Diversity, Pragmatic Informativeness and Semantic Adequacy in Character-Level Image Captioning: A Comparison of Decoding Strategies

Anonymous EMNLP submission

Abstract

1 Introduction

Neural models with encoder-decoder architectures and RNN-based sequence generation are used for a variety of problems in Language and Vision (L&V), e.g. Image Captioning or Referring Expression Generation.

So far, research has mainly focused on the training of the models (Zarrieß and Schlangen, 2018). L&V models are most commonly optimized to generate human-like descriptions for a given image: During training this is reflected by the usage of Maximum Likelihood Estimation objectives, in evaluation by aiming for the best possible results in metrics such as BLEU (Papineni et al., 2002) and CIDER (Vedantam et al.), which rely on the similarity to given ground-truth captions. More recently, however, the decoding process, i.e. the way in which word sequences can be derived from token probabilities for individual steps, has also received increasing attention.

A basic decoding strategy in sequence generation is to pick the token with the highest probability in each step until an end token is generated *greedy search*. However, in many cases this method does not allow for optimal results and often leads to repetitive sentences or sentences that are defective in some other way. For this reasons it has been extended in different ways. A common extension is to simultaneously expand a defined number of hypotheses in each step (*beam search*). While this often leads to improved results regarding metrics like BLEU or CIDEr, other issues with greedy decoding are not solved: Beam Search often leads to short and repetitive sentences which are very similar to each other and in which only a small part of the available vocabulary is used. These shortcomings are addressed by various attempts to enhance diversity through e.g. stochastic decoding strategies such as *Top-K Random Sampling* (Fan et al., 2018) or *Nucleus Sampling* (Holtzman et al.) (cf. Ippolito et al. (2019a) for a comprehensive overview). There appears to be a trade-off between likelihood and diversity: Models which where shaped to provide sequences as similar as possible to human annotations were shown to produce a less diverse output. Models with a optimized diversity achieve lower results on metrics like BLEU or CIDEr (Wang and Chan).

While diversity enhancing decoding strategies are reported to be effective in terms of more diverse outputs, their linguistic implications are to be viewed critically. Both Nucleus Sampling and Top-K Random Sampling are based on randomness - the language model is used to determine a set of possible candidates, then the tokens to be generated are randomly selected from this set. The diversity of the resulting utterances can thus be seen as being caused by blurring the predictions of the trained model with respect to specific tokens. This seems appropriate for tasks such as conditional story generation, in which the stylistic properties of the generated text play an important role. Strictly speaking, however, this form of linguistic diversity can be seen as an illusion, since it does not reflect the creative and intentional use of language that underlies the diversity of human utterances. (pragmatic aspects of vocabulary choice: e.g. gricean maxims, as described by Cruse (1977) (Reiter, 1990), (Reiter, 1991))

Moreover, essential aspects of language are disregarded in Beam Search as well as in Nucleus Sampling or Top-K Random Sampling. In actual language use, linguistic utterances are not only true and well-formed, but also goal-oriented, as formulated, for example, by speech act theory or the Gricean Cooperative Principle. Even if the output of diversity-focused decoding strategies is more varied both structurally and lexically, it remains unclear whether this diversity is used purposefully by the model, for example to describe a visual referent as meaningfully as possible. This pragmatic informativity can be enhanced during decoding e.g. by using strategies built on the Rational Speech Acts (RSA) framework (Cohn-Gordon et al.).

In this work, we want to explore the interactions between likelihood, diversity and pragmatic informativity. For this, we want to compare different decoding strategies that are focused on optimal results for these individual aspects. We want to compare Beam Search (which was shown to yield good results in likelihood-based evaluation metrics), Nucleus Sampling (which is designed to tackle the diversity issues that arise with Beam Search), and RSA-based greedy decoding (which is focused on improving discrimination between targets and competing distractors e.g. in REG-like tasks). We test these approaches by using evaluation metrics that reflect the agreement of the generated sentences with ground-truth annotations, the diversity of the resulting captions, or the pragmatic informativeness with which the referents are described in the context of similar distractors.

