FUZZY ATTENTION NETWORKS FOR MULTIMODAL AI: EMBEDDED INTERPRETABILITY IN TRANSFORMER ARCHITECTURE

Trofimov Yu.V., Averkin A.N.[1], Lebedev A.D.[1], Ilyin A.S.[2], Lebedev M.D.[3]

Dubna State University; MLIT JINR; Dubna, Russia; ura_trofim@bk.ru
[1]Dubna State University; Dubna, Russia; averkin2003@inbox.ru, lad.24@uni-dubna.ru
[2]Innopolis University; Dubna State University; Innopolis, Russia; a.ilin@innopolis.university
[3]NUST MISIS; Moscow, Russia; lebedevmisha2003@yandex.ru

Modern multimodal transformers (such as CLIP, BLIP) provide high accuracy in image and text analysis but remain largely opaque: tracing their reasoning and the contribution of individual features is nearly impossible. This significantly limits the application of such models in safety-critical domains [1,2].

To address this issue, we propose FAN (Fuzzy Attention Networks), which embed transparency directly into the neural network architecture. Instead of black-box computations, interpretable rules are used: ”IF a feature has value A, THEN the output is B”. The system employs flexible membership functions and special operations (t-norms) while maintaining accuracy and transparency.

Testing on four standard datasets showed the following results: Stanford Dogs (F1=95.74%), HAM10000 skin lesion images (F1=89.30%), Chest X-Ray (F1=78.0%), and CIFAR-10 (F1=88.0%). The accuracy remained on par with CLIP and BLIP models, but the model is now capable of explaining its decisions. Ablation analysis showed that learnable t-norms provide an increase of +2.65% F1, while cross-modal layers add +3.45% F1.

Thus, the neural network architecture itself ensures interpretability. This paves the way for applying such systems in real-world critical applications where not only accuracy but also trust is required [3].

The work was carried out within the framework of the state assignment of the Ministry of Science and Higher Education of the Russian Federation (theme No. 124112200072-2).

References.

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. *Trofimov Yu.V., Averkin A.N.* The Connection Between Trusted Artificial Intelligence and XAI 2.0: Theory and Frameworks // Soft Measurements and Computing. 90, 68-84. 2025.