Text Classification with Deep Learning

National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute
Speaker: D. L. Lazarenko
Supervisor: Ph.D. in Physico-mathematical Science, docent A. Maltsev

Kyiv, 2018

Introduction

Aim of this thesis is building an effective model which have high accuracy and an appropriate speed for classification of advertisements at the e-commerce platform Jiji.ng.
Object of study is advertisements at e-commerce platform
Subject of study is classification model for advertisements:

Relevance of the problem

- e-commerce sales are quickly increasing
- large online e-commerce websites serve millions of users' requests per day
- processes of registrations and purchases as much convenient and fast as possible
- users have to make a choice from more than hundred categories
- automatic category prediction is very important in terms of saving moderators' time and as a result, decreasing the number of necessary moderators to process them

Structure of the data files

lvl2	**titles**	**descriptions**
29	Clean Toyota Camry 2008 Silver	Fairly used Toyota 08 Camry with no problems V4 engine fabric seats and interior
25	Look Unique	Nice, quality, adorable,unique dress available now, whatsapp me

Existing approaches

Let's assume we have the following sentences:

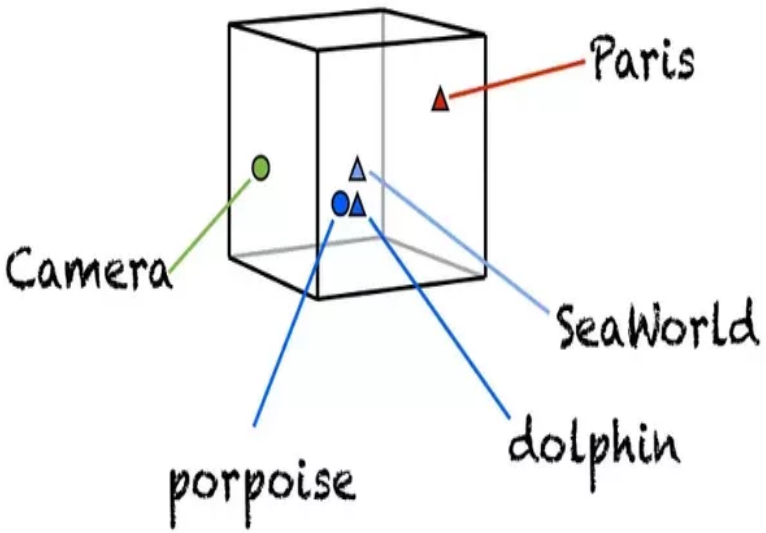
["The sun is yellow", "The sky is blue"]

