Regularizing RNNs for Caption Generation by Reconstructing The Past with The Present
–Supplementary Material

1. Image Captioning Examples

As shown in Fig. 1, we provide more examples generated by the attentive encoder-decoder models with and without our ARNet. Note that the model fine-tuned with our proposed ARNet gives more detailed and meaningful descriptions, such as the words or phases “soccer”, “fireplace”, “black and white photo”, and so on.

<table>
<thead>
<tr>
<th>Images</th>
<th>Generated Captions</th>
<th>Ground Truth Captions</th>
</tr>
</thead>
</table>
| ![Image](image1.png) | **Attentive Encode-Decoder:**
a bus that is sitting in the street.  
**Attentive Encode-Decoder-AR:**
a white bus **driving down a street next to a building.** | 1. a black and white bus some bushes and building.  
2. a white decorated bus is next to a building.  
3. a large white bus that is by a building.  
4. a large bus parked in a parking lot.  
5. a white bus driving past a tall building. |
| ![Image](image2.png) | **Attentive Encode-Decoder:**
a group of people standing in a field.  
**Attentive Encode-Decoder-AR:**
a group of **young children** playing with soccer. | 1. a man instructing a group of kids on a soccer field.  
2. dad coaches talking to the little soccer team players on the field.  
3. a soccer coach instructing the children on the field.  
4. a group of young children standing around a field.  
5. pair of adult males with group of small children with soccer balls. |
| ![Image](image3.png) | **Attentive Encode-Decoder:**
a living room with a flat screen tv.  
**Attentive Encode-Decoder-AR:**
a living room with a television and a **fireplace.** | 1. interior of a living room with furniture, plant, fireplace and a tv.  
2. white furniture and fireplace with a tv over it decorate this living room.  
3. a chair and a couch in a room.  
4. a tv mounted above a fireplace in a nicely furnished living room.  
5. a tv sitting above a fire place in a living room. |
| ![Image](image4.png) | **Attentive Encode-Decoder:**
a display case filled with lots of donuts.  
**Attentive Encode-Decoder-AR:**
a display case filled with lots of **different types** of donuts. | 1. a bakery with boxes of donuts and bread.  
2. a selection of donuts and pastries at an oriental bakery.  
3. this is a display of donuts on a couple shelves.  
4. assorted bakery goods are on display in a cabinet.  
5. fifteen different varieties of doughnuts in a display case. |
| ![Image](image5.png) | **Attentive Encode-Decoder:**
two men riding on a horse drawn carriage.  
**Attentive Encode-Decoder-AR:**
a **black and white photo** of a man riding a horse drawn carriage. | 1. a man on a cart racing on a back of a horse.  
2. a man is riding in a horse drawn carriage.  
3. a man rides behind a horse during a race.  
4. the man is driving the horse fast.  
5. a black and white photo of a horse running on a track with a man being pulled. |

Figure 1. Image captions generated by the attentive encoder-decoder with and without our proposed ARNet, along with their corresponding ground truth captions.
package org.apache.lucene.search.spell;

/**
 * SuggestWord, used in suggestSimilar method in SpellChecker class.
 * Default sort is first by score, then by frequency.
 * *
 */
public final class SuggestWord{
    public float score;
    public int freq;
    public String string;
}

2. Code Captioning

Code captioning was proposed in ReviewNet which used HabeaCorpus dataset. As illustrated in Fig.2, the aim is to produce a condensed representation of the source code file\(^1\) that captures its core meaning.

2.1. Evaluation Scores

As SPICE and CIDEr metrics are proposed specially for evaluating image captioning results, they are not suitable for the task of code captioning. The performances on the HabeaCorpus dataset are illustrated in Table 1. It can be observed that our proposed ARNet can significantly boost the performance.

Table 1. Performance comparison on the testing split of the HabeaCorpus dataset. The best results among all models are highlighted in boldface.

<table>
<thead>
<tr>
<th>Model Name</th>
<th>BLEU-1</th>
<th>BLEU-2</th>
<th>BLEU-3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Review Net</td>
<td>0.192</td>
<td>0.105</td>
<td>0.074</td>
</tr>
<tr>
<td>Encoder-Decoder</td>
<td>0.183</td>
<td>0.093</td>
<td>0.063</td>
</tr>
<tr>
<td>Encoder-Decoder + Zoneout</td>
<td>0.182</td>
<td>0.080</td>
<td>0.063</td>
</tr>
<tr>
<td>Encoder-Decoder + Scheduled Sampling</td>
<td>0.186</td>
<td>0.098</td>
<td>0.067</td>
</tr>
<tr>
<td>Encoder-Decoder + ARNet</td>
<td>0.196</td>
<td>0.107</td>
<td>0.075</td>
</tr>
<tr>
<td>Attentive Encoder-Decoder</td>
<td>0.228</td>
<td>0.140</td>
<td>0.106</td>
</tr>
<tr>
<td>Attentive Encoder-Decoder + Zoneout</td>
<td>0.227</td>
<td>0.140</td>
<td>0.105</td>
</tr>
<tr>
<td>Attentive Encoder-Decoder + Scheduled Sampling</td>
<td>0.229</td>
<td>0.142</td>
<td>0.108</td>
</tr>
<tr>
<td>Attentive Encoder-Decoder + ARNet</td>
<td><strong>0.255</strong></td>
<td><strong>0.173</strong></td>
<td><strong>0.139</strong></td>
</tr>
</tbody>
</table>

2.2. Examples

As shown in Fig. 3, Fig. 4, and Fig. 5, we provide examples of code captioning, with side-by-side comparisons of the ground truth captions and the captions produced by our ARNet.

\(^1\)https://lucene.apache.org/core/3_6_2/api/all/org/apache/lucene/search/spell/SuggestWord.html
Figure 3. Code captions generated by attentive encoder-decoder and attentive encoder-decoder-ARNet. Compared to the caption generated by the vanilla attentive encoder-decoder model, the model with our ARNet completely predicts the meaning of this code.
Figure 4. Code captions generated by attentive encoder-decoder and attentive encoder-decoder-ARNet. Red denote incorrect tokens in the generated captions. As we can see, the attentive encoder-decoder model with our ARNet achieve more accuracy than the vanilla one.
Figure 5. Code captions generated by attentive encoder-decoder and attentive encoder-decoder-ARNet. The attentive encoder-decoder model with our ARNet generates most of the true tokens in the ground truth caption successfully and achieves better performance than the vanilla model obviously.