Exploring the Generalizability of Sequence-to-Sequence Architecture

Kumar Abhishek and Nishant Kambhatla
Motivation

➢ The ultimate aim of AI is to reach Artificial General Intelligence (AGI).

➢ Deep learning has improved the performance on many NLP tasks individually.

➢ But the generalization of NLP models remains a hard problem within an approach that focuses on the particularities of a single metric, dataset, and task.
Generalizability

- The extent to which research findings can be applied to settings other than that in which they were originally tested.

- One task → many architectures  
  One architecture → many tasks
The Encoder-Decoder Architecture

"le chat est noir" <EOS>  
[ 02 85 03 12 99 ]

Encoder

Context

Decoder

"the cat is black" <EOS>  
[ 00 42 82 16 04 ]

[ 42 82 16 04 99 ]

"the cat is black" <EOS>
The Encoder-Decoder Architecture

Encoder

Decoder

Sean Robertson, *Translation With A Sequence To Sequence Network And Attention*, PyTorch Tutorials
Encoder-Decoder with Attention

Attention Decoder

Sean Robertson, *Translation With A Sequence To Sequence Network And Attention*, PyTorch Tutorials
Encoder-Decoder with Attention

Translation Example:
Datasets
Neural Machine Translation

IWSLT ’14 dataset

- **De-En**
  - 63.9k sentence pairs for training, 930 for validation, and 1660 for testing

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>German</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>text</td>
<td>#sent.</td>
<td>$</td>
<td>W</td>
</tr>
<tr>
<td>parallel</td>
<td>63.9k</td>
<td>1.16M</td>
<td>63.1k</td>
</tr>
<tr>
<td>dev2010</td>
<td>930</td>
<td>19.1k</td>
<td>4.2k</td>
</tr>
<tr>
<td>tst2010</td>
<td>1660</td>
<td>30.3k</td>
<td>5.2k</td>
</tr>
</tbody>
</table>

Text Summarization

CNN/Daily Mail dataset.

Online news articles paired with multi-sentence (average 3.75 sentences) summaries.

<table>
<thead>
<tr>
<th></th>
<th>CNN</th>
<th></th>
<th></th>
<th>Daily Mail</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>train</td>
<td>valid</td>
<td>test</td>
<td>train</td>
<td>valid</td>
<td>test</td>
</tr>
<tr>
<td># months</td>
<td>95</td>
<td>1</td>
<td>1</td>
<td>56</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td># documents</td>
<td>90,266</td>
<td>1,220</td>
<td>1,093</td>
<td>196,961</td>
<td>12,148</td>
<td>10,397</td>
</tr>
<tr>
<td># queries</td>
<td>380,298</td>
<td>3,924</td>
<td>3,198</td>
<td>879,450</td>
<td>64,835</td>
<td>53,182</td>
</tr>
<tr>
<td>Max # entities</td>
<td>527</td>
<td>187</td>
<td>396</td>
<td>371</td>
<td>232</td>
<td>245</td>
</tr>
<tr>
<td>Avg # entities</td>
<td>26.4</td>
<td>26.5</td>
<td>24.5</td>
<td>26.5</td>
<td>25.5</td>
<td>26.0</td>
</tr>
<tr>
<td>Avg # tokens</td>
<td>762</td>
<td>763</td>
<td>716</td>
<td>813</td>
<td>774</td>
<td>780</td>
</tr>
<tr>
<td>Vocab size</td>
<td>118,497</td>
<td></td>
<td></td>
<td>208,045</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Text Summarization

Example:

Text: The BBC producer allegedly struck by Jeremy Clarkson will not press charges against the “TopGear” host, his lawyer said Friday. Clarkson, who hosted one of the most-watched television shows in the world, was dropped by the BBC Wednesday after an internal investigation by the British broadcaster found he had subjected producer Oisin Tymon “to an unprovoked physical and verbal attack.”

Summary: Producer Oisin Tymon will not press charges against Jeremy Clarkson, his lawyer says.
Dialogue Systems

Movie subtitles in English from OpenSubtitles

446,612 documents, 3.2G tokens

Examples:

- “Yeah, I want to tell him I'm okay.”
- “I can't make you believe it.”
- “We are going to the hospital.”
Image Captioning

Microsoft COCO dataset

164k images, 5 captions per image

The man at bat readies to swing at the pitch while the umpire looks on.

A large bus sitting next to a very tall building.
The Plan
The Roadmap

1. Data collection and preprocessing
2. Building an initial architecture
3. Seq2Seq experiments
4. Model optimization
5. Evaluation of the model across tasks
## Evaluation

<table>
<thead>
<tr>
<th>Tasks</th>
<th>NMT</th>
<th>Summarization</th>
<th>Dialog</th>
<th>Image Captioning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arch1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arch2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Graph: Comparing Performance Across Architectures and Tasks

- **Arch1**
- **Arch 2**

![Graph comparing performance across architectures and tasks](image)
Progress So Far
Initial Architecture
Initial Results: NMT

- En-Fr dataset from Tatoeba
  - 164,327 sentence pairs.
  - Example: “I am cold.” ↔ “J'ai froid.”

[Training]

[2019-03-26 20:57:58,586 INFO] Extracting features...
[2019-03-26 20:57:58,593 INFO] Building `Fields` object...
[2019-03-26 20:57:58,593 INFO] Building & saving training data...
[2019-03-26 20:58:01,335 INFO] * saving 0th train data shard to data/fr-en.train.0.pt.
[2019-03-26 20:58:04,846 INFO] Building & saving validation data...
[2019-03-26 20:58:04,849 INFO] Building shard 0.
[2019-03-26 20:58:04,900 INFO] Building & saving vocabulary...
Initial Results: NMT

[Demo]

SENT 696: ['You', 'should', 'apologize.]
PRED 696: Tu devrais présenter tes excuses.
PRED SCORE: -0.8661

SENT 697: ['I', 'want', 'you', 'to', 'be', 'my', 'friend', 'again.]
PRED 697: Je veux que vous soyez à nouveau mon ami.
PRED SCORE: -0.8589

SENT 698: ['It', 'looks', 'as', 'if', 'you're', 'right.]
PRED 698: On dirait que tu as raison.
PRED SCORE: -0.9138

SENT 699: ['Show', 'me', 'how', 'to', 'do', 'that.]
PRED 699: Montre-moi comment faire ça!
PRED SCORE: -0.3248

SENT 700: ['When', 'he', 'came', 'to', 'he', 'was', 'tied', 'to', 'a', 'chair', 'in', 'the', 'basement.]
PRED 700: Lorsqu'il reprit ses esprits, il était ligoté à une chaise dans la cave.
PRED SCORE: -0.0473
PRED AVG SCORE: -0.1249, PRED PPL: 1.1330
Overview

✓ Data for all the tasks:
  ✓ Collected and pre-processed

➢ Training:
  ✓ NMT
  ✓ Summarization
  ✓ Dialog
  ➢ Image Captioning

➢ Hyperparameter optimization

➢ Model evaluation
Thank You