Motivation

News feed applications are growing exponentially in the big data area. Even though there are a lot of news to explore, there are challenges to find news that meet the users’ interest. One of the major tasks for those intelligent applications is to recommend personalized feeds to users. The core of a recommendation system is usually a ML-based model that categorizes news feeds based on its content and metadata to make recommendations based on a user’s activities. This project proposes a news content classifier that classifies text news content that sets predefined labels. The input to the system is text format news content (a combination of title and content) and the output is one of the predefined labels of the news categories.

Data

The dataset is General News Category Dataset from Kaggle. The JSON file contains 202,372 records and a total of 41 distinct categories. They are JSON files with pure text attributes.

Features

We extracted headline and short_description and cateenate them together to form the input texts. The category attribute is used for labels.

Models

Models we have explored: Bidirectional RNN, Bi-Directional RNN with Attention and MT-DNN. We calculate the loss with tensorflow sparse_sofmax_cross_entropy_with_logits loss function.

<table>
<thead>
<tr>
<th>Model</th>
<th>Training Accuracy</th>
<th>Test Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bidirectional RNN Accuracy</td>
<td>0.5453</td>
<td>0.5352</td>
</tr>
<tr>
<td>Bidirectional RNN with Attention Accuracy</td>
<td>0.6228</td>
<td>0.5967</td>
</tr>
<tr>
<td>MT-DNN</td>
<td>0.3765</td>
<td>0.3948</td>
</tr>
</tbody>
</table>

Future

MT-DNN is a novel and effective model for NLP tasks. We did not have enough bandwidth and computing resources to run it with original hyperparameters. We recommend migrating it to Docker to AWS platform for future experiments.

References