1. DCNet

An overview of our DCNet is shown in Figure 2. We consider the existing caption as a "noisy" caption, and wish to encode it into a compressed representation and decode the compressed representation to the desired output. Notably, an LSTM-based de-noising auto-encoder is equivalent to a Sequence-to-Sequence model.

For the encoder, we encode the noisy existing caption using a bi-directional LSTM and set the dimension of each direction to 512. Specifically, an existing sequence of \( N \) words is first converted into a sequence of word vectors using an embedding layer:
\[
[w_1^S, w_2^S, \ldots, w_N^S]
\]
where \( w_t^S \in \mathcal{R}^{1024} \), and serve as input to a bi-directional LSTM:
\[
\overrightarrow{e}_t = Bi-LSTM (\overrightarrow{e}_{t-1}, w_t)
\]
\[
\overleftarrow{e}_t = Bi-LSTM (\overleftarrow{e}_{t+1}, w_t)
\]
(1)
(2)

Which results in a matrix \( E = [(\overrightarrow{e}_1; \overleftarrow{e}_1) \ldots (\overrightarrow{e}_N; \overleftarrow{e}_N)] \), where ; indicates concatenation and \( E \in \mathcal{R}^{1024} \). The last hidden states of the bi-directional LSTM are concatenated and fed into a single feed-forward layer with tanh activation function:
\[
E_N = \tanh (W_O \cdot [\overrightarrow{e}_N; \overleftarrow{e}_N])
\]
(3)

For the decoder, we use the Top-Down decoder [1] and set the dimension size to 1024. The input to the Attention-LSTM \( x_t^1 \) comprise of the current word embedding, the previous hidden state of the language LSTM and the last encoder hidden state, such that \( x_t^1 = [w_t; E_N; h_{t-1}^L] \). The output of the attention LSTM is used to compute an attention vector over the textual features \( E \):
\[
c_t^E = \sum_{i=1}^{N} \alpha_t E_i
\]
(4)

where
\[
\alpha_t = \text{softmax} \left( w_t^T \tanh (W_g E + W_h h_t^L) \right)
\]
(5)

![Figure 1. Visualization of the selected words from SCMA](image)
Figure 2. An overview of our DCNet sub-module, which is an LSTM-based de-noising autoencoder.

Table 1. Performance of our Single Model on the Online COCO Testing Server, where B-N, M, R, and C are short for BLEU-N, METEOR, ROUGE-L, and CIDEr-D. All values are reported as percentage (%).

<table>
<thead>
<tr>
<th>Model</th>
<th>B-1</th>
<th>B-2</th>
<th>B-3</th>
<th>B-4</th>
<th>M</th>
<th>R</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metric</td>
<td>c5</td>
<td>c40</td>
<td>c5</td>
<td>c40</td>
<td>c5</td>
<td>c40</td>
<td>c5</td>
</tr>
<tr>
<td>SCST [3]</td>
<td>78.1</td>
<td>93.7</td>
<td>61.9</td>
<td>86.0</td>
<td>47.0</td>
<td>75.9</td>
<td>35.2</td>
</tr>
<tr>
<td>LSTM-A [5]</td>
<td>78.7</td>
<td>93.7</td>
<td>62.7</td>
<td>86.7</td>
<td>47.6</td>
<td>76.5</td>
<td>35.6</td>
</tr>
<tr>
<td>StackCap [2]</td>
<td>77.8</td>
<td>93.2</td>
<td>61.6</td>
<td>86.1</td>
<td>46.8</td>
<td>76.0</td>
<td>34.9</td>
</tr>
<tr>
<td>Up-Down [1]</td>
<td>80.2</td>
<td>95.2</td>
<td>64.1</td>
<td>88.8</td>
<td>49.1</td>
<td>79.4</td>
<td>36.9</td>
</tr>
<tr>
<td>CAVP [6]</td>
<td>80.1</td>
<td>94.9</td>
<td>64.7</td>
<td>88.8</td>
<td>50.0</td>
<td>79.7</td>
<td>37.9</td>
</tr>
<tr>
<td>SGAE [4]</td>
<td>80.6</td>
<td>95.0</td>
<td>65.0</td>
<td>88.9</td>
<td>50.1</td>
<td>79.6</td>
<td>37.8</td>
</tr>
<tr>
<td>ETN (Ours)</td>
<td>80.3</td>
<td>94.7</td>
<td>64.8</td>
<td>88.8</td>
<td>50.2</td>
<td>79.9</td>
<td>38.3</td>
</tr>
</tbody>
</table>

The output of the Attention-LSTM and the context vector $c_t^e$ serve as input to the Language-LSTM, such that: $x_t^e = [h_t^1; c_t^e]$. The output of the Language LSTM $h_t^2$ is then fed to the output layer which predicts a word from the vocabulary.

2. SCMA Decision Visualization

In this section, we include more results to visualize the decisions from the SCMA mechanism. Figure 1 shows the alignment plot between the existing caption and the generated caption.

3. Single Model Results on COCO Online Testing Server

We submitted our results evaluated on the official MSCOCO testing set to the online testing server. Table 1 shows the performance of our Single Model (which includes EditNet and DCNet) on the Online COCO Test Server. For fair comparison, we directly compare our method (ETN) with other state-of-art methods which also report the scores of their single model on the Online COCO Test Server (e.g. SGAE [4]). Note that the scores reported for SGAE are for the single model (and not ensemble).

References


