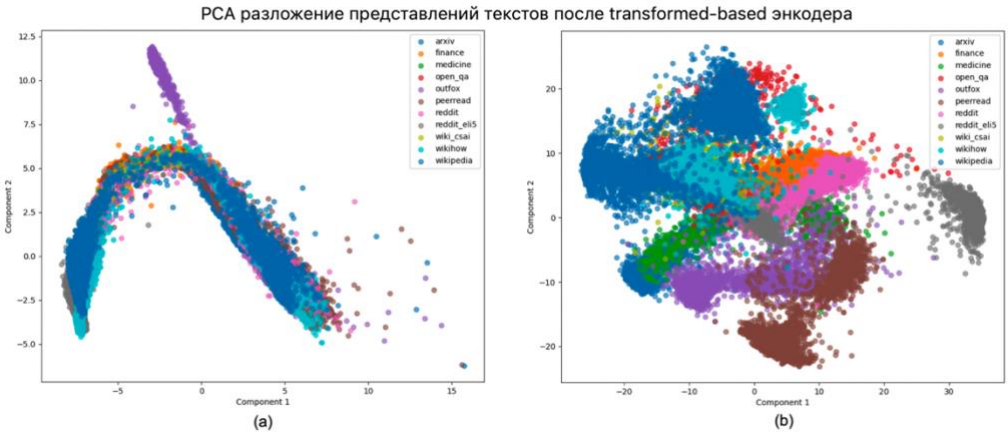

Рис. 1. Двукомпонентное PCA разложение для текстов из dev подвыбрки. Расположение примеров после обучения DeBERTta-v3-base в single-task режиме (a), расположение примеров после обучения той же модели, но в режиме multi-task с двумя дополнительными сетями (b)

В данном эксперименте подход многозадачного обучения позволил улучшить качество детектирования сгенерированных фрагментов на наборе данных с бинарной классификацией. Классификаторы, уточняющие внутренние особенности данных, позволили сблизить эмбеддинги текстов одного автора в векторном пространстве, а также сформировать кластерную структуру, что положительно повлияло на качество целевых метрик.

Литература

1. *Ruder S.* An Overview of Multi-Task Learning in Deep Neural Networks // arXiv preprint. [2017]. arXiv: 1706.05098.
2. *Gritsay G., Grabovoy A., Chekhovich Y.* Automatic Detection of Machine Generated Texts: Need More Tokens // Ivannikov Memorial Workshop (IVMEM). 2022. P. 20-26.
3. *Gritsay G., Voznyuk A., Khabutdinov I., Grabovoy A.* Advacheck at GenAI Detection Task 1: AI Detection Powered by Domain-Aware Multi-Tasking // Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect). 2025. P. 236-243.