УДК 004.89

ПРИМЕНЕНИЕ КОНЦЕПТУАЛЬНОГО РАЗЛОЖЕНИЯ ДЛЯ ИНТЕРПРЕТАЦИИ ЛАТЕНТНЫХ ПРЕДСТАВЛЕНИЙ

Егорова Л.Р., студент гр. БКНАД212, IV курс
Национальный исследовательский университет «Высшая школа экономики»,
г. Москва

Современные модели машинного обучения, особенно глубокие нейронные сети, обладают исключительной способностью к автоматическому извлечению латентных признаков из исходных данных – будь то изображения, тексты, звуковые сигналы или табличная информация. Эти признаки, как правило, не имеют непосредственного интерпретируемого значения с точки зрения человека, но являются фундаментом для вывода модели. Они располагаются в скрытых слоях сети, формируя так называемое латентное представление, или эмбеддинг [1].

В условиях растущей потребности в интерпретируемости и доверии к ИИ-моделям всё больше внимания уделяется методам интерпретации таких латентных представлений, в частности – концептуальному разложению (concept activation decomposition), представляющему собой мощный инструмент для связи абстрактных признаков модели с понятными и осмысленными концепциями из реального мира [2].

Концептуальное разложение – это класс методов, позволяющих проанализировать внутреннее представление модели через призму заранее определённых или автоматически извлечённых концептов, соответствующих семантически значимым категориям. В отличие от методов интерпретации, ограничивающихся входными признаками, концептуальное разложение стремится понять: в каких смысловых измерениях «думает» модель на скрытом уровне? Оно строится на предпосылке, что хотя латентные признаки могут не совпадать напрямую с исходными, они всё же структурированы и соотносятся с осмысленными категориями (например, «раздражение», «форма круга», «фамилия политика», «индикатор финансового риска») [3].

Один из самых известных подходов в этом направлении – TCAV (Testing with Concept Activation Vectors). Суть метода состоит в том, что для заданной концепции (например, «полосатость» или «тревожность») подбирается множество примеров, репрезентирующих её.

Далее обучается линейный классификатор, отделяющий эти примеры от остальной выборки в латентном пространстве модели. Полученное

направление в этом пространстве – вектор активации концепта (CAV) – служит индикатором того, насколько сильна представленность концепции в конкретном эмбеддинге.

Измеряя градиент вывода модели по направлению этого вектора, можно количественно оценить вклад концепции в конкретное предсказание. Это позволяет переходить от «непонятных» активаций нейронов к осмысленным категориям – например, утверждать, что "решение классифицировать изображение как зебру было на 70% обусловлено наличием концепта «полосы»".

Развитие TCAV и аналогичных подходов открыло путь к целому спектру семантических декомпозиций: от интерпретации эмоций в текстах до анализа политических метафор, от распознавания визуальных паттернов в медизображениях до выделения маркеров социальной принадлежности. Подобные методы особенно актуальны в случаях, когда модель может обучаться на смещённых или чувствительных данных – например, ассоциировать женские имена с гуманитарными профессиями, а афроамериканские лица с определёнными классами в системах безопасности. Концептуальное разложение позволяет выявить такие скрытые корреляции до того, как они проявятся в предсказаниях [4].

Следующим шагом в этом направлении стало автоматическое извлечение концептов из данных. Исследователи разрабатывают методы кластеризации эмбеддингов, визуализации скрытых слоёв (например, с помощью UMAP, t-SNE, PCA), а также алгоритмы агрегации схожих активаций для выделения потенциальных концептуальных направлений без предварительной разметки. Такой подход позволяет строить карты семантического пространства модели, выявлять зоны специализации, латентные категории, зоны конфликта и неоднозначности. Визуально это может выражаться в виде интерактивных графов, карт внимания, плотностных диаграмм [5].

Особую важность концептуальное разложение приобретает в задачах, где отсутствует прозрачная причинно-следственная модель. Например, в клинической диагностике, где модель классифицирует снимок или симптом по классу заболевания, но врачи требуют не просто результата, а понимания, на основании каких физиологических маркеров была сделана классификация. Здесь концепты могут соответствовать известным синдромам, участкам изображения, типичным проявлениям – и разложение вывода на вклад этих категорий даёт возможность оценить соответствие логики модели врачебной практике.

