TOYOTA TECHNOLOGICAL INSTITUTE AT CHICAGO

ANALYSIS OF DIVERSITY-ACCURACY TRADEOFF IN IMAGE CAPTIONING

Ruotian Luo — *TTI-Chicago* — rluo@ttic.edu

Greg Shakhnarovich — *TTI-Chicago* — greg@ttic.edu

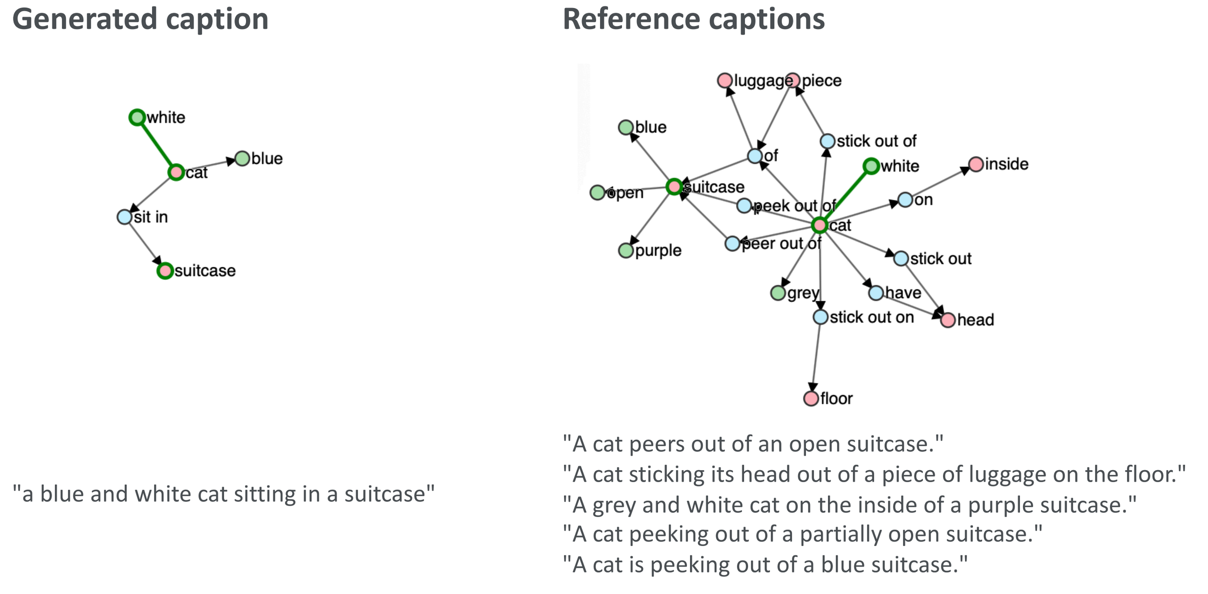
Introduction

We systematicly evaluate the role of different choices – training objectives; hyperparameter values; sampling/decoding procedure – play in the resulting tradeoff betweeen accuracy and the diversity of generated caption sets.

