

**Wymagający złamania wierszy
tytuł pracy w języku polskim**

(English title)

Grzegorz Ciesielski

Praca licencjacka

Promotor: dr Jan Chorowski

Uniwersytet Wrocławski
Wydział Matematyki i Informatyki
Instytut Informatyki

31 lipca 2019

Streszczenie

Poniższa praca sprawdza hipotezę, że można nauczyć sieć neuronową rozpoznawać tekst bez korzystania z podpisanych próbek tekstu, a jedynie z rozkładów prawdopodobieństw wystąpienia pewnych sekwencji liter (n-gramów) w danych uczących.

Below thesis validates hypothesis whether it is possible to teach a neural network text recognition without labeled data, but using sequential output statistics (in form of n-grams) only.

Spis treści

1. Introduction	7
1.1. Problem formulation	7
2. Output Distribution Match	9
Bibliografia	11

Rozdział 1.

Introduction

Most machine learning models require two main factors to function: well designed model and (usually a lot of) data. Preparing a model usually does not raise problems - there are a lot of them available online, and their performance usually allows them to learn on semi-advanced, affordable machines. But to do so, we also need a second factor - data. Especially in text recognition it seems that we cannot build a decent model without any labeled samples, and gathering such data requires a lot of human effort. Hence in this work I will study possibility of training neural net to recognize written text without labeled data, but only using sequential statistics regarding output, as in [1], except that we will use higher-variance dataset and more complex model.

1.1. Problem formulation

As mentioned before, we consider problem of learning text classifier using dataset consisting (pure, not labeled) sequences of data and their output statistics. In language processing obtaining such dataset is largely simpler than labeling data, as it is enough to get images of written text in specific language and for output statistics use easily obtainable n-grams of this language. Specifically: our classifier predicts sequences (y_1, \dots, y_n) (labels) from input sequences (x_1, \dots, x_n) , and has access to dataset $D = \{(x_1^k, \dots, x_n^k) : k = 1, \dots, M\}$.

Rozdział 2.

Output Distribution Match

Bibliografia

- [1] Yu Liu, Jianshu Chen, Li Deng. "Unsupervised Sequence Classification using Sequential Output Statistics"