UNIVERSITÄT BIELEFELD

Diversity as a By-Product:

Goal-oriented Language Generation Leads to Linguistic Variation

Simeon Schüz[1], Ting Han[2], Sina Zarrieß[1]

[1]Bielefeld University [2]Artificial Intelligence Research Center (AIST), Tokyo

SIGDIAL 2021

Scope of this work

- Task: **Image Captioning**
  - generating descriptions for images
  - usually non-interactive
- Key question: How does pragmatic reasoning affect the **diversity** of generated captions?
- Focus on **decoding** stage

Outline

1) Background
2) Decoding Methods
3) Experimental Set-Up
   - Data
   - Model
   - Evaluation
4) Results
5) Conclusion & Future Work

Introduction

Human Annotations:

- “**Boats** are traveling in the **large open water**”
- “A **cruise ship** travelling out of an **expansive harbor**.”
- “There is a **boat** going across the **waterway**.”

Introduction

Captioning Model (Beam Search):

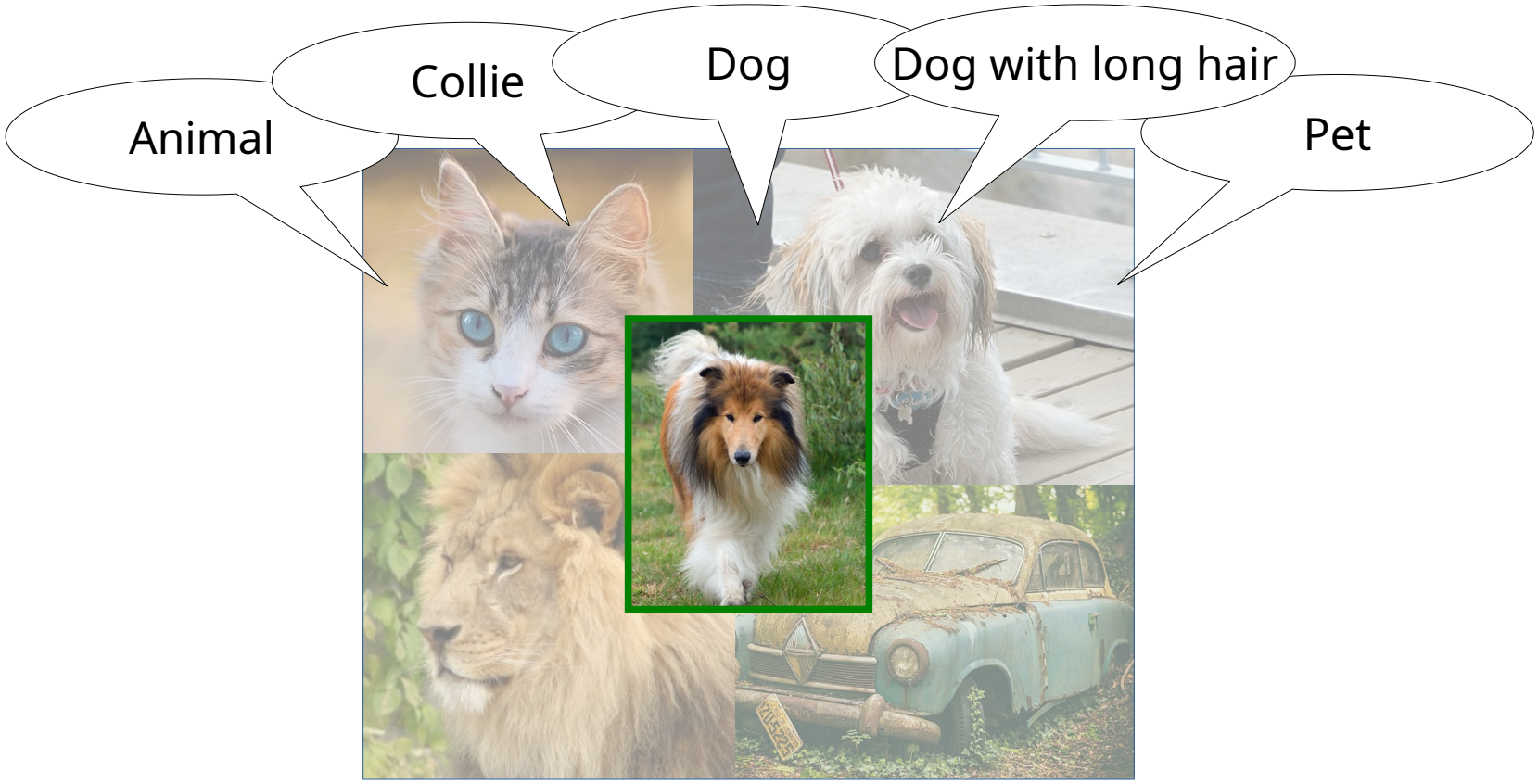
- `a boat that is floating in the water`