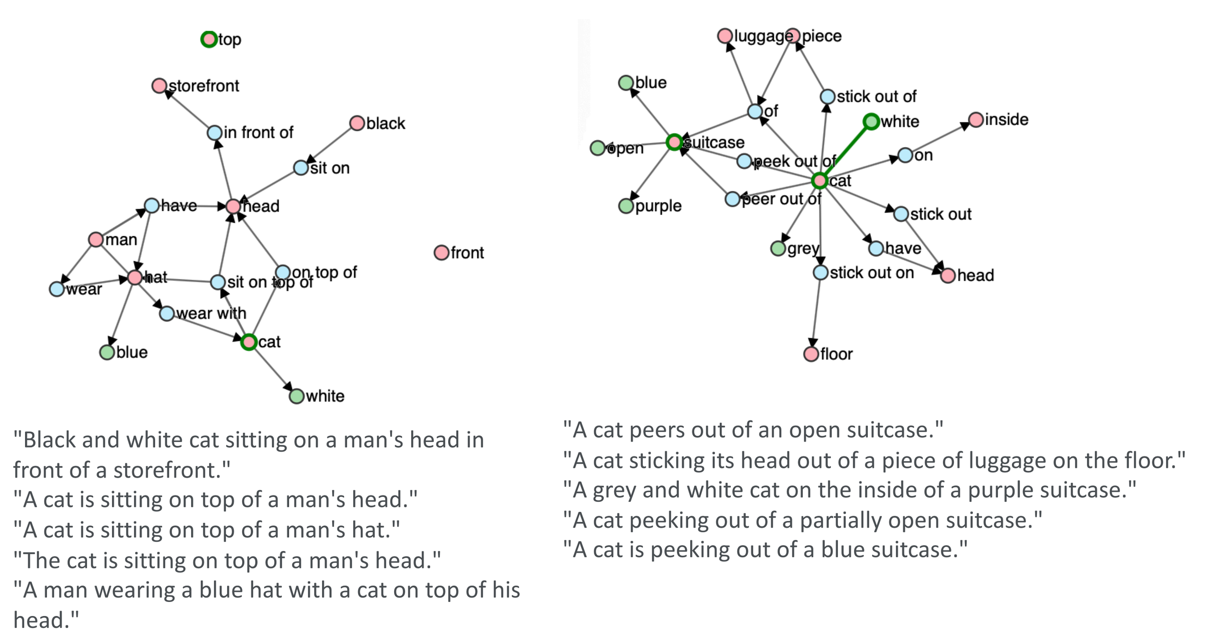
In addition, we introduce AllSPICE, a new metric for evaluating caption set on both accuracy and diversity.

AllSPICE

SPICE:



AllSPICE:



$$AllSPICE(S, S^*) = F_1(S, S^*) = \frac{2 \cdot P(S, S^*) \Delta R(S, S^*)}{P(S, S^*) + R(S, S^*)}$$

Properties:

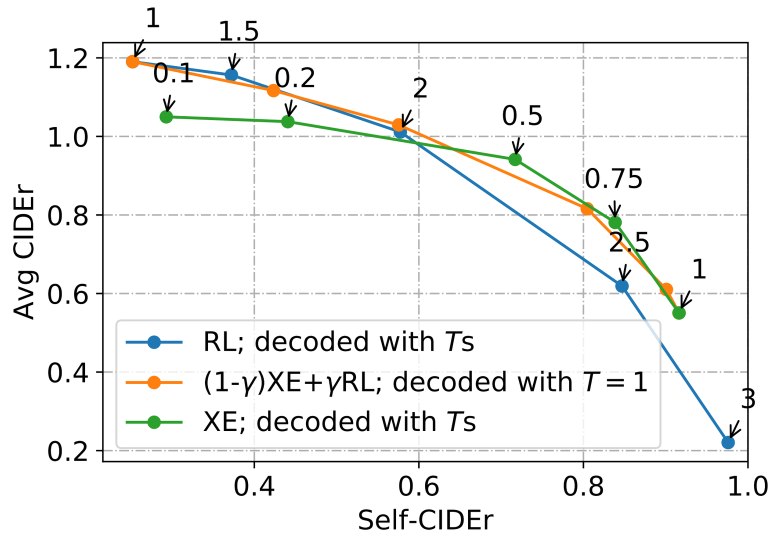
- repetitions across captions in the set won't change the score, because during the scene graph generation, synonymous vertices are merged.
- adding a caption that captures part of the reference not captured by previous captions in the set may improve the score (by increasing recall). This encourages semantic diversity.
- wrong content in any caption in the set will harm the score (by reducing precision). This encourages accuracy of the whole sets.
  (In contrary, oracle scores only require one caption in the set to be good)

Is RL-trained model really bad at diversity?

