The Role of Syntactic Planning in Compositional Image Captioning

Emanuele Bugliarello, Desmond Elliott
University of Copenhagen

EACL 2021
Image Captioning

a dog licks its lips while riding in a car
Image Captioning

Image Encoder → Text Decoder → a dog licks its lips while riding in a car
Compositional Image Captioning (Nikolaus et al., 2019)

A white cat sitting on a laptop computer

A white dog running along a beach

A big brown dog sitting on a couch

Compositional Image Captioning (Nikolaus et al., 2019)

RNN-based captioning models do not compositionally generalise

Compositional Image Captioning (Nikolaus et al., 2019)

RNN-based captioning models do not compositionally generalise

Due to the text decoder

Syntactic Planning for Compositional Generalisation
Syntactic Planning for Compositional Generalisation

Syntactic Planning for Compositional Generalisation

Reiter. *Has a consensus NL generation architecture appeared, and is it psycholinguistically plausible? INLG 1994.*
Syntactic Granularity & Planning Approaches
Syntactic Granularity & Planning Approaches

- CHUNK
- POS
- DEP
- CCG
Syntactic Granularity & Planning Approaches

- CHUNK
- POS
- DEP
- CCG
- IDLE
Syntactic Granularity & Planning Approaches

Image Encoder

Standard Decoder

Standard Decoder

Image Encoder

\(<S>\) a dog licks its lips while riding in a car \(<E>\)
Syntactic Granularity & Planning Approaches

- **Standard Decoder**: 
  
  _ chunk: a dog licks its lips while riding in a car _

- **Sequential Decoder**: 
  
  _ chunk: a dog licks its lips while riding in a car _

**Image Encoder**
Syntactic Granularity & Planning Approaches

Standard Decoder

\(<S>\) a dog licks its lips while riding in a car \(<E>\)

Sequential Decoder

\(<S>\) DET NOUN VERB PRON AUX NOUN SCONJ VERB ADP DET NOUN \(<T>\)
a dog licks its lips while riding in a car \(<E>\)

Interleave Decoder

\(<S>\) DET a NOUN dog VERB licks PRON it AUX s NOUN lips SCONJ
while VERB riding ADP in DET a NOUN car \(<E>\)
Syntactic Granularity & Planning Approaches

- **Standard Decoder**
  
  `<S>` a dog licks it s lips while riding in a car `<E>`

- **Sequential Decoder**
  
  `<S>` DET NOUN VERB PRON AUX NOUN SCONJ VERB ADP DET NOUN `<T>`
  a dog licks it s lips while riding in a car `<E>`

- **Interleave Decoder**
  
  `<S>` DET a NOUN dog VERB licks PRON it AUX s NOUN lips SCONJ while VERB riding ADP in DET a NOUN car `<E>`

- **Multi-task Decoder**
  
  `<S>` a dog licks it s lips while riding in a car `<E>`

- **Image Encoder**
Syntax Awareness

Average Recall@5 of unseen concepts by BUTD (Anderson et al., 2018)

Syntax Awareness

Average Recall@5 of unseen concepts by BUTD (Anderson et al., 2018)

Syntactic planning helps compositional image captioning

Syntax Awareness

Average Recall@5 of unseen concepts by BUTD (Anderson et al., 2018)

- Syntactic planning helps compositional image captioning
- Directly mapping an image onto words is sub-optimal

Generalisation across categories
Generalisation across categories
Generalisation across categories

SEQUENTIAL
![Graph showing generalisation across categories for SEQUENTIAL]  
- **COLORS** [R@5]
- **SIZES** [R@5]
- **VERBS** [R@5]

INTERLEAVE
![Graph showing generalisation across categories for INTERLEAVE]  
- **COLORS** [R@5]
- **SIZES** [R@5]
- **VERBS** [R@5]

MULTI-TASK
![Graph showing generalisation across categories for MULTI-TASK]  
- **COLORS** [R@5]
- **SIZES** [R@5]
- **VERBS** [R@5]
Generalisation across categories
Generalisation across categories
Results: BUTD & BUTR

