УДК 004.89

ПОДХОДЫ К ИНТЕРПРЕТАЦИИ ЛАТЕНТНЫХ ПРЕДСТАВЛЕНИЙ В ПРЕДОБУЧЕННЫХ МОДЕЛЯХ

Тюльпанов А.И., Оператор научной роты, III курс
Военная академия связи имени Маршала Советского Союза С. М. Буденного,
г. Санкт-Петербург

Современные предобученные модели машинного обучения – в особенности крупные трансформерные архитектуры – функционируют за счёт сложных многоуровневых представлений входных данных, которые формируются в скрытых слоях модели в виде так называемых латентных (скрытых) пространств [1].

Эти представления, будучи высокоорганизованными, но непрозрачными для человека, играют фундаментальную роль в принятии решений моделью. Однако по мере роста масштабов моделей и их применения в чувствительных областях всё более остро встаёт вопрос: насколько мы способны понять, что именно «видит» и «решает» модель на промежуточных этапах обработки информации? Ответ на этот вопрос требует разработки подходов к интерпретации латентных представлений, которые могли бы пролить свет на внутреннюю работу предобученных моделей, повысить доверие к ним и обеспечить возможность их контролируемой адаптации [2].

Латентное пространство в нейросетевой модели – это абстрактное многомерное пространство, в котором данные представлены в виде векторов, отражающих их семантические, структурные и статистические свойства.

Например, в языковых моделях токены и предложения кодируются в виде эмбеддингов, которые проходят серию трансформаций через слои внимания и нормализации, формируя представления, всё более отдалённые от исходного текста и всё более приближённые к решаемой задаче. Эти представления крайне важны: именно на их основе формируется финальное решение – будь то классификация, генерация, поиск или предсказание. Однако, несмотря на их значимость, прямое семантическое осмысление латентных векторов остаётся затруднительным: размеры этих пространств могут достигать сотен и тысяч измерений, а сами координаты не имеют очевидной интерпретации.

Подходы к интерпретации латентных представлений стремятся уменьшить этот разрыв между вычислительными структурами и человеческим пониманием. Один из первых и до сих пор распространённых методов – это проекция скрытых векторов в пространство пониженной размерности. Такие

методы, как t-SNE, UMAP и PCA, позволяют визуализировать кластеры векторов, выявлять группы данных, обрабатываемых моделью схожим образом, и находить паттерны, соответствующие семантическим категориям [3].

К примеру, в задачах обработки естественного языка часто удаётся наблюдать, как эмбеддинги слов, связанных по значению, образуют плотные и устойчивые группы, даже если обучение проходило без явного задания такой структуры. Тем не менее, такие визуализации – скорее эвристический инструмент, так как они чувствительны к параметрам метода и не всегда сохраняют глобальную структуру [4].

Более глубокий уровень интерпретации предполагает изучение индивидуальных измерений латентных векторов. В некоторых случаях удаётся связать определённые координаты с конкретными семантическими признаками: например, в обученных эмбеддингах могут обнаруживаться оси, соответствующие грамматическим родам, временам, категориям тональности или степени уверенности. Однако такая осмысленность возникает далеко не всегда, и, как правило, является результатом особой архитектурной настройки модели или специализированного обучения. Более того, в трансформерах представления формируются не просто в линейных подпространствах, а в сложных нелинейных конфигурациях, где значимость координат и их интерпретация может зависеть от положения в предложении, структуры входа и предшествующего контекста [5].

Перспективным направлением является анализ траекторий данных в латентном пространстве: как именно преобразуется представление входа при прохождении через слои модели. Это позволяет выявлять этапы, на которых модель «принимает решение», где возникает категоризация, где усиливаются или подавляются отдельные аспекты данных. Такие исследования особенно актуальны в системах принятия решений, где важно понимать, какие характеристики объекта послужили основанием для классификации. Наблюдение за трансформацией латентных векторов может показать, например, что на ранних этапах доминирует обработка формы, а на поздних – извлечение смысловой информации [6].

Ещё одним подходом становится обратное преобразование из латентного пространства в пространство наблюдаемых признаков. Это реализуется с помощью методов декодирования, когда по заданному латентному вектору пытаются реконструировать исходные данные или сгенерировать аналогичные. Если удаётся построить декодер, способный интерпретируемо и осмысленно восстанавливать входы, это служит сильным индикатором того, что латентное пространство содержит в себе релевантную и интерпретируемую информацию. Подобные методы используются, например, в вариационных автоэнкодерах

или трансформерных генеративных моделях, где латентный вектор становится ключевым элементом управления выходом.

Наконец, в последние годы всё больше внимания уделяется сравнению латентных пространств разных моделей – как способу интерпретировать их архитектурные и поведенческие различия. Исследуется изоморфизм представлений, степень их искажённости, согласованность между слоями и совместимость в условиях переноса обучения. Это особенно важно в инженерной практике, где предобученные модели используются как основа для дообучения на специализированных данных. Понимание структуры и семантики латентного пространства позволяет прогнозировать, насколько эффективно модель адаптируется к новой задаче и где могут возникнуть риски.

Таким образом, подходы к интерпретации латентных представлений в предобученных моделях представляют собой многогранное и активно развивающееся направление, объединяющее методы визуализации, анализа траекторий, реконструкции и сопоставления. Их развитие важно не только для повышения прозрачности и доверия к ИИ-системам, но и для практических целей: от улучшения архитектуры и обучения до отбора моделей, контроля качества и адаптации к новым задачам.

В перспективе интерпретируемость латентных представлений может стать основой для формирования более объяснимых, регулируемых и управляемых моделей, способных не только действовать эффективно, но и аргументировать свои действия в понятной для человека форме.

Список литературы:

1. Bengio, Y. Representation Learning : A Review and New Perspectives / Y. Bengio, A. Courville, P. Vincent // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2013. – Vol. 35, № 8. – P. 1798–1828. – DOI 10.1109/TPAMI.2013.50.

2. Alain, G. Understanding Intermediate Layers using Linear Classifier Probes / G. Alain, Y. Bengio // International Conference on Learning Representations Workshop. – 2017.

3. Raghu, M. SVCCA : Singular Vector Canonical Correlation Analysis for Deep Learning Dynamics and Interpretability / M. Raghu, J. Gilmer, J. Yosinski, J. Sohl-Dickstein // Advances in Neural Information Processing Systems. – 2017. – Vol. 30.

4. Kornblith, S. Similarity of Neural Network Representations Revisited / S. Kornblith, M. Norouzi, H. Lee, G. Hinton // Proceedings of the 36th International Conference on Machine Learning. – 2019. – Vol. 97. – P. 3519–3529.

5. Higgins, I. beta-VAE : Learning Basic Visual Concepts with a Constrained Variational Framework / I. Higgins, L. Matthey, A. Pal [et al.] // International Conference on Learning Representations. – 2017.

6. Locatello, F. Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations / F. Locatello, S. Bauer, M. Lucic [et al.] // Proceedings of the 36th International Conference on Machine Learning. – 2019. – Vol. 97. – P. 4114–4124.