Introduction

We systematically evaluate the role of different choices – training objectives, hyperparameter values, sampling/decoding procedure – in the resulting tradeoff between accuracy and the diversity of generated captions. In addition, we introduce AllSPICE, a new metric for evaluating caption set on both accuracy and diversity.

AISPIEC

\[
\text{AISPIEC} = \frac{\sum_{s \in S} \text{ISPIEC}(s)}{|S|} \quad \text{where ISPIEC}(s) = \frac{\sum_{r \in R} \text{ISPIEC}(r)}{|R|}
\]

where ISPIEC represents the accuracy score of a given sentence.

Properties

1. \( \text{AISPIEC}(S, S) = 2 \times F(S, S) \times (1 - F(S, S)) \)

2. \( \text{P.S.R}(S, S) = 2 \times F(S, S) = 1 - F(S, S) \)

Is RL-trained model really bad at diversity?

Previous work evaluates model accuracy/diversity tradeoff by running random sampling with temperature 1. Doing so, the result would be:

- Cross entropy loss(XE) trained model get low accuracy but high diversity.
- RL (or specifically self-critical sequence training) would achieve high accuracy, but every sample would be very similar.

Therefore interpolating RL objective and XE objective was proposed to achieve better tradeoff.

However, a simple alternative for trading diversity for accuracy (or vice versa) is to modulate the sampling temperature.

Different sampling methods

Random sampling with \( T = 0.9 \) outperforms better than other settings on AllSPICE, so no need to carefully tune the XE-RL weight in XE+RL method.

Biased sampling is marginally better than random sampling. Benefits are more prominent when trained with RL.

Beam search in different from sampling methods, higher temperature leads to less diverse set. However, due to the expanding nature, beam search is generally less diverse.

Comparison between methods(XE):

- Diverse beam search is the best algorithm with high AllSPICE and Self-CIDEr, including both semantic and syntactic diversity.
- Beam search performs best on oracle CIDEr and average CIDEr, and it performs well on AllSPICE too. However although all the generated captions are accurate, the syntactic diversity is missing, shown by Self-CIDEr.
- Sampling methods (RS, Top-K, Top-p) are reasonably competitive. (And they are also fast)

Qualitative results

Sample size

Oracle CIDEr tends to increase with sample size, because more captions mean more chances to fit the reference.

AISPIEC drops with more samples, because additional captions are more likely to hurt (say something wrong) than help (say something correct but not said).

Average CIDEr Sampling methods' average scores are largely invariant to sample size. BS and especially DBS suffer a lot with more samples, because diversity constraints and the properties of the beam search force the additional captions to be lower quality.

Conclusion

- Simple random sampling, coupled with rather low temperature, is competitive with the best previously proposed decoding methods in terms of speed and diversity accuracy tradeoff.
- Diverse beam search exhibits the best tradeoff, but is also the slowest.

- Decoding parameters, in particular temperature, affect the resulting diversity/accuracy tradeoff more significantly than the choice of training objectives.

- Using CIDEr-based reward is detrimental to the diversity properties of the resulting generator, reducing diversity in a way that is not mitigated by manipulating decoding parameters.

- Finally, we introduce AllSPICE, a new metric that reflects both accuracy and diversity of caption sets.