Partial Off-policy Learning: Balance Accuracy and Diversity for Human-Oriented Image Captioning
- Supplementary Material

Jiahe Shi  Yali Li*  Shengjin Wang
Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
shijhl8@mails.tsinghua.edu.cn  {liyal13, wgsgj}@mail.tsinghua.edu.cn

1. Accuracy-Diversity Balance of Human Performance

In our paper, we evaluate human performance based on the annotations within the MSCOCO dataset. To be specific, we follow the implementation in [16, 17] where the leave-one-out CIDEr score is used as an evaluation for accuracy. Mathematically, the score is calculated as:

\[ \text{CIDEr}(G(I)) = \frac{1}{|G(I)|} \sum_{s \in G(I)} \text{CIDEr}^*(s), \]

\[ \text{CIDEr}^*(s) = \frac{1}{|G(I)|} \sum_{j \in G(I)} \cos(g_n(s), g_n(j)) \]

We notice that human acquires CIDEr score far lower than the state-of-the-art image captioning methods. This issue may be caused by some inaccurate annotations within the dataset. Recently, Wang et al. [18] reports a similar issue that some of the annotations are of "low qualities", i.e. acquire CIDEr score significantly lower than other annotations of the image. To provide a comprehensive and fair evaluation of human performance, we exclude the offbeat annotations from evaluating human performance. Specifically, we calculated \( \text{CIDEr}^*(s) \) for each \( s \in G(I) \). The annotation \( s \) with relatively lower \( \text{CIDEr}^*(s) \) among the 5 ground truths \( G(I) \) is regarded as the semantically different one. By excluding the most or the most two different annotations of the image, we derive a curve indicating human performance on accuracy-diversity balance in Fig. 1.

In Fig. 1, we gradually exclude the most different annotations according to the leave-one-out CIDEr score. With offbeat sentences excluded from the evaluation, the remaining sentences become semantically similar to each other, resulting in high accuracy scores and relatively low diversity performance. Note that our method with \( \epsilon \) set as 0.1 still locates the closest to the curve of human performance.

![Figure 1. Accuracy-Diversity balance of human performance by excluding certain offbeat annotations. Model trained by the proposed partial off-policy strategy with \( \epsilon \) set as 0.1 is still the closest to the curve of human performance.](image)

<table>
<thead>
<tr>
<th>Model</th>
<th>( |\rho_{\text{all}}| )</th>
<th>( |\rho_{\text{ex}-1}| )</th>
<th>( |\rho_{\text{ex}-2}| )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Att2in[10]</td>
<td>0.243</td>
<td>0.189</td>
<td>0.132</td>
</tr>
<tr>
<td>Up-Down[2]</td>
<td>0.253</td>
<td>0.200</td>
<td>0.139</td>
</tr>
<tr>
<td>AdaAtt[9]</td>
<td>0.236</td>
<td>0.181</td>
<td>0.123</td>
</tr>
<tr>
<td>ReTrans[6]</td>
<td>0.295</td>
<td>0.238</td>
<td>0.170</td>
</tr>
<tr>
<td>AoA[8]</td>
<td>0.289</td>
<td>0.232</td>
<td>0.171</td>
</tr>
<tr>
<td>CVAE[15]</td>
<td>0.200</td>
<td>0.168</td>
<td>0.138</td>
</tr>
<tr>
<td>GMM-CVAE[15]</td>
<td>0.280</td>
<td>0.181</td>
<td>0.129</td>
</tr>
<tr>
<td>CapGAN[11]</td>
<td>0.193</td>
<td>0.157</td>
<td>0.102</td>
</tr>
<tr>
<td>Ours (( \epsilon = 0.1 ))</td>
<td><strong>0.337</strong></td>
<td><strong>0.304</strong></td>
<td><strong>0.255</strong></td>
</tr>
</tbody>
</table>

Table 1. Pearson’s correlation coefficient between predicted captions and subsets of human annotations, where \( \text{ex}-1 \) denotes excluding the most different annotations from evaluation and \( \text{ex}-2 \) denotes excluding the most two different annotations.

Accordingly, we also report Pearson’s correlation coefficient between model-generated captions and the subset of human annotations in Table 1. Our method achieves correlation coefficients as 0.304 and 0.255 respectively excluding...
Figure 2. The distribution of the performance acquired by different methods in the Diversity-Accuracy space. We sample 100 images randomly from Karpathy’s validation set for visualization. Compared with the on-policy trained baseline, our approach not only locates closer to human performance but also manages to mimic the human distribution.

the most or the most two different annotations, which is still the best performance compared with other methods.

The curve of human performance demonstrated in Fig. 1 also acts as the performance boundary for which works on the accuracy-diversity balance issue are supposed to approach. Currently, diversity performance derived by the proposed partial off-policy strategy declines significantly when accuracy increases. However, since our work is agnostic to model structure, we would expect to narrow the gap by deploying the proposed learning scheme on models with better accuracy in the future.

2. What does Pearson’s Correlation Coefficient Stand for?

We use Pearson’s correlation coefficient as the quantitative evaluation in our paper. It actually serves as a complement to the diversity-accuracy diagram. In the diagram, we locate closest to human performance, indicating that we approach human performance in general. However, we are also concerned about the quality of the modeled posterior in detail, as illustrated in Fig. 2.

