1. Summary

- Conceptually simpler models like MLPs promise to be more sustainable because they are easier to train and require less data.
- We propose HyperMixer, an MLP-based neural architecture with inductive biases suited for natural language processing.
- HyperMixer is substantially better at text classification tasks than alternative MLP-based models.
- HyperMixer is less costly than Transformers in terms of processing time, training data, and hyperparameter tuning.

2. Motivation

- Simpler models promise to be less costly than MLPs.
- Existing models lack important inductive biases of Transformers: variable binding, variable length and positional invariance.

3. Model

See figure and pseudo-code at the top!

- General Transformer-like architecture: apply token mixing and feature mixing (FF-MLP) per token → variable binding
- MLP-Mixer: uses a fixed token mixing MLP to mix positions → fixed length and positional invariance

4. Experiments

Results:
1. HyperMixer performs better at text classification tasks than MLP-Mixer and similar MLP-based alternatives.
2. HyperMixer is less costly than Transformers in terms of processing time, training data, and hyperparameter tuning, while achieving competitive results.

Scope of results:
- Low-resource scenarios: relatively small models, no pretraining, medium-size datasets
- We only cover text classification datasets (no text generation) mostly from the GLUE benchmark

4.1. Comparison to other models

Test set results on 5 tasks from the GLUE benchmark [8]:

<table>
<thead>
<tr>
<th>Model</th>
<th>MNLI</th>
<th>SNLI</th>
<th>QQP</th>
<th>SST</th>
<th>[H] Params</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lin. Transformer [3]</td>
<td>81.0</td>
<td>81.0</td>
<td>78.0</td>
<td>79.0</td>
<td>11 M</td>
</tr>
<tr>
<td>Transformer [7]</td>
<td>66.6</td>
<td>66.6</td>
<td>15.2</td>
<td>15.2</td>
<td>11 M</td>
</tr>
<tr>
<td>MLP-Mixer [6]</td>
<td>62.9</td>
<td>62.9</td>
<td>70.8</td>
<td>71.8</td>
<td>11 M</td>
</tr>
<tr>
<td>gMPL [4]</td>
<td>61.2</td>
<td>61.2</td>
<td>75.0</td>
<td>76.0</td>
<td>11 M</td>
</tr>
<tr>
<td>HyperMixer (ours)</td>
<td>48.8</td>
<td>48.8</td>
<td>84.1</td>
<td>85.1</td>
<td>11 M</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Task</th>
<th>Relative Validation Acc. [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNLI</td>
<td>80.0</td>
</tr>
<tr>
<td>SNLI</td>
<td>79.0</td>
</tr>
<tr>
<td>QQP</td>
<td>79.0</td>
</tr>
<tr>
<td>SST</td>
<td>79.0</td>
</tr>
</tbody>
</table>

4.2. Cost comparison with Transformers

Cost of an AI result according to Schwartz et al. [5]:

\[\text{Cost}(C) \propto C \cdot D \cdot H \]

- \(E\): processing time
- \(D\): dataset size
- \(H\): hyperparameters

E: HyperMixer has complexity of \(O(N)\) vs Transformers \(O(N^2)\)

References


Acknowledgement

Florian Mail was supported by the Swiss National Science Foundation under grant number 32003B_178862. François Marelli was supported by the Swiss National Centre of Competence in Research (NCCR) under grant number “32432.1 IP-ICT”. Fabio Fehr was supported by the Swiss National Science Foundation under grant number “179217”. Arnaud Pannatier was supported by the Swiss Innovation Agency Innosuisse under the project MALAT, grant number “179217”.