НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
сентябрь–октябрь 2026 Том 26 № 5 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
September–October 2026 Vol. 26 No 5 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2026-26-5-1080-1083
УДК 004.85

Метод формирования мультимодального векторного пространства графа доступности для поддержки невизуального восприятия информации

Алексей Владимирович Дегтярь✉

Публичное акционерное общество «Банк «Санкт-Петербург», Санкт-Петербург, 195112, Российская Федерация
Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
degtyar.avlad@gmail.com✉, https://orcid.org/0009-0003-5646-9182

Аннотация
Рассмотрена задача формирования мультимодального векторного пространства, устойчивого к недоступности отдельных каналов восприятия информации, характерной для незрячих и слабовидящих пользователей. Предложен метод, основанный на представлении узлов в общем векторном пространстве, получаемом обучаемыми адаптерными проекциями четырехканальных представлений (текст, изображение, программный код, оптическое распознавание), их взвешенным слиянием с учетом маски доступности и обогащением посредством механизма внимания над графом ближайших соседей. Обучение проводится с совместной функцией потерь, содержащей классификационный, согласовывающий и контрастный компоненты. Показано, что на коллекции 2864 статей arXiv-CS, размеченных по 10 ветвям онтологии Computer Science Ontology, в сценарии недоступности текстового канала предложенный метод достигает точности классификации 0,790 против 0,498 у прототипа и 0,560 у версии без графового обогащения.

Ключевые слова
мультимодальные представления, контроль деградации, граф доступности, выпадение модальностей, ассистивные системы, невизуальное восприятие, графовое обогащение, механизм внимания над графом

Ссылка для цитирования: Дегтярь А.В. Метод формирования мультимодального векторного пространства графа доступности для поддержки невизуального восприятия информации // Научно-технический вестник информационных технологий, механики и оптики. 2026. Т. 26, № 5. С. 1080–1083. doi: 10.17586/2226-1494-2026-26-5-1080-1083

Constructing the multimodal vector space of the accessibility graph for non-visual information access

Alexey V. Degtyar✉

“Bank “Saint-Petersburg” Public Joint-Stock Company, Saint Petersburg, 195112, Russian Federation
ITMO University, Saint Petersburg, 197101, Russian Federation
degtyar.avlad@gmail.com✉, https://orcid.org/0009-0003-5646-9182

Abstract
A multimodal vector space robust to the unavailability of individual perception channels — characteristic for blind and visually impaired users — is constructed. The proposed method represents nodes in a common vector space obtained by learnable adapter projections of four channel representations (text, image, program code, and optical character recognition), their weighted fusion with the availability mask, and enrichment via attention over the graph of nearest neighbours. Training uses a joint loss with classification, consistency and contrastive components. On 2,864 arXiv-CS papers labelled by 10 Computer Science Ontology branches and under text channel unavailability, the method reaches 0.790 accuracy against 0.498 for the baseline and 0.560 for the no-enrichment version.

Keywords
multimodal representations, degradation control, accessibility graph, missing modalities, assistive systems, non-visual perception, graph enrichment, graph attention

© Дегтярь А.В., 2026

For citation: Degtyar A.V. Constructing the multimodal vector space of the accessibility graph for non-visual information access. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2026, vol. 26, no. 5, pp. 1080–1083 (in Russian). doi: 10.17586/2226-1494-2026-26-5-1080-1083

Известно, что для совместного представления разнородных данных используются мультимодальные модели, такие как Contrastive Language-Image Pre-training Model (CLIP) [1], выравнивающие модальности различной природы в общем векторном пространстве. Такие модели позволяют задавать общую структуру представлений, но не допускают компенсировать недоступность отдельного канала: при выпадении модальности представление собирается только из оставшихся каналов, без восстановления утраченной информации, а структура связей между документами при этом не используется. Вместе с тем ситуация недоступности отдельных модальностей является типичной для пользователей ассистивных технологий: незрячим и слабовидящим пользователям фактически недоступны изображения и формулы, а пользователям программ экранного доступа – визуально-структурные модальности. По этой причине требуется метод, сохраняющий семантику представления при выпадении части модальностей. Существует несколько подходов, специально разработанных для решения этой проблемы. Так в модели Multimodal Learning with Severely Missing Modality Model (SMI) [2] применяется байесовская процедура мета-обучения, восстанавливающая недоступные каналы на отдельных образцах, а в модели Shared-Specific Model (ShaSpec) [3] представление разделяется на общую и специфическую для каждой модальности компоненты. Данные подходы восстанавливают недостающую информацию на уровне отдельных образцов, но не используют графовой структуры связей между ними. Известные работы в области ассистивных представлений [4] предлагают структурное обогащение информации на уровне отдельных диаграмм для программ экранного доступа, но не на уровне векторного пространства модели.

