Text Is an Image: Augmentation Via Embedding Mixing

Kevin Zhao
Mentors: Vladislav Lialin and Dr. Rumshisky
Text Data Augmentation Techniques

- Data Augmentation: create new data based on existing data, without actually collecting more data.
- In general, increasing the amount of data will improve performance of the model.

- Synonym replacement
  - The quick brown fox jumps over the lazy brown dog.
  - The quick brown fox leaps over the lazy brown dog.

- Backtranslation
Image Data Augmentation Techniques

- Basic augmentation techniques
  - Crop, reflect, rotate
- Mixing based augmentation techniques
  - Cutout, Mixup, CutMix
Image Data Augmentation Techniques

- Basic augmentation techniques
  - Crop, reflect, rotate

- Mixing based augmentation techniques
  - Cutout, Mixup, CutMix

\[
X_{mixed} = \alpha X_1 + (1 - \alpha) X_2 \\
y_{mixed} = \alpha y_1 + (1 - \alpha) y_2
\]

\[
X_{mixed} = M \circ X_1 + (1 - M) \circ X_2 \\
y_{mixed} = \alpha y_1 + (1 - \alpha) y_2
\]
Word Embeddings

- Pretrained
- Embedding Matrix

<table>
<thead>
<tr>
<th>Word</th>
<th>Embedding</th>
<th>Embedding</th>
<th>Embedding</th>
<th>Embedding</th>
</tr>
</thead>
<tbody>
<tr>
<td>aardvark</td>
<td>-0.0216, 0.0426, 0.3035, ...</td>
<td>0.1276, 0.0043, 0.0502</td>
<td></td>
<td></td>
</tr>
<tr>
<td>abacus</td>
<td>0.0815, 0.1033, -0.0632, ...</td>
<td>0.0426, -0.0103, 0.0075</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hello</td>
<td>-0.0434, -0.0914, -0.1399, ...</td>
<td>-0.0414, -0.1471, 0.3977</td>
<td></td>
<td></td>
</tr>
<tr>
<td>world</td>
<td>0.0724, -0.0536, 0.0984, ...</td>
<td>-0.0011, 0.0453, 0.0937</td>
<td></td>
<td></td>
</tr>
<tr>
<td>zyzzyva</td>
<td>0.1748, 0.0247, -0.1461, ...</td>
<td>-0.0119, 0.0618, 0.1744</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Sentence 1

Sentence 2

Cutout
Word Embeddings

- Pretrained
- Embedding Matrix
Word Embeddings

- Pretrained
- Embedding Matrix

<table>
<thead>
<tr>
<th>Word</th>
<th>Embedding Vectors</th>
</tr>
</thead>
<tbody>
<tr>
<td>aardvark</td>
<td>-0.0216, 0.0426, 0.3035, ... 0.1276, 0.0043, 0.0502</td>
</tr>
<tr>
<td>abacus</td>
<td>0.0815, 0.1033, -0.0632, ... 0.0426, -0.0103, 0.0075</td>
</tr>
<tr>
<td>hello</td>
<td>-0.0434, -0.0914, -0.1399, ... -0.0414, -0.1471, 0.3977</td>
</tr>
<tr>
<td>world</td>
<td>0.0724, -0.0536, 0.0984, ... -0.0011, 0.0453, 0.0937</td>
</tr>
<tr>
<td>zyzzyva</td>
<td>0.1748, 0.0247, -0.1461, ... -0.0119, 0.0618, 0.1744</td>
</tr>
</tbody>
</table>

Sentence 1: 768 × 256
Sentence 2: 768 × 256
CutMix: 768 × 256
<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>The movie was funny and inspiring. The book had a terrible ending.</td>
</tr>
<tr>
<td>Mixup</td>
<td>The book had a terrible ending.</td>
</tr>
<tr>
<td>CutMix</td>
<td>The book was funny and inspiring.</td>
</tr>
<tr>
<td>Cutout</td>
<td>The movie was inspiring. The book had a terrible ending.</td>
</tr>
<tr>
<td>Mixup + Cutout</td>
<td>The terrible ending.</td>
</tr>
<tr>
<td>Nonlinear Mixup</td>
<td>The book had a funny and inspiring ending.</td>
</tr>
</tbody>
</table>
RoBERTa

- Bidirectional
- Better long term memory
- Parallelism
- Pretrained

- Mixing on the static embeddings vs. contextualized embeddings
RoBERTa

- Bidirectional
- Better long term memory
- Parallelism
- Pretrained

- Mixing on the static embeddings vs. contextualized embeddings
GLUE

- 8 datasets

Results of augmentation techniques on GLUE tasks

<table>
<thead>
<tr>
<th>RoBERTa_{large}</th>
<th>CoLA</th>
<th>SST-2</th>
<th>MRPC</th>
<th>STS-B</th>
<th>QNLI</th>
<th>RTE</th>
<th>QQP</th>
<th>MNLI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reported Results</td>
<td>68.0</td>
<td>96.4</td>
<td>90.9/-</td>
<td>92.4/-</td>
<td>94.7</td>
<td>86.6</td>
<td>92.2/-</td>
<td>90.2</td>
</tr>
<tr>
<td>Our implementation</td>
<td>68.7</td>
<td>96.7</td>
<td>90.2/93.1</td>
<td>92.5/92.2</td>
<td>94.6</td>
<td>86.6</td>
<td>91.9/89.2</td>
<td>90.7</td>
</tr>
<tr>
<td>Mixup</td>
<td>71.2</td>
<td>96.7</td>
<td>90.0/93.0</td>
<td>92.3/92.2</td>
<td></td>
<td>79.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CutMix</td>
<td>70.7</td>
<td></td>
<td>89.7/92.7</td>
<td>91.8/91.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cutout</td>
<td>67.8</td>
<td></td>
<td>89.0/92.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>81.2</td>
</tr>
<tr>
<td>Mixup + Cutout</td>
<td>72.8</td>
<td>96.7</td>
<td>92.4/94.5</td>
<td>92.6/92.5</td>
<td>94.5</td>
<td>86.6</td>
<td>92.1/89.3</td>
<td>90.6</td>
</tr>
<tr>
<td>NonlinearMixup</td>
<td>65.2</td>
<td>96.1</td>
<td>90.2/93.0</td>
<td></td>
<td>94.4</td>
<td>86.6</td>
<td>91.2/88.3</td>
<td></td>
</tr>
</tbody>
</table>
Conclusion

- Computer vision augmentations such as Mixup, Cutout and CutMix are applicable to natural language processing tasks when used on a word embedding level.
- Combination of Mixup and Cutout can considerably improve the results of RoBERTa on the GLUE benchmark.
- A wider range of augmentations is applicable to NLP and potentially opens a way to a broader adoption of computer vision techniques in NLP.
Future Work

- Nonlinear Mixup
  - AdaMixup, MetaMixup, FMix, Attentive CutMix, PuzzleMix
- Scheduler for alpha
  - No longer sample mixing weight from a fixed distribution
  - Start with weaker augmentations and progress to stronger ones
Acknowledgements

Thanks to:

- My mentors Vladislav Lialin and Dr. Rumshinsky for their guidance and support
- Dr. Gerovitch and the PRIMES program for giving me this opportunity
- My parents
Questions?