Другой важный кейс – правоприменительные системы: классификация текста судебного решения, определение риска нарушения, аннотирование жалоб и т.д. Здесь концепты могут быть связаны с правовыми категориями, типами речевых актов, степенями модальности или стилистическими

признаками. Возможность декомпозировать вывод модели по этим признакам критически важна для обеспечения подотчётности и соблюдения процедурных прав.

Существуют также мультиконцептуальные разложения, в которых используются сразу несколько ортогональных направлений, соответствующих разным аспектам объяснения. Например, в обработке изображений это могут быть форма, цвет, текстура, освещённость; в тексте – стиль, тон, жанр, экспрессивность. Такие разложения позволяют строить многомерные профили предсказания, визуализировать вклад каждого аспекта и проводить сравнительный анализ между примерами.

Инженерные реализации концептуального разложения начинают интегрироваться в современные библиотеки интерпретации: Lucid, NetDissect, InterpretML, Alibi, и другие. Некоторые платформы MLOps включают механизмы отслеживания изменений концептов при обновлении модели, что важно при мониторинге дрейфа и повторной сертификации. Также активно развивается направление интерактивных интерфейсов, в которых пользователь может вводить или корректировать концепты, запрашивать объяснения в терминах своей предметной области и получать не только количественную, но и качественную интерпретацию.

Несмотря на значительный прогресс, остаются вызовы: необходимость построения обоснованных и независимых наборов концептов (чтобы избежать круговой логики), проблемы мультиколлинеарности в латентных пространствах, сложности с интерпретацией высокоабстрактных или перекрывающихся концепций. Кроме того, не все архитектуры легко поддаются семантическому разложению – особенно сложные ансамбли, гибридные модели и self-supervised-системы [6].

Тем не менее, концептуальное разложение становится мощным мостом между эмпирическим поведением модели и понятийной структурой человеческого мышления. Оно позволяет не просто интерпретировать модель в техническом смысле, но и понять, в каких терминах она «рассуждает», какие категории «усваивает», как меняется её внутреннее представление мира в зависимости от данных. Это делает такие методы неотъемлемой частью развития доверенного ИИ, особенно в тех областях, где интерпретация означает не только понимание, но и ответственность.

Список литературы:

1. Kim, B. Interpretability Beyond Feature Attribution : Quantitative Testing with Concept Activation Vectors (TCAV) / B. Kim, M. Wattenberg, J. Gilmer [et al.] // Proceedings of the 35th International Conference on Machine Learning. – 2018. – Vol. 80. – P. 2668–2677.

2. Fong, R. Net2Vec : Quantifying and Explaining How Concepts are Encoded by Filters in Deep Neural Networks / R. Fong, A. Vedaldi // 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. – Salt Lake City : IEEE, 2018. – P. 8730–8738. – DOI 10.1109/CVPR.2018.00910.

3. Bau, D. Network Dissection : Quantifying Interpretability of Deep Visual Representations / D. Bau, B. Zhou, A. Khosla [et al.] // 2017 IEEE Conference on Computer Vision and Pattern Recognition. – Honolulu : IEEE, 2017. – P. 6541–6549. – DOI 10.1109/CVPR.2017.354.

4. Chen, Z. Concept Whitening for Interpretable Image Recognition / Z. Chen, Y. Bei, C. Rudin // Nature Machine Intelligence. – 2020. – Vol. 2. – P. 772–782. – DOI 10.1038/s42256-020-00265-z.

5. Ghorbani, A. Towards Automatic Concept-Based Explanations / A. Ghorbani, J. Wexler, J. Y. Zou, B. Kim // Advances in Neural Information Processing Systems. – 2019. – Vol. 32.

6. Bau, D. GAN Dissection : Visualizing and Understanding Generative Adversarial Networks / D. Bau, J.-Y. Zhu, H. Strobelt [et al.] // International Conference on Learning Representations. – 2019.