Целью настоящей работы является создание метода формирования мультимодального векторного пространства, обладающего следующими свойствами: представление узлов получено посредством специализированных канальных кодировщиков и обучаемых адаптерных проекций в общее пространство; при недоступности части каналов используется механизм компенсации, основанный на агрегации представлений соседей в графе ближайших соседей; специализация модели к условиям недоступности каналов обеспечивается обучением с искусственным выпадением модальностей.

Графом доступности называется $G = (\mathbf{V}, \mathbf{E}, \mathbf{M}, \mathbf{U})$, где $\mathbf{V}$ — конечное множество узлов.

Каждому узлу $v \in \mathbf{V}$ сопоставлен набор канальных представлений $\mathbf{E}_v = \{\mathbf{e}_{v,1}, \ldots, \mathbf{e}_{v,4}\} \subset \mathbb{R}^{512}$ (текст, изображение, программный код, оптическое распознавание) и набор масок фактически доступных каналов $\mathbf{M}_v = \{m_{v,1}, \ldots, m_{v,k}\}$, где $m \in \{0, 1\}$; а пространство $\mathbf{U}$ содержит объединенные представления узла $\mathbf{u}_v \in \mathbb{R}^{512}$, формируемые посредством взвешенного слияния доступных каналов согласно

$$\mathbf{u}_v = \Sigma_m \, w_{v,m} \cdot \mathbf{e}_{v,m}, \quad w_{v,m} = \mathrm{softmax}_m(m_{v,m} \cdot l_m),$$

где l_m — обучаемые коэффициенты слияния; $m_{v,m} \in \{0, 1\}$ — маска доступности канала. Канальные представления $\mathbf{e}_{v,m}$ формируются с использованием специализированных кодировщиков: для текстового канала и канала оптического распознавания используется Qwen3-Embedding-0,6B [5], для канала изображения — SigLIP [6], для канала программного кода и фрагментов на LaTeX — CodeBERT [7]. Каждое канальное представление проецируется обучаемыми адаптерными слоями (Linear → GELU → LayerNorm) в общее пространство $\mathbb{R}^{512}$ и нормируется на единичную сферу.

Множество ребер $\mathbf{E}$ задается порогом косинусной близости $^*\tau_e = 0{,}5$ с ограничением сверху $k_{\max} = 50$ соседей; на этапе обучения исключаются ребра, соединяющие узлы одной тематической группы, во избежание утечки информации о целевой метке. Понятие графа доступности следует отграничивать от понятия дерева доступности спецификации ARIA, которое задает структуру документа в терминах языка разметки. В настоящей работе граф доступности задает структуру векторного пространства модели.

Графовое обогащение представления узла основывается на механизме внимания над графом ближайших соседей: для центрального узла v и соседнего узла n с доступным каналом m вычисляется коэффициент внимания $\alpha_{v,n,m} = \mathrm{softmax}_n(\mathrm{LeakyReLU}(\mathbf{a}_m^{\top}[\mathbf{W}_m u_v \,\|\, \mathbf{W}_m \mathbf{e}_{n,m}]))$, причем для каждой из четырех модальностей используются отдельные матрицы $\mathbf{W}_m$ и векторы $\mathbf{a}_m$. Обучение проводится с использованием совместной функции потерь:

$$L = L_{\text{кл}} + \lambda_{\text{м}} \cdot L_{\text{м}} + \lambda_{\text{к}} \cdot L_{\text{к}},$$

где $L_{\text{кл}}$ — перекрестная энтропия классификации по 10 ветвям онтологии Computer Science Ontology; $L_{\text{м}}$ — компонент согласованности представлений при выпадении модальностей; $L_{\text{к}}$ — контрастная функция потерь типа NT-Xent; коэффициенты $\lambda_{\text{м}} = 0{,}3$, $\lambda_{\text{к}} = 0{,}1$ подобраны по результатам экспериментов на валидационной выборке.

