0.1. *Козинец Р.М.* **Анализ работы сети глубокого обучения с использованием логических решающих функций**

В работе рассматривается проблема интерпретируемости сверточных нейронных сетей в задаче классификации изображений. В задаче классификации множество объектов представимо в виде совокупности объектов с известными метками класса и подмножества объектов, классовую принадлежность которых требуется установить с минимальной вероятностью ошибки на всем множестве. Данные представляются в виде изображений. Модель — классификатор, способный интерпретировать предсказание в понятной для человека форме, может помочь специалистам в медицине проводить диагностику патологий на основе компьютерной томограммы с помощью нейросетей с большим доверием и пониманием процесса.

В качестве модели-классификатора была разработана новая архитектура нейросети – Neural Pattern Tree. Идея заключается в использовании прототипов [1] — специфичных частей изображения, наличие которых на изображении частично или полностью определяет категорию на изображении, и использовании дерева решений [2], которое использует значения сходства прототипов модели и патчей распознаваемого изображения в качестве признаков. Модель может представить предсказание в виде решающего пути вдоль дерева, визуально отображая части изображения, которые совпали с обученными прототипами и выделяя наиболее важные участки [3]. Разработанный метод сравнивался с классической архитектурой сверточной нейронной сети с полносвязным слоем в качестве классификатора. По результатам сравнения разработанный метод показывает сопоставимую точность классификации, при этом предоставляя интерпретируемый процесс принятия решения.

Работа выполнена при финансовой поддержке РНФ (грант № 19-29-01175).

Научный руководитель — д.т.н. Бериков Владимир. Борисович.

Список литературы

[1] Chen C., Li O., Barnett A., Su J.K., Rudin C. This looks like that: deep learning for interpretable image recognition // NeurIPS. 2019.

[2] Frosst N., Hinton G. Distilling a neural network into a soft decision tree // arXiv preprint arXiv:1711.09784. 2017.

[3] Selvaraju R.R., Das A., Vedantam R., Cogswell M., Parikh D., Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization // International Journal of Computer Vision. 2019. Vol. 128. P. 336–359.