We hypothesize that neither the increased (lexical) diversity in utterances produced using Nucleus Sampling nor the likelihood to human utteranced of captions produced using Beam Search lead to a higher pragmatic informativeness as compared to a greedy decoding baseline. Conversely, we assume that the introduction of additional pragmatic constraints in the RSA-based decoding leads to increased diversity compared to both to Greedy and Beam Search.

- research questions:
    1. how does pragmatic decoding relate to greedy, beam + nucleus? (does it increase/decrease scores on likelihood metrics? does it increase/decrease scores on diversity metrics?)
    2. are there structural differences (e.g. word types used) between the decoding strategies? (how does diversity look like if we look at it on a more detailed level?
    3. what are the differences between the decoding strategies if tested with a neural listener model?
    4. how do diversity enhancing decoding strategies work for character level decoding?

2 Related Work

- pragmatics in image descriptions / captioning: van Miltenburg et al. (2016), Cohn-Gordon et al.
- decoding + diversity (Ippolito et al., 2019b) (Wang and Chan) (van Miltenburg et al., 2018) structural properties of language (Ghodsi and DeNero, 2016) (Lippi et al.)
    - (Zarrieß and Schlangen, 2018)
- diversity
- reinforcement learning
- rational speech acts

3 Models, Methods, Data

3.1 Model

model description

3.2 Decoding Strategies

all decoding strategies: character level (because of RSA, other reasons?) character level for rsa: if made on word level, it would require some kind of pruning in order to be computationally feasible (Cohn-Gordon et al.). if we assess the lexical diversity, we should allow the model at least in theory to produce all words seen during the training

Greedy Decoding We use a simple Greedy decoding algorithm as our baseline: At each tim step, the word with the highest probability is selected and appended to the output sequence. The algorithm terminates after the generation of the end token or when the maximal sequence length is reached (cf. e.g. Zarrieß and Schlangen, 2018).

Beam Search In Beam Search, a fixed number of hypotheses is kept and expanded simultaneously at each step (cf. e.g. Graves). According to Zarrieß and Schlangen (2018), beam search algorithms can be modified in numerous ways, e.g. by specifying constant or dynamic values for the number of hypotheses considered at the same time (beam size), restricting possible next candidates (pruning), defining more sophisticated finishing conditions (termination) or normalizing candidates with different lengths. Here, we use a rather standard approach - we use static beam widths, refrain from pruning or length normalization, and terminate the beam search if the top candidate has the end token as its final segment.

Nucleus Decoding

- "In practice this means selecting the highest probability tokens whose cumulative probability massexceeds the pre-chosen thresholdp. The size of the sampling set will adjust dynamically based onthe shape of the probability distribution at each time step. For high values ofp, this is a small subsetof vocabulary that takes up vast majority of the probability mass — the nucleus." Holtzman et al.

Greedy RSA Decoding

- model implemeted as a "pragmatic speaker": a RSA model is combined with the neural image captioning model to produce captions that distinguish targets from similar images

- rsa speaker reasons about how the produced captions would be understood by a listener, to assess whether the utterances produced are capable of distinguishing the target

- Cohn-Gordon et al.

- whereas Cohn-Gordon et al. report results for a beam search variant, we focus on a greedy search method. The reasons for that are 1) computational constraints (we use a much larger test set compared to the 100 images / image clusters in the original paper) and 2) that the comparison to captions generated using beam search is less concise if the decoding strategy itself is a kind of beam search. this way, we have beam search, nucleus sampling and rsa decoding as three greedy search extensions, which have a minimal overlap.

- RSA approach used in REG (Zarrieß and Schlangen, 2019) and other language generation tasks (Shen et al.)

3.3 Evaluation Metrics

Likelihood BLEU / CIDEr

Diversity

Pragmatic Informativity a listener model reproduces captions produced by a speaker model in a greedy-like fashion, given a set of potential target images. for each token the model updates the probabilities for every image - the candidate with the highest probability is chosen as the target image. The accuracy and MRR of the choice of target images is compared between the decoding strategies.

3.4 Data

- images and annotations from MSCOCO (Lin et al.)