By acquiring a higher Pearson’s correlation coefficient, the performance of our method varies over different image inputs in a similar manner compared with humans. In other words, models trained by the proposed scheme behave well where humans give better annotations and behave relatively poorly where humans fail to provide decent captions too. This indicates that we fully utilize the supervision annotations and succeed in modeling a human-like posterior.

Furthermore, we would like to emphasize that the correlation coefficient cannot represent the quality of the posterior by itself. It is supposed to be combined with the diversity-accuracy diagram for a comprehensive evaluation, where the diagram illustrates that our method achieves human-like performance from a holistic perspective, and Pearson’s correlation coefficient illustrates such a conclusion considering the detail characteristic of the posterior.

3. Additional Quantitative Results

More Comparison Results with Sampling Methods. Our method is a training-side improvement which differs from sampling methods like temperature scaling, top-k sampling [4], nucleus sampling [7], and DBS [14]. In our paper, we provide a comparison with DBS [14]. We would like to provide more comparison results with the above-mentioned sampling strategies on the balance effect of accuracy and diversity. The result is demonstrated in Fig. 3, where different sampling methods are performed on both CE-trained model and on-policy RL-trained model.

Our method forms an upper envelope over the sampling approaches in Fig. 3, which suggesting better performance. This is because that the sampling methods achieve balance effects based on manipulation over a learned posterior, while our method manages to derive a posterior of better quality with the training-side improvements.

Comparison with Other On-policy Exploration Methods. Compared with the mainstream on-policy approaches in image captioning, our method can be viewed as an exploration-enhancing strategy. There is also some literature focusing on the exploration issue under the on-policy framework, i.e., distributional entropy regularization [5] and count-based state visitation [12]. However, as we mentioned in our paper, such traditional exploration strategies may not perform well in image captioning. To illustrate
such an issue, we conduct experiments and observe performance degradation on both diversity and accuracy. The experimental result can be seen in Fig. 4.

The ineffectiveness of traditional exploration methods may due to the enormous searching space in the task of image captioning. The action space of image captioning in each step consists of 9,487 different actions (words). Typically the maximum length of each trajectory is set as 16. Therefore there are up to \(9487^{16} \approx 4.3 \times 10^{63}\) different states within the searching space. Simply encouraging exploration indiscriminately in such searching space may be inefficient and hinder the model from acquiring better performance. On the contrary, our method restricts the searching space by introducing the sampling model, which circumvents such an issue.

**Does Weighted Combination of Reward Components Works?** As we mentioned in our paper, the proposed partial off-policy learning scheme can be interpreted under the scheme of multi-objective reinforcement learning. An ordinary solution to such problems is to use a weighted combination of the multiple rewards as the optimization target. We conduct experiments to provide a comparison between our method and the above-mentioned solution. The results can be found in Fig 5.

Our method acquires better performance than the weighted combination solutions. This is because in our work we not only optimize the additional reward \(\text{max-CIDEr}\) but also arrange a specific way (i.e. the off-policy strategy) to generate appropriate training trajectories. Thus, the goal is optimized more effectively.

**Other combination of diversity and accuracy evaluation.**

In our paper, we evaluate the proposed method using CIDEr [13] as accuracy evaluation and self-CIDEr [16] as diversity evaluation. We would like to present more evaluation results with different combinations of diversity-accuracy metrics for comprehensive illustration. For accuracy scores, we report METEOR [3] in addition to CIDEr, for its comprehensive evaluation on both recall and precision. We also report SPICE [1] for its insensitivity on n-gram overlap. For diversity evaluations, we report self-CIDEr and mBLEU-4. Since lower mBLEU-4 represents more diversity, we reverse it as \(100 - \text{mBLEU-4}\) so that it is consistent with self-CIDEr. Moreover, since mBLEU-4 varies dramatically, we plot the figure using logarithmic coordinates for more details. The results are illustrated in Fig. 6, where our method locates close to human performance according to all metric selections.

We also compute the Pearson’s correlation coefficient using different combinations of metrics. The results are shown in Table 2. We achieve the best correlation with human scores according to most selections of metrics. For example, our method obtains correlation coefficients as 0.324, 0.260, and 0.333 respectively using METEOR, BLEU-4, and SPICE for accuracy evaluation and self-CIDEr as diversity metric, which exceeds the performance of other methods by a significant margin. We also notice that some methods outperform ours when mBLEU-4 is chosen to represent diversity. This may due to the significant variation of mBLEU-4 as reported by [16]. Since we normalize the diversity and accuracy scores respectively before calculating correlation coefficients, the influence of diversity may be greatly reduced due to the large variance of mBLEU-4, leading the coefficients to incline towards accuracy more. To provide fair evaluation, we calculate the correlation co-
efficient using \( \exp(100 - \text{mBLEU-4}) \) in Table 3, as the logarithmic coordinates tends to demonstrate more details in Fig. 7. Our method achieves the best performance compared with other works under such implementation.

### 4. Additional Qualitative Results

We also present more samples for qualitative evaluation. The results are shown in Fig. 7. Descriptive semantics which the on-policy baseline omits are highlighted. Moreover, we provides some samples in the last row of Fig. 7 where incorrect descriptions are predicted, which results in decline of accuracy performance. The incorrect concepts include false attribute (e.g., "white" of the left image), false relation (e.g., "reach" of the middle image) and false object (e.g., "carrot" of the right image). We expect future works to deal with such cases.

## References


[10] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for


Figure 6. Accuracy-Diversity balance according to multiple combinations of metrics. Our method shows human-oriented performance over different metric selections.
Figure 7. More qualitative results.