Encode words with the Bag-of-words method

Text	the	sun	is	yellow	sky	blue
T_1	1	1	1	1	0	0
T_2	1	0	1	0	1	1

1. Naive Bayes
2. Logistic Regression
3. Support Vector Machines (SVMs)
4. Decision Trees and Random Forests

Embeddings



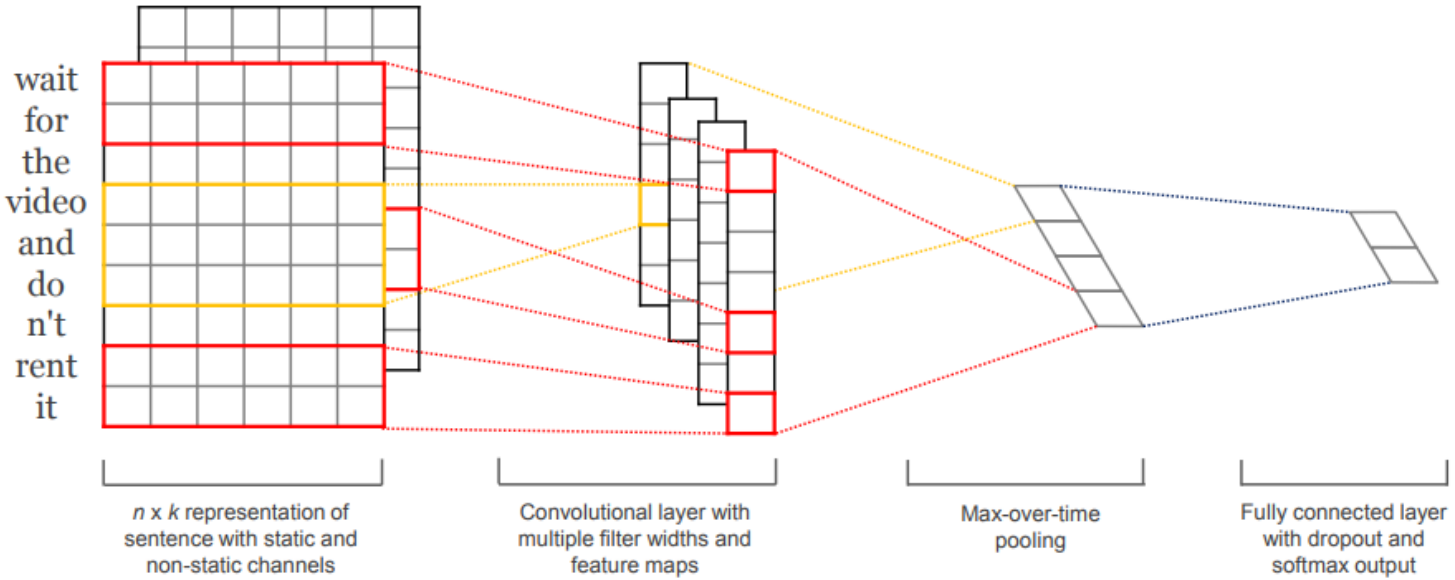
An **embedding** is a mapping from discrete objects, such as words, to vectors of real numbers. For example, a 300-dimensional embedding for English words could include:

blue: $(0.059, 0.7597, ...)$

Bi-LSTM Neural Network

Metric	**Train**	**Test**
categorical accuracy	0.7975	0.8203
category crossentropy	0.8532	0.7478
top k accuracy	0.9189	0.9219.

