Discovery and Scaling of Text Models for Authorship Attribution

Porter Bagley
Computing - DSSI
Collaborators: Juanita Ordonez, Rafael Rivera Soto
Introduction

• Motivation:
  • As more of us rely on online sources for news and information, it becomes increasingly important to verify their authenticity
  • Authorship attribution can help us determine trustworthiness

• Questions:
  • How to scale a CNN model strategically?
  • Is this the best architecture?
Methods

• Compound Scaling
  • “EfficientNet: Rethinking Model Scaling for CNNs” (Tan & Le, 2019)
  • Rather than searching for the best dimensions with a huge model, find the best dimensions on a small model, then scale up proportionally

• We apply the methods from EfficientNet to text-based data

  depth: $d = \alpha^\phi$
  width: $w = \beta^\phi$
  resolution: $r = \gamma^\phi$

  s.t. $\alpha \cdot \beta^2 \cdot \gamma \approx 2$

  $\alpha \geq 1, \beta \geq 1, \gamma \geq 1$

• Dataset
  • 100 authors from Reddit
  • Documents of 100-500 characters
  • Concatenated according to timestamp and subreddit to make 2000 & 5000 character length datasets
Results and Conclusion

• Results

<table>
<thead>
<tr>
<th>Scaling amount</th>
<th>Average Author Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0 (original)</td>
<td>0.8286</td>
</tr>
<tr>
<td>2.0</td>
<td>0.9254</td>
</tr>
<tr>
<td>4.0</td>
<td>0.9399</td>
</tr>
<tr>
<td>8.0</td>
<td>0.9601</td>
</tr>
<tr>
<td>16.0</td>
<td>0.9728</td>
</tr>
<tr>
<td>20.0</td>
<td>0.9900</td>
</tr>
</tbody>
</table>

Random search results:
Accuracies for all depths

Accuracy w.r.t. resolution

Top model results:

• Conclusion

  • Compound scaling is effective for text-based CNNs
  • Increasing the resolution of text-based data can lead to large gains in performance
  • Achieved 99.0% (+16.1%) accuracy on authorship attribution task

• Next Steps

  • Discover alternate architectures using TextNAS (Wang et al., 2019)
  • Apply similar methods to open-world authorship attribution task