УДК 004.912

ИСПОЛЬЗОВАНИЕ МЕТОДА АНАЛИЗА НЕЗАВИСИМЫХ КОМПОНЕНТ ДЛЯ ВЫДЕЛЕНИЯ В РАЗРЕЖЕННЫХ ЭМБЕДДИНГАХ ДОКУМЕНТОВ ОБОСОБЛЕННЫХ ТЕМАТИК

Шерман М.Л. (ИТМО), **Добрынин В.Ю.** (ИТМО), **Абрамович Р.К.** (ИТМО)
Научный руководитель – кандидат технических наук, доцент Платонов А.В.
(ИТМО)

Введение. Использование плотных векторных представлений остаётся популярным при проектировании моделей на базе архитектуры трансформер. Будь то уже ставшим эталоном BERT [1], или более современная LLaMa [2] - в их основе лежит обучение плотных представлений. Такой подход позволяет достаточно легко сравнивать друг с другом векторы - достаточно простого косинусного расстояния, однако компоненты таких векторов оказываются неотделимы друг от друга, из-за чего выделение отдельных смысловых сущностей представляется трудновыполнимой задачей. Так, при использовании плотных представлений в задаче тематического моделирования, приходится работать со всем вектором сразу, из-за невозможности выделить отдельные смысловые составляющие, которые мы могли бы соотнести с тематиками. Чтобы избавиться от таких ограничений предлагается перейти к разреженному векторному пространству с заранее определёнными свойствами, которое позволит представить объект как набор независимых компонент, что позволит эффективно группировать их по категориям.

Основная часть. Итоговое пространство обладает следующими ограничениями:

1) Разреженность. Для каждого текстового документа будет сопоставлено небольшое число тематик, порядка 7-10, которые однозначно характеризуют его.
2) Независимость. Каждая тематика должна быть максимально независима от другой, чтобы в дальнейшем мы могли выделять обособленные категории и производить выборку по ним.
3) Переход в большую размерность. При векторизации объекта мы сначала преобразуем его моделью на основе плотных векторов, для учитывания скрытых семантических связей. Однако большинство таких моделей работает с пространством небольшой размерности, мы же хотим научиться выделять большое число независимых тематик. Для этого потребуется увеличить размерность пространства.

Первое ограничение задаётся путём использования L1-регуляризации в функции потерь. Второе требует использование одного из нелинейных методов анализа независимых компонент для overcomplete случая – когда число источников данных (элементы итогового векторного представления объекта) много больше, чем число их смесей (элементы плотного эмбеддинга).

Конечная архитектура модели состоит из двух частей – модели на основе плотных векторов ColBERTer [3] и автокодировщика, который сначала переводит данные в желаемое разреженное пространство, а затем восстанавливает исходное представление как на выходе из первой части модели. Это позволяет сохранить информацию, полученную из ColBERTer, во время нелинейных преобразований данных. В результате функция потерь для обучения модели включает подсчёт среднеквадратичной ошибки для восстановления данных, переданных в автокодировщик, L1-норму для задания скрытому пространству разреженности и регуляризационную функцию, обеспечивающую независимость компонент этого же пространства.

Выводы. Спроектирована функция потерь, индуцирующая разреженность пространства и независимость его компонент в задаче тематического моделирования. Продолжаются исследования по оптимизации пространства.

Список использованных источников:

1. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings of the 2019 Conference of the North Proceedings of the 2019 Conference of the North. – Minneapolis, Minnesota: Association for Computational Linguistics, 2019. – C. 4171-4186

2. Touvron H., Lavril T., Izacard G., Martinet X., Lachaux M.-A., Lacroix T., Rozière B., Goyal N., Hambro E., Azhar F., Rodriguez A., Joulin A., Grave E., Lample G. LLaMA: Open and Efficient Foundation Language Models. LLaMA/arXiv:2302.13971 [cs]. - arXiv, 2023.

3. Hofstätter S., Khattab O., Althammer S., Sertkan M., Hanbury A. Introducing Neural Bag of Whole-Words with ColBERTer: Contextualized Late Interactions using Enhanced Reduction. Introducing Neural Bag of Whole-Words with ColBERTer/arXiv:2203.13088 [cs]. - arXiv, 2022.