Sheffield Hallam University

Text classification by Convolution Networks for Data-Driven Decision Making

RODRIGUES, Marcos <http://orcid.org/0000-0002-6083-1303>

Published version

The extremely short descriptions are significant challenges to classification. The designed network has 15 convolution layers followed by 2 fully connected layers. The network was implemented using the Torch Framework on a Mac Pro running macOS Sierra 3.5GHz 6-core Intel Xeon E5 processor with 16GB of memory. After 36 hours of training, results for unseen test data are depicted in the confusion matrix below.

bakery	allinson wholemeal batch bread medium sliced 800g
chilled	alf turner sausage roll
dairy	actimel yogurt drink 0.1% fat strawberry 8x100g
drinks	35 south latitude pinot noir 75cl
fruit &veg	birds eye field fresh select mixed vegetable 690g
meat &fish	bernard matthews breaded ham and cheese chicken escalope 285g

The extremely short descriptions are significant challenges to classification. The designed network has 15 convolution layers followed by 2 fully connected layers. The network was implemented using the Torch Framework on a Mac Pro running macOS Sierra 3.5GHz 6-core Intel Xeon E5 processor with 16GB of memory. After 36 hours of training, results for unseen test data are depicted in the confusion matrix below.

	bakery	chilled	dairy	drinks	Fruit &veg	Meat &fish	Total
Validation:							
bakery	**104**	7	2	0	0	1	114
chilled	4	**173**	7	0	5	9	198
dairy	2	3	**171**	0	0	1	177
drinks	0	2	0	**107**	0	0	109
Fruit &veg	5	2	1	0	**105**	4	117
Meat &fish	1	13	4	0	5	**123**	146
Total	116	200	185	107	115	138	861
TP	91%	87%	97%	98%	90%	84%	91%

The columns of the matrix show the predicted values, while the rows are actual values. The percentage accuracies quoted are for TP-true positives only. The largest misclassification is for meat and fish, where 13 samples were classified as chilled. The overall accuracy of 91% is impressive given that the classification features were extracted from character sequences only and that descriptions are extremely short. It is shown that

character-based classification is a valid solution for short descriptions and we are now investigating alternative optimal network designs and the possibility of using a larger training set.