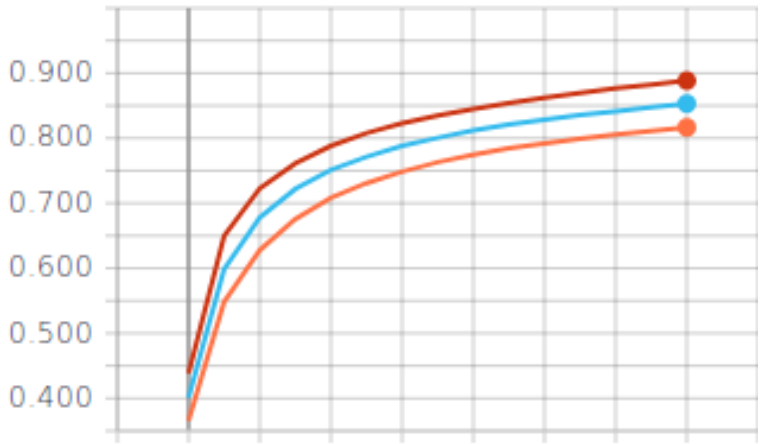
Convolution Neural Network



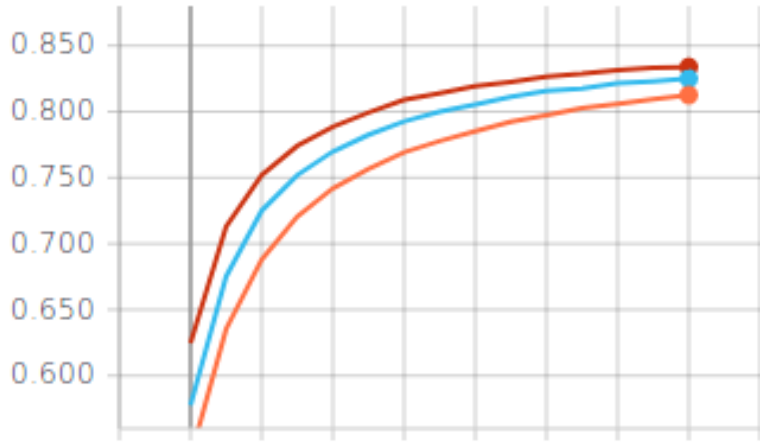
- 300 filters
- size of filter: 3, 4, 5
- l2-regularization equals to 0.01
- dropout equals to the rate 0.5