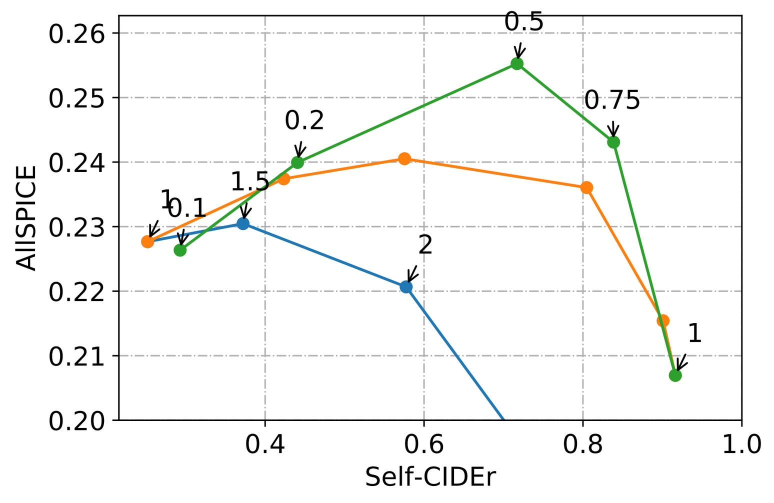
Previous work evaluates model accuracy/diversity tradeoff by running random sampling with temperature 1. Doing so, the result would be:

- 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.

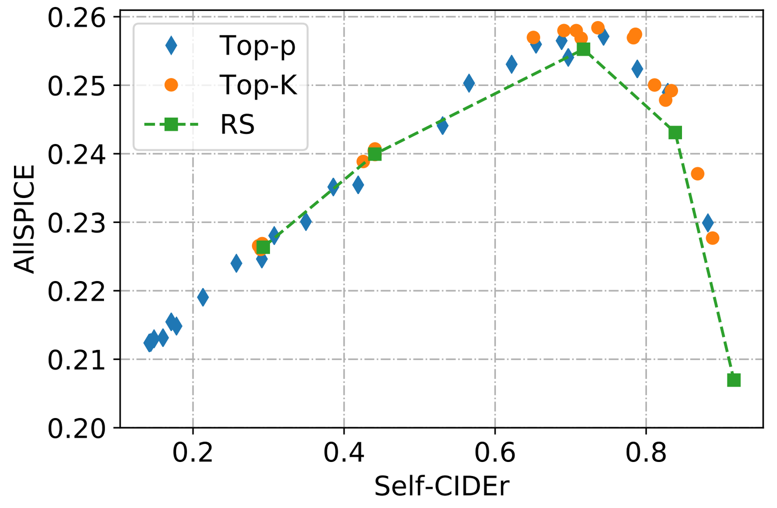
Therefore interpolating RL objective and XE objective was proposed to achieve better tradeoff. However, a simple alternative for trading diversity for accuracy (or vice versa) is to **modulate the sampling temperature**.



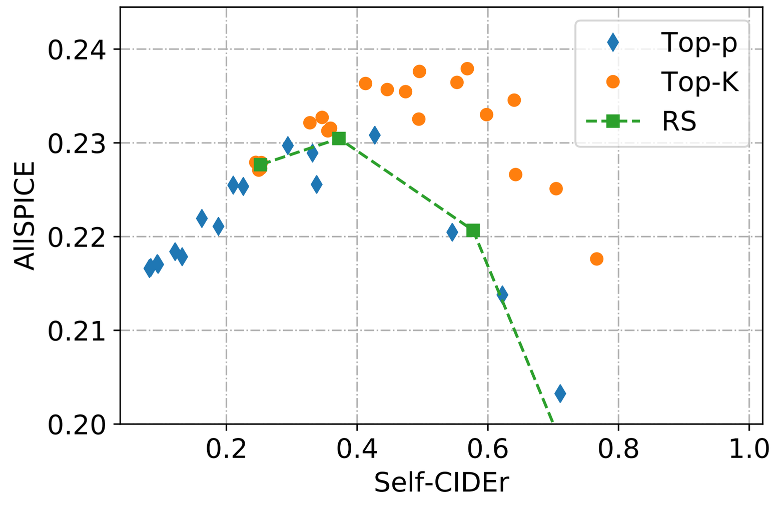
Avg-CIDEr vs. Self-CIDEr of XE, RL, XE+RL



