Summary: Sequence classification problems often arise in such areas as bioinformatics and natural language processing. In the last few years, best results in this field were achieved by the deep learning methods, especially by architectures based on recurrent neural networks (RNN). However, the common problem of such models is a lack of interpretability, i.e., extraction of key features from data that affect the most the model’s decision. Meanwhile, using of less complicated neural network leads to decreasing predictive performance thus limiting usage of state-of-art machine learning methods in many subject areas. In this work we propose a novel interpretable deep learning architecture based on extraction of principal sets of short substrings – sequence motifs. The presence of extracted motif in the input sequence is a marker for a certain class. The key component of proposed solution is differential alignment algorithm developed by us, which provides a smooth analog of classical string comparison methods such as Levenshtein edit distance, and Smith-Waterman local alignment. Unlike previous works devoted to the motif based classification, which used CNN for shift-invariant searching, ours model provide a way to shift and gap invariant extraction of motifs.

MSC:

68T05 Learning and adaptive systems in artificial intelligence
62H30 Classification and discrimination; cluster analysis (statistical aspects)
68T07 Artificial neural networks and deep learning
68T10 Pattern recognition, speech recognition

Keywords:
sequence classification; machine learning; neural network; motif extraction

Software:
LSTMVis; CUDA

Full Text: DOI MNR

References:


This reference list is based on information provided by the publisher or from digital mathematics libraries. Its items are heuristically matched to zbMATH identifiers and may contain data conversion errors. It attempts to reflect the references listed in the original paper as accurately as possible without claiming the completeness or perfect precision of the matching.