Overfitting

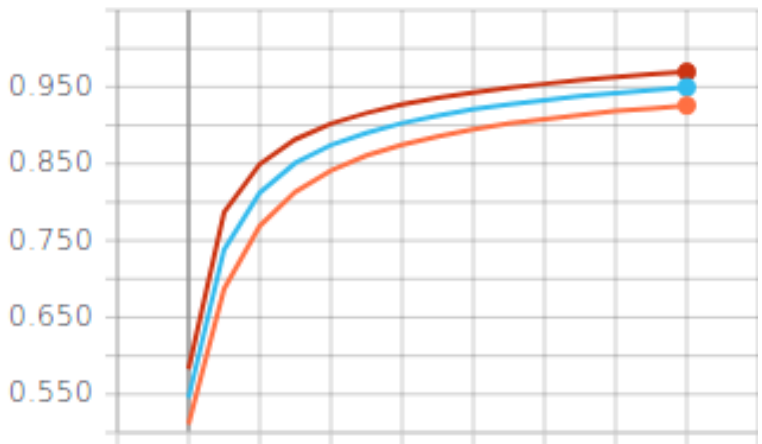
categorical_accuracy



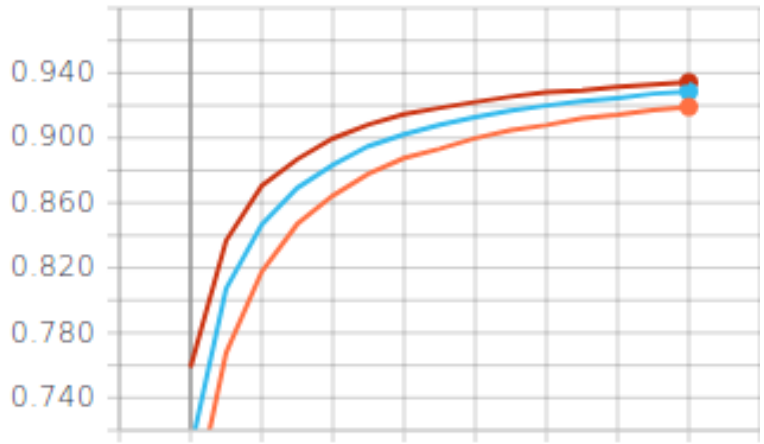
val_categorical_accuracy



top_k_categorical_accuracy



val_top_k_categorical_accuracy

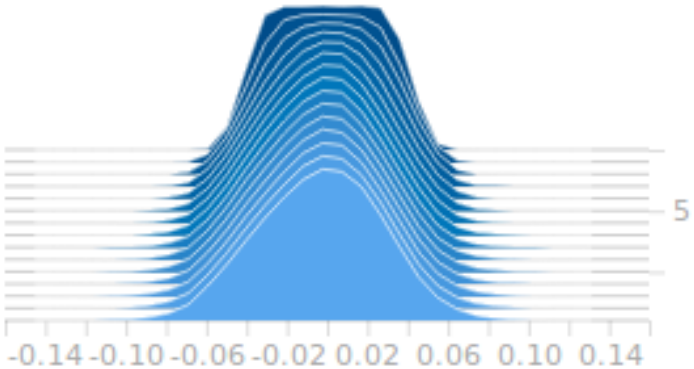


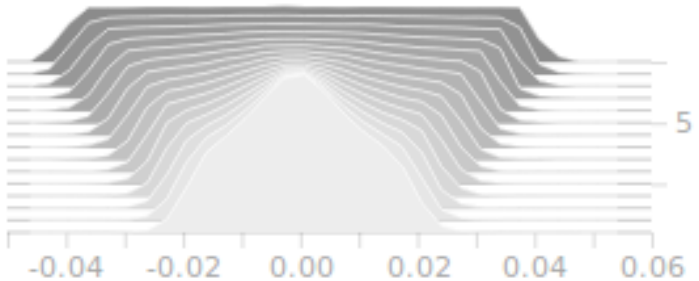
Regularizations

Modifications:

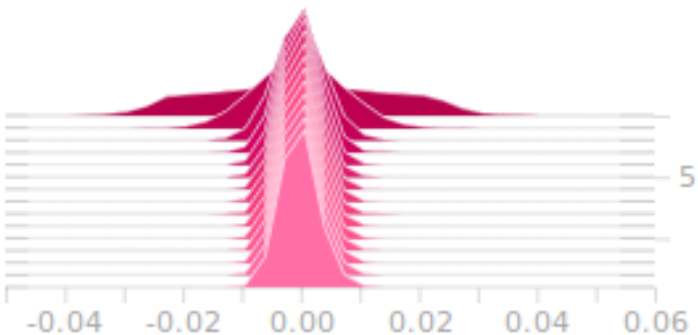
1. Dropout rate decreased it in two times.
2. l2-regularization equals to 0.01 both for convolution layers and dense layer. Dropout = 0.5 both for dense and convolution layers. training algorithm - Adam with learning rate 1e-4.
3. l2-regularization equals to = rate 1e-3. Dropout = 0.25 both for dense and convolution layers.
4. l2-regularization = 0.001 for convolution layers and 0.01 for dense layers. Dropout rate = 0.25 for convolution layers and 0.5 for dense layers.