Introduction

Dog

Collie

Introduction



Experimental Set-Up

	Likelihood	Diversity	Informativity
Decoding Methods	Greedy, Beam	Top-K, Nucleus	Emitter/Supressor, RSA
Evaluation Metrics	$BLEU_4$, CIDEr, SPICE	Metrics from Van Miltenburg et al. 2018, frequency ranks	Recall from external retrieval model

Decoding / Likelihood

1) Greedy Search:

- Pick the most probable token at each step

2) Beam Search

- Expand multiple hypotheses simultaneously

Decoding / Diversity

1) Top-K Sampling (Fan et al. 2018)
   - Sample from the *k* most probable candidates
2) Nucleus Sampling (Holtzman et al. 2020)
   - Sample from the top *p* part of the cumulative probability mass

- Additional parameter: Temperature *t*

Decoding / Informativity

- RSA Decoding (Cohn-Gordon et al. 2018)
  - Based on the *Rational Speech Acts* Framework
  - Integrates pragmatic reasoning into the iterative unrolling of the generation model: How would a listener interpret the utterance?

Decoding / Informativity

- *Literal Speaker (S_0):* Generation Model; assigns initial probabilities to token candidates
- *Pragmatic Listener (L_0):* Assesses whether token candidates allow identifying the target

$$L_0(w|u) \propto \frac{S_0(u|w) * P(w)}{\sum_{w' \in W} S_0(u|w') * P(w')}$$

Decoding / Informativity

- *Pragmatic Speaker (S_1):* Chooses the most informative token candidates

$$S_1(u|w) \propto \frac{L_0(w|u)^\alpha * P(u)}{\sum_{u' \in U} L_0(w|u')^\alpha * P(u')}$$

- α : Rationality Parameter

Decoding / Informativity

- Emitter / Supressor (Vedantam et al. 2017)
  - Base model as *Emitter*
  - *Supressor* function rates informativity

Decoding / Informativity

- Slightly modified for multiple distractor images

$$\Delta(I_t, D) =$$

$$\arg\max_{s} \sum_{\tau=1}^{T} \sum_{i=1}^{|D|} \log \frac{p(s_\tau | s_{1:\tau-1}, I_t)}{p(s_\tau | s_{1:\tau-1}, D_i)^{1-\lambda}}$$

- λ: Rationality Parameter

Evaluation

1)Likelihood

- $BLEU_4$, CIDEr, SPICE

2)Diversity

- TTR1, TTR2, % novel, word types, % coverage (Van Miltenburg et al. 2018); avg. frequency rank

3)Informativity

- Recall from external retrieval model (Faghri et al. 2018)

Data

- COCO Caption Dataset
- Image clusters needed for discriminative decoding & informativity evaluation

