Выбор интерпретируемых сверточных моделей глубокого обучения

A Preprint

Мурадов Тимур
МФТИ

Олег Бахтеев
МФТИ

Константин Яковлев
МФТИ

Abstract

В статье рассматривается проблема слабой интерпретируемости сверточных нейронных сетей. Слабая интерпретируемость затрудняет выделение наиболее важных признаков, а также определение кластеров схожих объектов. Для улучшения интерпретируемости в статье ведётся модификация доказавшего свою эффективность метода OpenBox работающего с кусочно-линеными нейронными сетями. Метод обобщается на работу с более широким классом нейронных сетей - сверточными нейронными сетями. Предлагается математически эквивалентная замена слоев: свертка, пулинг, нормализация на линейные, что позволяет значительно улучшить интепретируемость.

Keywords Machine Learning · CNN · OpenBox · Explicit

1 Introduction

2 Headings: first level

2.1 Headings: second level

3 Examples of citations, figures, tables, references

3.1 Citations

3.2 Figures

3.3 Tables

3.4 Lists