

Обучение представлениям электронных документов

Андрей Евгеньевич Алексеев¹, Андрей Анатольевич Михайлов^{1,2}

¹ Институт динамики систем и теории управления им. В.М. Матросова Сибирского отделения Российской академии наук, Россия, 664033, г. Иркутск, ул. Лермонтова, 134

² Институт системного программирования им. В. П. Иванникова Российской академии наук, Россия, 109004, г. Москва, ул. Александра Солженицына, д. 25

Аннотация

В работе рассматривается метод обучения представлений электронных документов с использованием самообучения и инвариантной кластеризации. Основное внимание уделяется применению метода ИС (Invariant Information Clustering) для анализа электронных документов. Эксперименты показывают, что метод ИС способен выявлять информативные признаки из изображений документов и снижать зависимость от ручной разметки. Результаты классификации демонстрируют высокую точность при минимальном количестве размеченных данных.

Ключевые слова

самообучение, инвариантная кластеризация, электронные документы, обучение без учителя, анализ документов, классификация документов, контрастное обучение.

1. Введение

В современном мире работа с документами является неотъемлемой частью многих сфер деятельности. Однако автоматизация обработки документов всё ещё требует значительных усилий. Одной из ключевых задач является обучение компьютера анализу и интерпретации документов. В данной работе рассматривается метод обучения представлений электронных документов с использованием подхода инвариантной кластеризации (ИС).

2. Обзор работ

Самообучение включает различные подходы, среди которых выделяются методы на основе предварительных задач [2,3,4], контрастное обучение [5,6,7], генеративные методы и комбинированные подходы [8,9,10,11,12]. В области обучения представлений электронных документов модель LayoutLM стала первой трансформерной моделью, обученной с совместным использованием информации на уровне текста документа и двумерного позиционного представления текста. Метод ИС реализует подход к обучению без учителя, основанный на максимизации взаимной информации между кластерными назначениями парных образцов данных. Проведённые эксперименты демонстрируют значительное превосходство ИС по сравнению с существующими методами.

3. Метод обучения представлениям электронных документов

Наш подход основан на работе [1]. Авторы представили метод обучения представлениям, основанный на контрастном обучении. Отличительная особенность данного метода заключается в использовании взаимной информации в качестве функции потерь.

7th International Workshop on Information, Computation, and Control Systems for Distributed Environments (ICCS-DE 2025), July 7–11, 2025, Irkutsk, Russia

EMAIL: solmir505@gmail.com (A. 1); mikhailov@icc.ru (A. 2)



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

ICCS-DE 2025 Workshop Proceedings

DOI: 10.47350/ICCS-DE.2025.21

Основная идея заключается в обучении отображения $\Phi : X \rightarrow Y$, которое сохраняет общие для x и x' черты и отбрасывает уникальные для каждого экземпляра особенности.

4. Эксперименты

Архитектура фреймворка, предназначенного для сопровождения научных мероприятий, Для нашего подхода важно, чтобы классы хорошо кластеризовались. Следовательно, нужно отобрать классы, которые визуальнo отличаются друг от друга и содержат чётко выраженные уникальные признаки, характерные для данного класса. Исходя из этого отобраны следующие три класса: чек, паспорт и водительское удостоверение. Эксперименты проводились на компьютере с GPU NVIDIA GeForce RTX 4070 (12GB) и 8-ядерным процессором AMD Ryzen 7 5700X. Конфигурация включала 32GB оперативной памяти. Для оценки качества кластеризации использовались следующие метрики: Adjusted Rand Index (ARI) – 0.909, Homogeneity – 0.909, Completeness – 0.893. На основе кластеризации была решена задача классификации документов. Для этого понадобились метки классов 5% от общего размера датасета. Соотношение меток классов с результатами кластеризации показывает точность модели (accuracy) 0.969, что является хорошим результатом для классификации с учётом сбалансированной выборки.

5. Заключение

В ходе проведённого исследования был изучен и адаптирован метод ПС для задачи обучения представлений электронных документов. Эксперименты подтвердили высокую эффективность метода в условиях отсутствия размеченных данных, что является значительным преимуществом в контексте анализа сложных неструктурированных данных. Метод ПС продемонстрировал способность выявлять информативные признаки из изображений документов, что позволяет снизить зависимость от ручной разметки и повысить качество векторных представлений. Результаты экспериментов показали высокую точность классификации даже при минимальном количестве размеченных данных, что подтверждает практическую применимость метода в области анализа документов. Таким образом, метод ПС представляет собой перспективное решение для построения масштабируемых и устойчивых систем анализа документов на основе методов обучения без учителя. Полученные результаты открывают новые возможности для автоматизации обработки и анализа электронных документов, что может найти широкое применение в различных сферах деятельности.

6. Благодарности

Работа выполнена в рамках государственного задания Министерства науки и высшего образования Российской Федерации (тема № 1023110300006-9).

7. Список использованных источников

- [1] Xu Ji, Joao F Henriques, and Andrea Vedaldi. 2019. Invariant information clustering for unsupervised image classification and segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 9865–9874.
- [2] S. Gidaris, P. Singh, and N. Komodakis, “Unsupervised representation learning by predicting image rotations,” in Int. Conf. Learn. Represent., pp. 1–14, 2018.
- [3] R. Zhang, P. Isola, and A. A. Efros, “Colorful image colorization,” in Eur. Conf. Comput. Vis., pp. 649–666, 2016.
- [4] T. Nathan Mundhenk, D. Ho, and B. Y. Chen, “Improvements to context based self-supervised learning,” in IEEE Conf. Comput. Vis. Pattern Recognit., pp. 9339–9348, 2018.

- [5] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum contrast for unsupervised visual representation learning,” in *IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 9729–9738, 2020.
- [6] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *Int. Conf. Mach. Learn.*, pp. 1597–1607, 2020.
- [7] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, “Barlow twins: Self-supervised learning via redundancy reduction,” in *Int. Conf. Mach. Learn.*, 2021.
- [8] K. He, X. Chen, S. Xie, Y. Li, P. Doll’ar, and R. Girshick, “Masked autoencoders are scalable vision learners,” in *IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 16000–16009, 2022.
- [9] H. Bao, L. Dong, S. Piao, and F. Wei, “Beit: Bert pre-training of image transformers,” in *Int. Conf. Learn. Represent.*, pp. 1–13, 2022.
- [10] J. Zhou, C. Wei, H. Wang, W. Shen, C. Xie, A. Yuille, and T. Kong, “ibot: Image bert pre-training with online tokenizer,” in *Int. Conf. Learn. Represent.*, pp. 1–12, 2022.
- [11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *NAACL*, 2019.
- [12] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations,” in *NeurIPS*, 2020.