- **Lang LSTM**
- **Softmax**
- **Attention**
- **TD LSTM**
- **Embedding**

**BUTD (Anderson+, 2018)**
- 9.5
- +POS
- 11.8

```
a dog sitting.
```
Results: BUTD & BUTR

<table>
<thead>
<tr>
<th>Model</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>BUTD (Anderson+, 2018)</td>
<td>9.5</td>
</tr>
<tr>
<td>+POS</td>
<td>11.8</td>
</tr>
<tr>
<td>BUTR (Nikolaus+, 2019)</td>
<td>15.0</td>
</tr>
</tbody>
</table>

Diagram showing a dog sitting with attention, ranking, and weighted pool components.
Results: BUTD & BUTR

- **BUTD (Anderson+, 2018)**
  - +POS: Score of 11.8

- **BUTR (Nikolaus+, 2019)**
  - +POS: Score of 12.0

Diagram:
- **Embedding**
- **Lang LSTM**
- **TD LSTM**
- **Softmax**
- **Attention**
- **Ranking**
- **Weighted Pool**

Sentence: 
- a dog
Results: BUTD & BUTR

Syntax hurts retrieval

<table>
<thead>
<tr>
<th></th>
<th>BUTD (Anderson+, 2018)</th>
<th>BUTR (Nikolaus+, 2019)</th>
</tr>
</thead>
<tbody>
<tr>
<td>+POS</td>
<td>11.8</td>
<td>12.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>BUTD (Anderson+, 2018)</th>
<th>BUTR (Nikolaus+, 2019)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>9.5</td>
<td>15.0</td>
</tr>
</tbody>
</table>
Results: BUTRweight

<table>
<thead>
<tr>
<th>Model</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>BUTD (Anderson+, 2018)</td>
<td>9.5</td>
</tr>
<tr>
<td>+POS</td>
<td>11.8</td>
</tr>
<tr>
<td>BUTR (Nikolaus+, 2019)</td>
<td>15.0</td>
</tr>
<tr>
<td>+POS</td>
<td>12.0</td>
</tr>
</tbody>
</table>

Diagram showing the architecture of BUTRweight with layers for Embedding, TD LSTM, Lang LSTM, Softmax, Attention, Ranking, and Weighted Pool.
Results: BUTRweight

<table>
<thead>
<tr>
<th>Model</th>
<th>Score 1</th>
<th>Score 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>BUTD (Anderson+, 2018)</td>
<td>9.5</td>
<td>11.8</td>
</tr>
<tr>
<td>+POS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BUTR (Nikolaus+, 2019)</td>
<td>15.0</td>
<td>12.0</td>
</tr>
<tr>
<td>+POS</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

```
Results: BUTRweight

<table>
<thead>
<tr>
<th>Embedding</th>
<th>TD LSTM</th>
<th>Lang LSTM</th>
<th>Softmax</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>dog</td>
<td>sitting</td>
<td></td>
</tr>
</tbody>
</table>
```

Diagram:
- Embedding
- TD LSTM
- Lang LSTM
- Attention
- Softmax
- BUTD (Anderson+, 2018): 9.5
  - +POS: 11.8
- BUTR (Nikolaus+, 2019): 15.0
  - +POS: 12.0
Results: BUTRweight

<table>
<thead>
<tr>
<th>Method</th>
<th>BUTD (Anderson+, 2018)</th>
<th>+POS</th>
<th>BUTR (Nikolaus+, 2019)</th>
<th>+POS</th>
<th>BUTRweight (ours)</th>
<th>+POS</th>
</tr>
</thead>
<tbody>
<tr>
<td>BUTD (Anderson+, 2018)</td>
<td>9.5</td>
<td></td>
<td>BUTR (Nikolaus+, 2019)</td>
<td>15.0</td>
<td>BUTRweight (ours)</td>
<td>14.9</td>
</tr>
<tr>
<td>+POS</td>
<td>11.8</td>
<td></td>
<td>+POS</td>
<td>12.0</td>
<td></td>
<td>16.4</td>
</tr>
<tr>
<td>BUTRweight (ours)</td>
<td>14.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Results: M2 Transformer (Cornia et al., 2020)

BUTD (Anderson+, 2018) 9.5
+POS 11.8

BUTR (Nikolaus+, 2019) 15.0
+POS 12.0

BUTRweight (ours) 14.9
+POS 16.4

Results: M2 Transformer (Cornia et al., 2020)

