Improving product classification using generative recurrent networks

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

[Athens Institute for Education and Research](#)

Abstract Submitting Form

Conference	**A Stream on "Data Science", 23-26 July 2018, Athens, Greece**
Title of Paper	Improving Product Classification using Generative Recurrent Networks

For more than one author, please copy and paste the following eight rows for each additional author.

Title	x Dr ☐ Mr ☐ Ms ☐ Other Specify: Professor
First Name	Marcos
Family Name	Rodrigues
Position	Professor of Computer Science
University/ Organization	Sheffield Hallam University
Country	UK
E-mail	m.rodrigues@shu.ac.uk
Telephone(s)	+44 (0)114 225 6911
Fax	+44 (0)114 225 6702

Abstract

The issue addressed in this paper is related to machine learning techniques for automatic classification of product descriptions. The problem arises when database entries do no perfectly match and so it is questionable whether a description is related or not to the same item, product, or service. A typical example is merging disparate databases that is required, for instance, when one business buys off a competitor. An obvious solution would be to train an AI system to perform classification. The problem is that AI deep learning networks require vast amounts of training data, normally in tens or hundreds of thousand samples and normally such data are not available. The specific classification problem we are addressing can be illustrated as follows.

Product	Category	Level 2	Level 3
Actimel Yogurt Drink 0.1% Fat Original 12x100g	Dairy	Yogurts	Actimel
Actimel Yogurt Drink 0.1% Fat Strawberry 8x100g	Dairy	Yogurts	Actimel
Actimel Yogurt Drink Blueberry 8x100g	Dairy	Yogurts	Actimel
Actimel Yogurt Drink Coconut 8x100g	Dairy	Yogurts	Actimel
Actimel Yogurt Drink Kids Strawberry and Raspberry 6x100g	Chilled	Easy Lunches	Lunchbox favourites

Note that while the first four records have been manually classified as 'Dairy', the last entry was classified as 'Chilled' (classification is accepted as correct for all entries). In order to learn the nuances of classification, an AI system needs a vast number of additional samples to be able to distinguish what characterizes Dairy and Chilled. Therefore, we have investigated network models to augment the training data set in a flexible but reliable way. The principle is to train a network with the objective of generating new data similar but not exactly the same as the input data. Validation of the newly generated data is performed by a second network which has been trained on the original data. A simple binary decision (yes/no) is output whether or not generated data has enough or acceptable similarity with the original data. Accepted data would eventually make part of an augmented training set, improving the network ability to classify unseen data. We designed and implemented a recurrent network with Keras, an open source neural network library written in Python. The network is based on the LSTM-Long-Short Term Memory model which has proved useful to a large number of problems with time dependencies. The encoding of product description is character-based so, once trained, the network outputs a character and tries to predict what the next character would be. With an appropriate training set to learn the structure of the data, such networks can output valid vectors. We set the network to train over 20 epochs outputting the description (with a limited number of characters) at the end of each epoch. At epoch 0 (before training) it can only output random characters:

```
R22QQQOOVVVV0000000aa33aKTTTTTTTTTT**eLLLePPPPCJJllmvao
```

At epoch 2, things start to get better as the net begins to learn to separate words properly:

```
X Crisps and snacks
```

```
Supermarket's Crisps and Crisps and Cream
```

At epoch 3 the data now starts to resemble the training file with one description per line (ignoring the nonsense meaning of generated data such as chicken yogurt):

```
Chilled > Fresh pasta and sauces > Fresh pasta
British Chicken and Strawberry and Corner Yogurt 4x125g|Dairy > Yogurts > Muller
British Pork Sausages x8 200g|Meat and fish > Fish and seafood > All fish and seafood
Supermarket's British Pork Light and Coconut and Cheese
```

Network outputs get increasingly better and, at the end of training, valid samples are generated for an augmented database. Note that the generated data are not the same as the original. The main outcome of such generative recurrent network is that it works for text generation, giving us the ability to generate valid data from a limited set of samples. In this paper, we provided a justification for using recurrent networks to solve a significant limitation of small data sets in deep learning. We also showed that LSTMs are a good solution to the problem together with character-based text encoding and these represent the state-of-the-art in recurrent neural networks. Future work involves improvements to the network design model and testing SimpleRNN or GRU-Gate Recurrent Unit in place of LSTMs and fine-tuning of network parameters.

Keywords	AI, Deep Learning, Recurrent Networks

Please email to: atiner@atiner.gr as an attached file or fax it to +30 210 3634209