НЕЧЁТКИЕ СЕТИ ВНИМАНИЯ ДЛЯ МУЛЬТИМОДАЛЬНОГО ИИ: ВСТРОЕННАЯ ИНТЕРПРЕТИРУЕМОСТЬ В АРХИТЕКТУРЕ ТРАНСФОРМЕРОВ

Трофимов Ю.В., Аверкин А.Н.[1], Лебедев А.Д.[1], Ильин А.С.[2], Лебедев М.Д.[3]

Государственный университет «Дубна»; ЛИТ им. М.Г. Мещерякова ОИЯИ; Россия, 141980, г. Дубна; ura_trofim@bk.ru
[1]Государственный университет «Дубна»; Россия, 141980, г. Дубна; averkin2003@inbox.ru, lad.24@uni-dubna.ru
[2]Университет Иннополис; Государственный университет «Дубна»; Россия, 420500, г. Иннополис; a.ilin@innopolis.university
[3]НИТУ «МИСиС»; Россия, 119049, г. Москва; lebedevmisha2003@yandex.ru

Современные мультимодальные трансформеры (CLIP, BLIP) обеспечивают высокую точность при анализе изображений и текста, однако остаются во многом непрозрачными: ход их рассуждений и вклад отдельных признаков восстановить практически невозможно. Это существенно ограничивает применение таких моделей в критически важных областях [1,2].

Для решения указанной проблемы мы предлагаем собственную разработку FAN — нечеткие сети внимания, которые встраивают прозрачность в архитектуру нейросети. Вместо скрытых вычислений используются понятные правила: «ЕСЛИ признак имеет значение А, ТО выход будет В». Система применяет гибкие функции принадлежности и специальные операции (t-нормы), сохраняя точность и прозрачность.

Тестирование на четырех стандартных наборах данных показало: Stanford Dogs (F1=95.74%), медицинские снимки кожи HAM10000 (F1=89.30%), рентгенограммы грудной клетки (F1=78.0%), CIFAR-10 (F1=88.0%). Точность осталась на уровне CLIP и BLIP, однако теперь модель объясняет свои решения. Абляционный анализ показал: обучаемые t-нормы дают +2.65% F1, кросс-модальные слои +3.45% F1.

Таким образом, архитектура нейросети сама обеспечивает интерпретируемость. Это открывает дорогу к применению таких систем в реальных критических приложениях, где нужна не только точность, но и доверие [3].

Работа выполнена в рамках государственного задания Министерства науки и высшего образования Российской Федерации (тема № 124112200072-2).

Литература.

1. *Lanham T., Chang K., Rajkomar A., et al.* Measuring faithfulness in chain-of-thought reasoning // arXiv preprint arXiv:2307.13702. 2023.
2. *Pahud de Mortanges A., Duane A.M., Hardman C.A.* Orchestrating explainable artificial intelligence for multimodal and longitudinal data in medical imaging // npj Digital Medicine. 7, 195. 2024.
3. *Трофимов Ю.В., Аверкин А.Н.* Связь доверенного искусственного интеллекта и XAI 2.0: Теория и фреймворки // Мягкие измерения и вычисления. 90, 68-84. 2025.