Transformers do not compositionally generalise


<table>
<thead>
<tr>
<th>Model</th>
<th>BUTD (Anderson+, 2018)</th>
<th>BUTR (Nikolaus+, 2019)</th>
<th>BUTRweight (ours)</th>
<th>M2 TRM (Cornia+, 2020)</th>
</tr>
</thead>
<tbody>
<tr>
<td>+POS</td>
<td>11.8</td>
<td>12.0</td>
<td>16.4</td>
<td>10.6</td>
</tr>
</tbody>
</table>

BUTD (Anderson+, 2018) 9.5  
BUTR (Nikolaus+, 2019) 15.0  
BUTRweight (ours) 14.9  
M2 TRM (Cornia+, 2020) 10.6
**Results: M2 Transformer (Cornia et al., 2020)**

Transformers do not compositionally generalise

Interleaving POS tags with words is model-agnostic

<table>
<thead>
<tr>
<th>Model</th>
<th>Score (+POS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BUTD (Anderson+, 2018)</td>
<td>9.5</td>
</tr>
<tr>
<td>+POS</td>
<td>11.8</td>
</tr>
<tr>
<td>BUTR (Nikolaus+, 2019)</td>
<td>15.0</td>
</tr>
<tr>
<td>+POS</td>
<td>12.0</td>
</tr>
<tr>
<td>BUTRweight (ours)</td>
<td>14.9</td>
</tr>
<tr>
<td>+POS</td>
<td>16.4</td>
</tr>
<tr>
<td>M2 TRM (Cornia+, 2020)</td>
<td>10.6</td>
</tr>
<tr>
<td>+POS</td>
<td>13.2</td>
</tr>
</tbody>
</table>

Qualitative Analysis: ride-woman

BUTD

BUTD+POS
Qualitative Analysis: ride-woman

BUTD
there is a woman that is on the floor

BUTD+POS
a woman riding a bike on a wooden floor
Qualitative Analysis: ride-woman

BUTD
there is a woman that is on the floor

BUTD+POS
a woman riding a bike on a wooden floor

a woman with a child sitting on a bench

a girl that is standing on a skateboard
Qualitative Analysis: ride-woman

BUTD
there is a woman that is on the floor

BUTD+POS
a woman riding a bike on a wooden floor

a woman with a child sitting on a bench

a girl that is standing on a skateboard

Higher quality captions with more unseen concept pairs
Conclusion
Conclusion

• Revisited traditional NLG pipeline
  • Syntactic planning stage for image captioning
Conclusion

• Revisited traditional NLG pipeline
  • Syntactic planning stage for image captioning

• Syntactic planning *consistently* improves compositional generalisation
Conclusion

• Revisited traditional NLG pipeline
  • Syntactic planning stage for image captioning

• Syntactic planning *consistently* improves compositional generalisation
  • With no loss for standard metrics (see paper)
Conclusion

• Revisited traditional NLG pipeline
  • Syntactic planning stage for image captioning

• Syntactic planning *consistently* improves compositional generalisation
  • With no loss for standard metrics (see paper)
  • In both RNN- and Transformer-based models
Conclusion

• Revisited traditional NLG pipeline
  • Syntactic planning stage for image captioning

• Syntactic planning *consistently* improves compositional generalisation
  • With no loss for standard metrics (see paper)
  • In both RNN- and Transformer-based models

• Future work
Conclusion

• Revisited traditional NLG pipeline
  • Syntactic planning stage for image captioning

• Syntactic planning *consistently* improves compositional generalisation
  • With no loss for standard metrics (see paper)
  • In both RNN- and Transformer-based models

• Future work
  • Syntax-aware captioning models
Conclusion

• Revisited traditional NLG pipeline
  • Syntactic planning stage for image captioning

• Syntactic planning *consistently* improves compositional generalisation
  • With no loss for standard metrics (see paper)
  • In both RNN- and Transformer-based models

• Future work
  • Syntax-aware captioning models
  • Fine-grained syntactic tags (e.g. CCG)
Conclusion

• Revisited traditional NLG pipeline
  • Syntactic planning stage for image captioning

• Syntactic planning *consistently* improves compositional generalisation
  • With no loss for standard metrics (see paper)
  • In both RNN- and Transformer-based models

• Future work
  • Syntax-aware captioning models
  • Fine-grained syntactic tags (e.g. CCG)

Code, models & data available online at [https://github.com/e-bug/syncap](https://github.com/e-bug/syncap)