Target Image

Distractor Images

Model

- Adaptive Attention Model (Lu et al. 2017)
  - Used for all decoding methods

Hypotheses

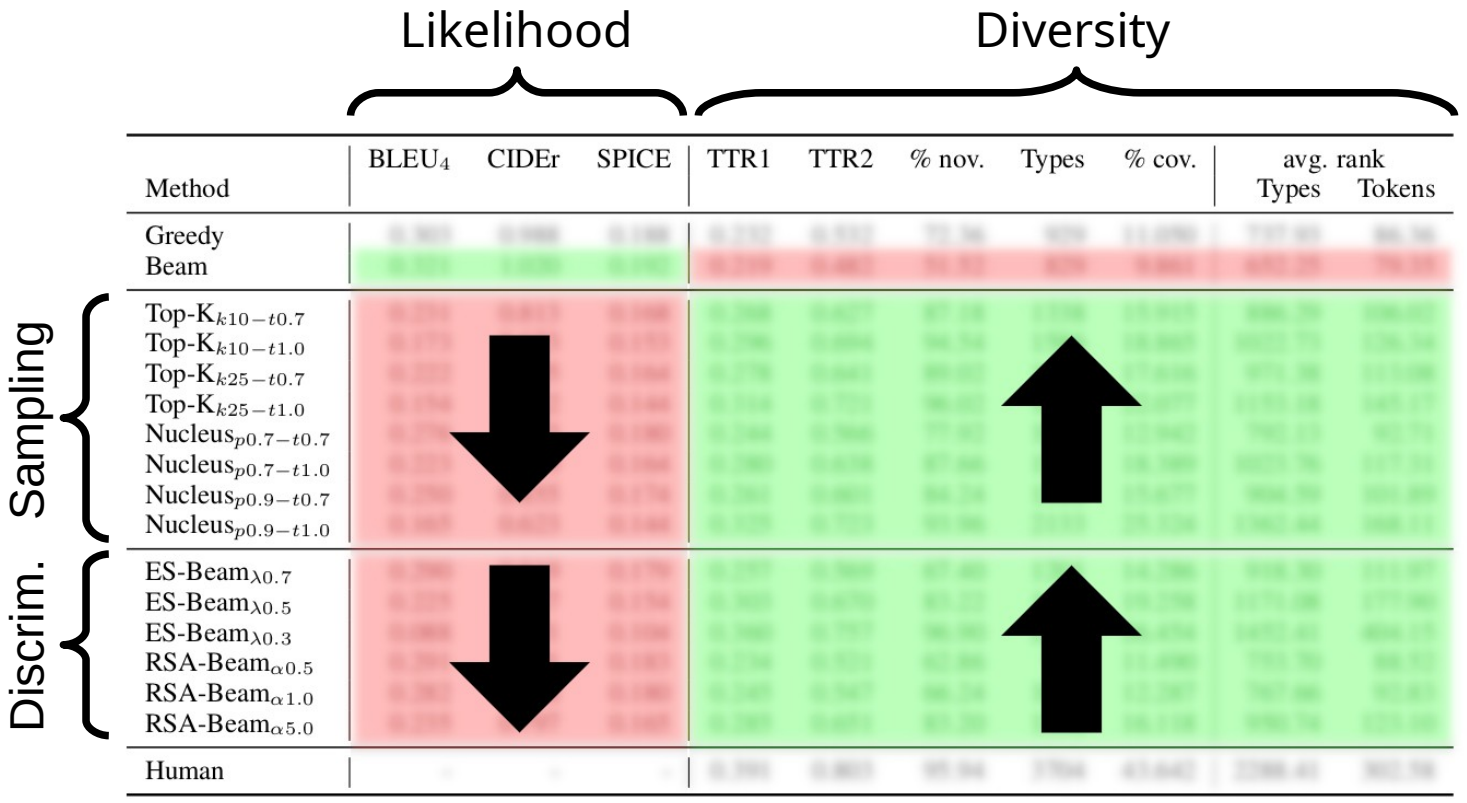
1) Discriminative Decoding → higher diversity
2) Pragmatically driven variation → improved retrieval results
3) Similar linguistic strategies as found in humans

