Sequence Length is a Domain:
Length-based Overfitting in Transformer Models

Dušan Váriš and Ondřej Bojar

Charles University, Institute of Formal and Applied Linguistics

October 11th, 2021
Overfitting in Transformers

- Recent models (e.g. GPT-3) increase both in size and in number of training instances.
- We suspect that an overlap in the train-test data could lead to overestimation of model generalization ability.
- Long-range dependencies in transformer:
  - result of poor modeling ability (?) ...  
  - ... or lack of data with long-range dependencies?
Mock Task: String Editing

- Easier evaluation:
  - clear distinction between examples,
  - no ambiguity in correct answers,
  - accuracy metric: exact match with correct solution

<table>
<thead>
<tr>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>push 1</td>
<td>1 0 1 0</td>
</tr>
<tr>
<td>reverse</td>
<td>1 0 0 1 1</td>
</tr>
<tr>
<td></td>
<td>1 1 0 0 1</td>
</tr>
</tbody>
</table>
String Editing: Results

- Training lengths: 10-15
- train/test sequence length mismatch → the models fail horribly

<table>
<thead>
<tr>
<th></th>
<th>0-10</th>
<th>11-15</th>
<th>16-20</th>
</tr>
</thead>
<tbody>
<tr>
<td>copy</td>
<td>62.6</td>
<td>100.0</td>
<td>0.0</td>
</tr>
<tr>
<td>push</td>
<td>59.1</td>
<td>100.0</td>
<td>0.0</td>
</tr>
<tr>
<td>pop</td>
<td>0.1</td>
<td>100.0</td>
<td>0.0</td>
</tr>
<tr>
<td>shift</td>
<td>52.5</td>
<td>100.0</td>
<td>0.0</td>
</tr>
<tr>
<td>unshift</td>
<td>41.2</td>
<td>100.0</td>
<td>0.0</td>
</tr>
<tr>
<td>reverse</td>
<td>0.0</td>
<td>84.4</td>
<td>0.0</td>
</tr>
<tr>
<td>all</td>
<td>15.822</td>
<td>97.5</td>
<td>0.978</td>
</tr>
</tbody>
</table>
Machine Translation

- Split CzEng 2.0 (Kocmi et al., 2020) into buckets based on target-side (or source-side) sequence length (after subword tokenization).
- Train a separate system on each training bucket.
- Evaluate on WMT newstest split in a similar way.
Machine Translation: Results (Target-length Buckets)

![Graph showing BLEU and Hyp/Ref Ratio for different lengths of target buckets.

- **BLEU** graph:
  - X-axis: Test Bucket
  - Y-axis: BLEU score
  - Lines represent different training bucket sizes:
    - TrainBucket = 10
    - TrainBucket = 20
    - TrainBucket = 30
    - TrainBucket = 40
    - TrainBucket = 50
    - TrainBucket = 60
    - TrainBucket = 70
    - TrainBucket = 80
    - Full CzEng

- **Hyp/Ref Ratio** graph:
  - X-axis: Test Bucket
  - Y-axis: Hyp/Ref Ratio
  - Lines represent different training bucket sizes:
    - TrainBucket = 10
    - TrainBucket = 20
    - TrainBucket = 30
    - TrainBucket = 40
    - TrainBucket = 50
    - TrainBucket = 60
    - TrainBucket = 70
    - TrainBucket = 80
    - Full CzEng

The graphs illustrate how BLEU scores and Hyp/Ref Ratios change with varying lengths of target buckets for different training bucket sizes.
Create a synthetic 60-bucket data using concatenation of:
▶ 6 × 10-bucket sentences,
▶ 3 × 20-bucket sentences,
▶ 2 × 30-bucket sentences.

We concatenate consecutive sentence pairs (after shuffling).

Compare with the system trained on 60-bucket data.
Synthetic Concatenation: Results (Target-length Buckets)
Machine Translation: Source-length Buckets

![Graph showing BLEU and Hyp/Ref Ratio across different Test Buckets for various Train Buckets.](image-url)
Source-length Buckets: Target-length Distributions

Figure: **Left:** Train Distribution, **Right:** Test Distribution
Sequence Length is a Domain: Length-based Overfitting in Transformer Models

Dušan Varis and Ondřej Bojar
varis.bojar@ufal.mff.cuni.cz

Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Prague, Czech Republic

Introduction

Transformers generalize poorly to longer AND shorter sequence editing examples. Similar trends can be observed on MT task.

Methods

Split data to buckets based on target-side length. Train a separate NMT system on each training bucket and evaluate it on the validation buckets.

Results

Strong implication of target-side-length overfitting in Transformers that use absolute position encoding.

• Higher train-test length difference → higher performance drop.
• Hypothesis length similar to that of training data.
• Length overfitting could be avoided with relative position embeddings (Neishi and Yoshinaga, 2019).

Transformers with absolute position encoding output sequences of length similar to sequences in training data.

This work was supported by the GA ČR NEUREM3 grant (Neural Representations in Multi-modal and Multi-lingual Modelling, 19-26934X (RIV: GX19-26934X)), by the SSHOC grant (Social Sciences & Humanities Open Cloud, 823782) and by SVV 260 453 grant. This template is based on Mike Morrison’s idea of Better Poster. It was modified to latex by Rafael Bailo and modified for UFAL purposes by Tom Kocmi.