Экспериментальная проверка предложенного метода проведена на коллекции 2864 статей репозитория arXiv-CS, размеченных по 10 ветвям онтологии Computer Science Ontology [8]. Целевая задача — оценка деградации мультимодальных представлений при недоступности отдельных каналов на основе классификации по ближайшему соседу в выученном пространстве представлений $^*\mathbf{u}_v$ с 10 классами (точность случайного выбора 0,1). Выборка разбивается на обучающую, валидационную и тестовую части в соотношении 60/20/20 при условии стратификации по классам; используется индуктивная постановка. Проведено 5 обучающих прогонов со случайными начальными значениями параметров; доверительные интервалы получены методом ресемплинга с 10 000 итераций.

Из таблицы следует, что в сценарии полной доступности каналов различия между мультимодальными

Таблица. Результаты сравнения предложенного метода с аналогами на коллекции arXiv-CS (2864 узла, 5 прогонов, ± стандартное отклонение)

Table. Comparison of the proposed method with analogues on the arXiv-CS dataset (2,864 nodes, 5 runs, ± standard deviation)

Метод	Результаты классификации выходных данных использованных методов			
	полный набор модальностей	без текстовой модальности	Δ	*p*
Случайный выбор	0,10	0,10	+0,690	менее 0,0001
Qwen3 (без обучения)	0,752	—	+0,066 / —	0,001 / —
Усреднение + внимание над графом	0,825 ± 0,009	0,498 ± 0,024	−0,007 / +0,292	не значимо / менее 0,0001
Без графового обогащения	0,814 ± 0,008	0,560 ± 0,026	+0,005 / +0,230	0,72 / менее 0,0001
Предложенный метод	0,818 ± 0,011	0,790 ± 0,014	—	—

Примечание: Δ — разность значений точности «предложенный метод–прототип/вариант» в сценариях полной доступности и недоступности текстового канала соответственно; *p* — статистическая значимость различия, полученная методом ресемплинга (10 000 итераций); «—» — случаи, в которых метод не работает.

методами оказываются статистически незначимыми (*p* более 0,7): классификационный сигнал переносится непосредственно доминирующим текстовым каналом. В сценарии недоступности текстового канала значение точности у прототипа снижается с 0,825 до 0,498, у версии без графового обогащения — с 0,814 до 0,560, у предложенного метода — с 0,818 до 0,790. Преимущество предложенного метода составляет +0,292 над прототипом и +0,230 над собственной версией без графового обогащения (*p* менее 0,0001 в обоих случаях).

Полученные результаты показывают, что графовое обогащение позволяет компенсировать недоступность отдельных каналов восприятия и снизить деградацию мультимодальных представлений. В условиях полной доступности каналов вклад графового обогащения находится в зоне статистического шума, тогда как в условиях недоступности доминирующего текстового канала обогащение обеспечивает значимое увеличение точности через агрегацию представлений узлов-соседей. Направления дальнейшей работы включают сопоставление с известными методами SMIL [2] и ShaSpec [3] в рамках единой методики и реализацию навигационного интерфейса программы экранного доступа на основе выученного пространства представлений.