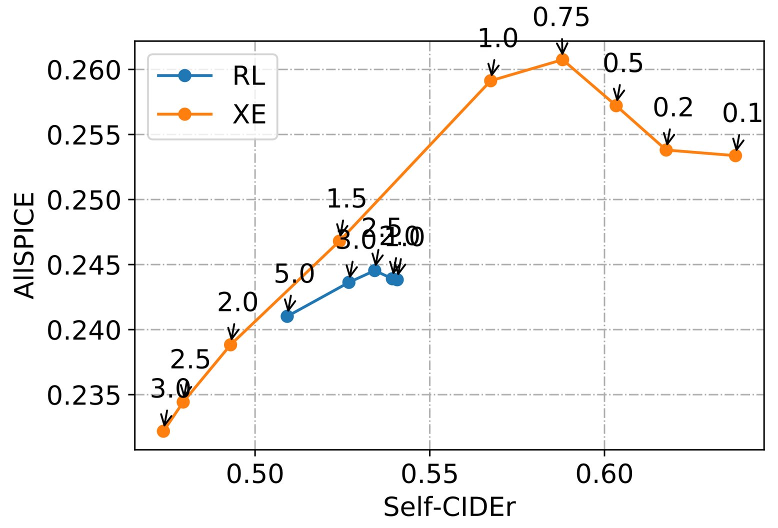
AllSPICE vs. self-CIDEr of XE, RL, XE+RL



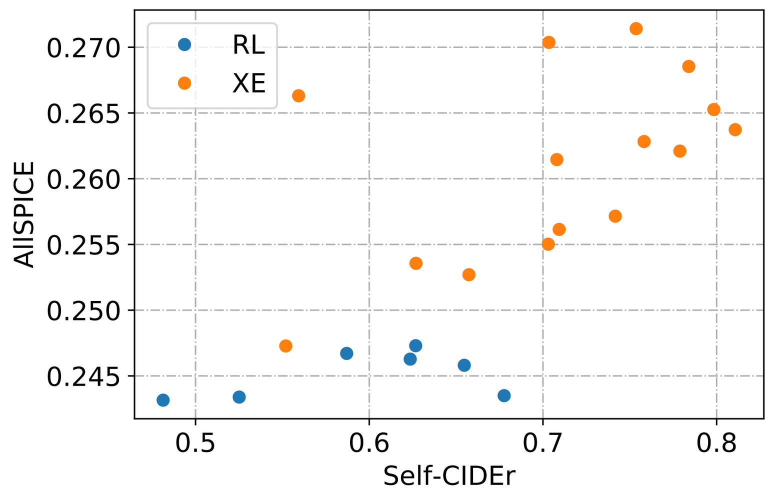
AllSPICE of XE with RS, Top-K and Nucleus



AllSPICE of RL with RS, Top-K and Nucleus



AllSPICE of XE, RL with Beam search



AllSPICE of XE, RL with Diverse Beam search

Different sampling methods

Random sampling with T =0.5 outperforms better than other settings on AllSPICE; no need to carefully tune the XE-RL weight in XE+RL method.

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

Beam search is 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):

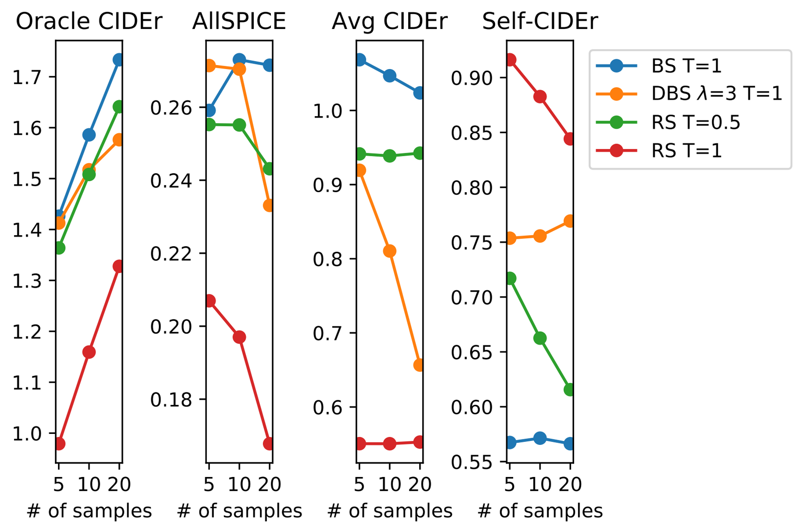
- Diverse beam search is the best algorithm with high AllSPICE and Self-CIDEr, indicating both semantic and syntactic diversity.
- Beam search performs best on oracle CIDEr and average CIDEr, and it performs well on AllSPICE too. However although all the generated captions are accurate, the syntactic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