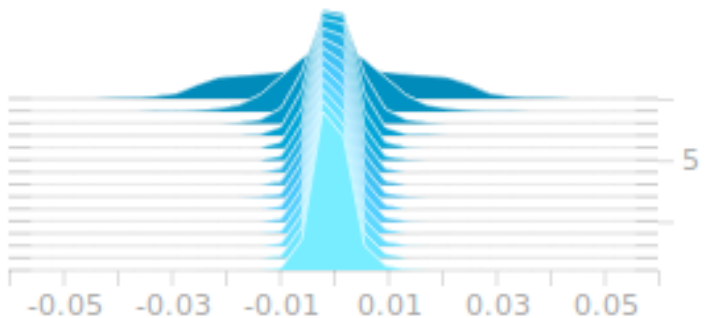
Regularizations





a, b

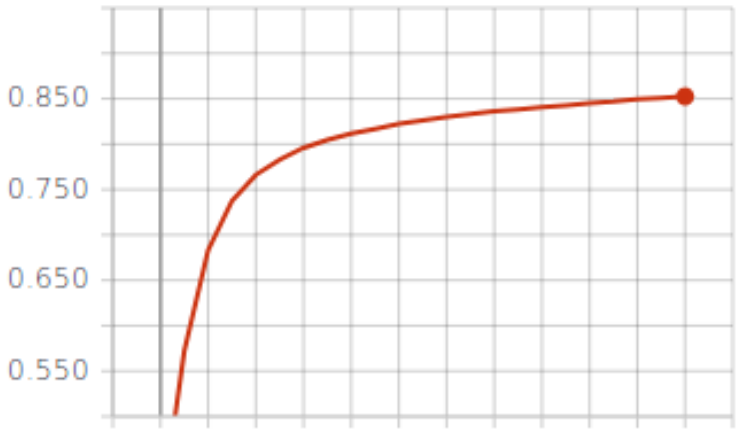




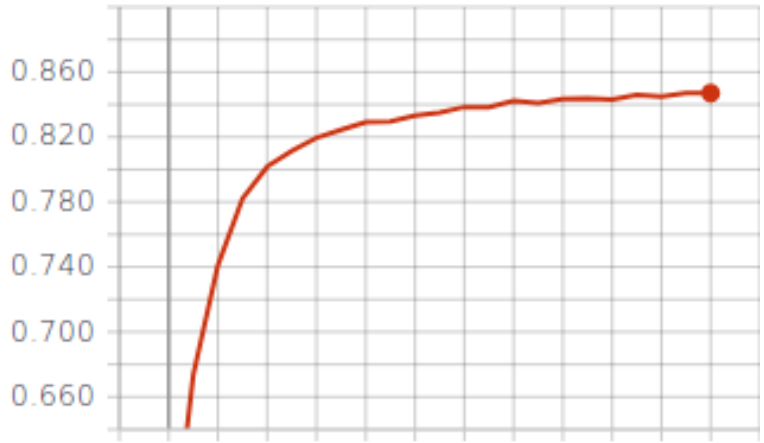
c, d

Results

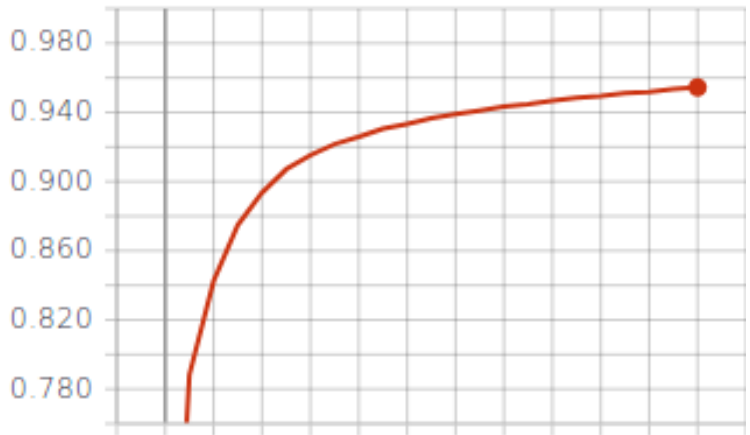
categorical_accuracy



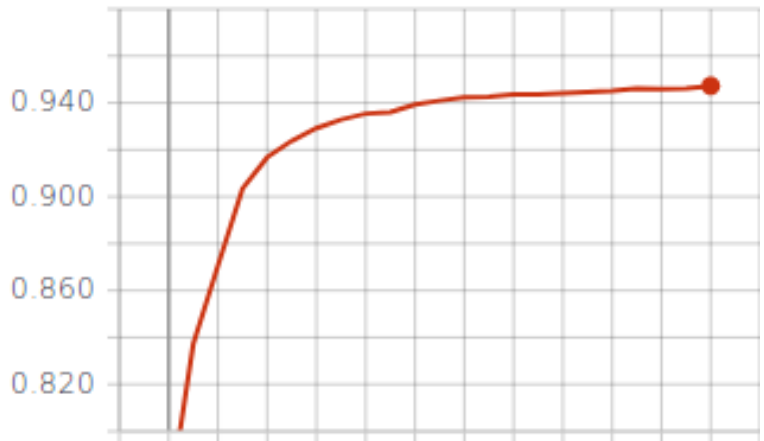
val_categorical_accuracy



top_k_categorical_accuracy



val_top_k_categorical_accuracy



Results

Metric	**Train**	**Test**
categorical accuracy	0.8250	0.8307
category crossentropy	0.5800	0.6612
top k accuracy	0.9545	0.9473

Future work

- train embeddings on the given dataset
- use pictures and attributes as additional source of information for better classification

Thank you for your attention!