Литература

1. Radford A., Kim J.W., Hallacy C., Ramesh A., Goh G., Agarwal S., et al. Learning transferable visual models from natural language supervision // Proc. of the 38th International Conference on Machine Learning. 2021. V. 139. P. 8748–8763.
2. Ma M., Ren J., Zhao L., Tulyakov S., Wu C., Peng X. SMIL: Multimodal learning with severely missing modality // Proc. of the AAAI Conference on Artificial Intelligence. 2021. V. 35. N 3. P. 2302–2310. doi: 10.1609/aaai.v35i3.16330
3. Wang H., Chen Y., Ma C., Avery J., Hull L., Carneiro G. Multi-modal learning with missing modality via shared-specific feature modelling // Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2023. P. 15878–15887. doi: 10.1109/CVPR52729.2023.01524
4. Morris M.R., Johnson J., Bennett C.L., Cutrell E. Rich representations of visual content for screen reader users // Proc. of the CHI Conference on Human Factors in Computing Systems. 2018. P. 1–11. doi: 10.1145/3173574.3173633
5. Zhang Y., Li M., Long D., Zhang X., Lin H., Yang B., et al. Qwen3 Embedding: Advancing text embedding and reranking through foundation models // arXiv. 2025. arXiv:2506.05176. doi: 10.48550/arXiv.2506.05176
6. Zhai X., Mustafa B., Kolesnikov A., Beyer L. Sigmoid loss for language image pre-training // Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV). 2023. P. 11941–11952. doi: 10.1109/ICCV51070.2023.01100
7. Feng Z., Guo D., Tang D., Duan N., Feng X., Gong M., et al. CodeBERT: A pre-trained model for programming and natural languages // Proc. of the EMNLP. 2020. P. 1536–1547. doi: 10.18653/v1/2020.findings-emnlp.139
8. Salatino A., Thanapalasingam T., Mannocci A., Birukou A., Osborne F., Motta E. The Computer Science Ontology: a comprehensive automatically-generated taxonomy of research areas // Data Intelligence. 2020. V. 2. N 3. P. 379–416. doi: 10.1162/dint_a_00055

References

1. Radford A., Kim J.W., Hallacy C., Ramesh A., Goh G., Agarwal S., et al. Learning transferable visual models from natural language supervision. *Proc. of the 38th International Conference on Machine Learning*, 2021, vol. 139, pp. 8748–8763.
2. Ma M., Ren J., Zhao L., Tulyakov S., Wu C., Peng X. SMIL: Multimodal learning with severely missing modality. *Proc. of the AAAI Conference on Artificial Intelligence*, 2021, vol. 35, no. 3, pp. 2302–2310. doi: 10.1609/aaai.v35i3.16330
3. Wang H., Chen Y., Ma C., Avery J., Hull L., Carneiro G. Multi-modal learning with missing modality via shared-specific feature modelling. *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 15878–15887. doi: 10.1109/CVPR52729.2023.01524
4. Morris M.R., Johnson J., Bennett C.L., Cutrell E. Rich representations of visual content for screen reader users. *Proc. of the CHI Conference on Human Factors in Computing Systems*, 2018, pp. 1–11. doi: 10.1145/3173574.3173633
5. Zhang Y., Li M., Long D., Zhang X., Lin H., Yang B., et al. Qwen3 Embedding: Advancing text embedding and reranking through foundation models. *arXiv*, 2025. arXiv:2506.05176. doi: 10.48550/arXiv.2506.05176
6. Zhai X., Mustafa B., Kolesnikov A., Beyer L. Sigmoid loss for language image pre-training. *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 11941–11952. doi: 10.1109/ICCV51070.2023.01100
7. Feng Z., Guo D., Tang D., Duan N., Feng X., Gong M., et al. CodeBERT: A pre-trained model for programming and natural languages. *Proc. of the EMNLP*, 2020, pp. 1536–1547. doi: 10.18653/v1/2020.findings-emnlp.139
8. Salatino A., Thanapalasingam T., Mannocci A., Birukou A., Osborne F., Motta E. The Computer Science Ontology: a comprehensive automatically-generated taxonomy of research areas. *Data Intelligence*, 2020, vol. 2, no. 3, pp. 379–416. doi: 10.1162/dint_a_00055

Автор

Дегтярь Алексей Владимирович — ведущий специалист, Публичное акционерное общество «Банк «Санкт-Петербург», Санкт-Петербург, 195112, Российская Федерация; преподаватель практики, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 59914986700, https://orcid.org/0009-0003-5646-9182, degtyar.avlad@gmail.com

Author

Alexey V. Degtyar — Leading Specialist, "Bank "Saint-Petersburg" Public Joint-Stock Company, Saint Petersburg, 195112, Russian Federation; Practice Teacher, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 59914986700, https://orcid.org/0009-0003-5646-9182, degtyar.avlad@gmail.com

Статья поступила в редакцию 08.06.2026
Одобрена после рецензирования 23.08.2026
Принята к печати 20.09.2026

Received 08.06.2026
Approved after reviewing 23.08.2026
Accepted 20.09.2026



Работа доступна по лицензии
Creative Commons
«Attribution-NonCommercial»