	avg CIDEr	oracle CIDEr	AllSPICE	Self-CIDEr
DBS λ=0.3, T=1	0.919	1.413	**0.271**	**0.754**
BS T=0.75	**1.073**	**1.444**	0.261	0.588
Top-K K=3 T=0.75	0.921	1.365	0.258	0.736
Top-p p=0.8 T=0.75	0.929	1.366	0.257	0.744
RS T=0.5	0.941	1.364	0.255	0.717

Best performing hyperparamters for each method, and the resulting performance.

Sample size

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



AllSPICE drops with more samples, because additional captions are more likely to hurt (say something wrong) than help (add something correct not yet said). BS, which explores the caption space more "cautiously" than other methods, is initially resilient to this effect, but with enough samples its AllSPICE drops as well.

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, hurting precision without improving recall.

Qualitative results

DBS λ=3 T=1:
a kitchen with a sink and a window
there is a sink and a window in the kitchen
an empty kitchen with a sink and a window
an image of a kitchen sink and window
a sink in the middle of a kitchen
BS T=0.75:
a kitchen with a sink and a window
a kitchen with a sink a window and a window
a kitchen sink with a window in it
a kitchen with a sink and a sink
a kitchen with a sink and a window in it
Top-K K=3 T=0.75:
a kitchen with a sink and a mirror
the kitchen sink has a sink and a window
a sink and a window in a small kitchen
a kitchen with a sink a window and a window
a kitchen with a sink a window and a mirror
Top-p p=0.8 T=0.75:
a kitchen with a sink a window and a window
a kitchen with a sink a sink and a window
a kitchen with a sink and a window
a kitchen with a sink and a window
a kitchen with a sink and a window
RS T=0.5:
a kitchen with a sink a window and a window
a sink sitting in a kitchen with a window
a kitchen sink with a window on the side of the counter
a kitchen with a sink and a window
a kitchen with a sink and a window

DBS λ=3 T=1:
a wooden table topped with lots of wooden boards
a bunch of different types of food on a cutting board
there is a wooden cutting board on the table
some wood boards on a wooden cutting board
an assortment of vegetables on a wooden cutting board
BS T=0.75:
a wooden table topped with lots of wooden boards
a wooden cutting board topped with lots of wooden boards
a wooden cutting board with a bunch of wooden boards
a wooden cutting board with a wooden cutting board
a wooden cutting board with a bunch of wooden boards on it
Top-K K=3 T=0.75:
a wooden cutting board with a bunch of wooden boards
a wooden table with several different items
a wooden cutting board with some wooden boards
a wooden cutting board with some wooden boards on it
a bunch of different types of food on a cutting board
Top-p p=0.8 T=0.75:
a bunch of wooden boards sitting on top of a wooden table
a wooden cutting board with several pieces of bread
a wooden cutting board with a bunch of food on it
a bunch of different types of different colored UNK
a wooden cutting board with a wooden board on top of it
RS T=0.5:
a wooden cutting board with knife and cheese
a wooden table topped with lots of wooden boards
a wooden cutting board with chopped up and vegetables
a wooden table topped with lots of wooden boards
a wooden cutting board with some wooden boards on it

Conclusion

- Simple random sampling, coupled with suitably low temperature, is competitive with the best previously proposed decoding methods with respect to speed and diversity/accuracy tradeoff.
- Diverse beam search exhibits the best tradeoff, but it 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.