- speaker and listener models each trained on one half of the train partition

- random sample of 5000 images from val partition used for testing

4 Experiments

4.1 Likelihood and Diversity Tradeoffs

- method
  - assess BLEU and CIDEr scores for different decoding strategies and hyperparameters
  - compare between each other and to greedy baseline

- goal
  - determining whether likelihood & diversity tradeoff holds for beam search vs. nucleus decoding and how RSA decoding performs in terms of likelihood and diversity

- results
  - likelihood: beam search best, pragmatic worst. greedy ¿ nucleus
  - diversity: higher for pragmatic than for beam search / greedy (nucleus: depends on hyperparameters)

4.2 Vocabulary

- method
  - more in-depth analysis of the vocabulary generated by the decoding strategies
  - types gained / lost in comparison to greedy decoding (or other decoding strategies)
  - display of the word types generated

  - out of vocabulary words generated (with respect to the training captions)
  - average word frequency (with respect to the training captions)
  - zipf
- goal
  - determine more detailed differences between decoding strategies (not only TTR, novel captions or numbers of types generated)
- results
  -

4.3 Listener Evaluation

- method
  - evaluate using listener model: for each caption the model tries to distinguish the right target image against a set of similar distractor images (method adopted from Cohn-Gordon et al.)
  - accuracy or MRR of decisions used to compare the models
- goal
  - check success of pragmatic decoding
  - determine whether the increased linguistic diversity in nucleus decoding is used purposefully
  - see how the non-RSA decoding strategies compare to RSA decoding and to each other
- results
  - best results for RSA decoding
  - no big differences between other strategies

5 General Discussion

- trainable decoding?
- combining decoding strategies (if compatible)

6 Conclusion

References

Reuben Cohn-Gordon, Noah Goodman, and Christopher Potts. Pragmatically informative image captioning with character-level inference.

D. A. Cruse. 1977. The pragmatics of lexical specificity. *Journal of Linguistics*, 13(2):153–164. - Einfluss des situativen Kontext auf Spezifizität von Referenz.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Aneiss Ghodsi and John DeNero. 2016. An analysis of the ability of statistical language models to capture the structural properties of language. In *Proceedings of the 9th International Natural Language Generation conference*, pages 227–231, Edinburgh, UK. Association for Computational Linguistics.

Alex Graves. Sequence transduction with recurrent neural networks.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration.

Daphne Ippolito, Reno Kriz, João Sedoc, Maria Kustikova, and Chris Callison-Burch. 2019a. Comparison of diverse decoding methods from conditional language models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3752–3762, Florence, Italy. Association for Computational Linguistics.

Daphne Ippolito, Reno Kriz, Joao Sedoc, Maria Kustikova, and Chris Callison-Burch. 2019b. Comparison of diverse decoding methods from conditional language models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3752–3762, Florence, Italy. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context.

Marco Lippi, Marcelo A Montemurro, Mirko Degli Esposti, and Giampaolo Cristadoro. Natural language statistical features of lstm-generated texts.

Emiel van Miltenburg, Desmond Elliott, and Piek Vossen. 2018. Measuring the diversity of automatic image descriptions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1730–1741, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Emiel van Miltenburg, Roser Morante, and Desmond Elliott. 2016. Pragmatic factors in image description: The case of negations. In *Proceedings of the 5th Workshop on Vision and Language*, pages 54–59, Berlin, Germany. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ehud Reiter. 1990. A new model for lexical choice for open-class words. In *Proceedings of the Fifth International Workshop on Natural Language Generation*.

Ehud Reiter. 1991. A new model of lexical choice for nouns. *Computational Intelligence*, 7(4):240–251.

Sheng Shen, Daniel Fried, Jacob Andreas, and Dan Klein. Pragmatically informative text generation.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation.

Qingzhong Wang and Antoni B. Chan. Describing like humans: on diversity in image captioning.

Sina Zarrieß and David Schlangen. 2018. Decoding strategies for neural referring expression generation. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 503–512, Tilburg University, The Netherlands. Association for Computational Linguistics.

Sina Zarrieß and David Schlangen. 2019. Know what you don't know: Modeling a pragmatic speaker that refers to objects of unknown categories. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 654–659, Florence, Italy. Association for Computational Linguistics.