Results: Likelihood / Diversity

Likelihood | Diversity

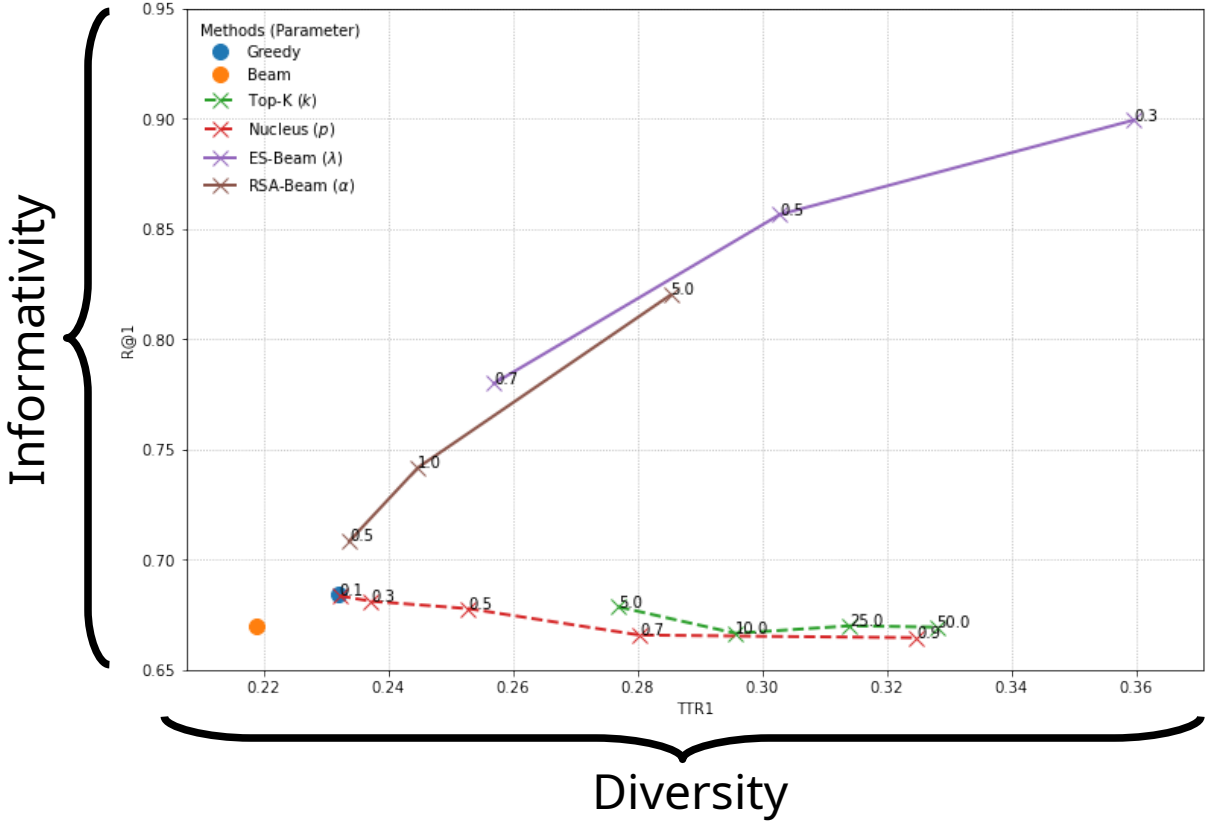
	Method	$BLEU_4$	CIDEr	SPICE	TTR1	TTR2	% nov.	Types	% cov.	avg. rank Types	avg. rank Tokens
	Greedy	0.303	0.988	0.188	0.232	0.532	72.36	929	11.050	737.93	86.36
	Beam	0.321	1.020	0.192	0.219	0.482	51.52	829	9.861	652.25	79.35
Sampling	Top-$K_{k10-t0.7}$	0.231	0.813	0.168	0.268	0.627	87.18	1338	15.915	886.29	106.02
Sampling	Top-$K_{k10-t1.0}$	0.173	0.673	0.153	0.296	0.694	94.54	1586	18.865	1022.73	126.34
Sampling	Top-$K_{k25-t0.7}$	0.222	0.785	0.164	0.278	0.641	89.02	1482	17.616	971.38	113.08
Sampling	Top-$K_{k25-t1.0}$	0.154	0.612	0.144	0.314	0.721	96.02	1857	22.077	1153.18	145.17
Sampling	$Nucleus_{p0.7-t0.7}$	0.276	0.923	0.180	0.244	0.566	77.92	1088	12.942	792.13	92.71
Sampling	$Nucleus_{p0.7-t1.0}$	0.223	0.779	0.164	0.280	0.638	87.66	1546	18.389	1023.76	117.31
Sampling	$Nucleus_{p0.9-t0.7}$	0.250	0.855	0.174	0.261	0.601	84.24	1319	15.677	904.59	101.89
Sampling	$Nucleus_{p0.9-t1.0}$	0.165	0.623	0.144	0.325	0.723	93.96	2133	25.324	1362.44	168.11
Discrim.	ES-$Beam_{\lambda 0.7}$	0.290	0.919	0.179	0.257	0.569	67.40	1201	14.286	918.30	111.97
Discrim.	ES-$Beam_{\lambda 0.5}$	0.225	0.727	0.154	0.303	0.670	83.22	1619	19.258	1171.08	177.90
Discrim.	ES-$Beam_{\lambda 0.3}$	0.088	0.371	0.104	0.360	0.757	96.90	2225	26.454	1452.41	404.15
Discrim.	RSA-$Beam_{\alpha 0.5}$	0.291	0.951	0.183	0.234	0.521	62.86	966	11.490	753.70	88.52
Discrim.	RSA-$Beam_{\alpha 1.0}$	0.282	0.928	0.180	0.245	0.547	66.24	1033	12.287	767.66	92.83
Discrim.	RSA-$Beam_{\alpha 5.0}$	0.235	0.797	0.165	0.285	0.651	83.20	1356	16.118	950.74	123.10
	Human	-	-	-	0.391	0.803	95.94	3704	43.642	2288.41	302.58

Results: Likelihood / Diversity

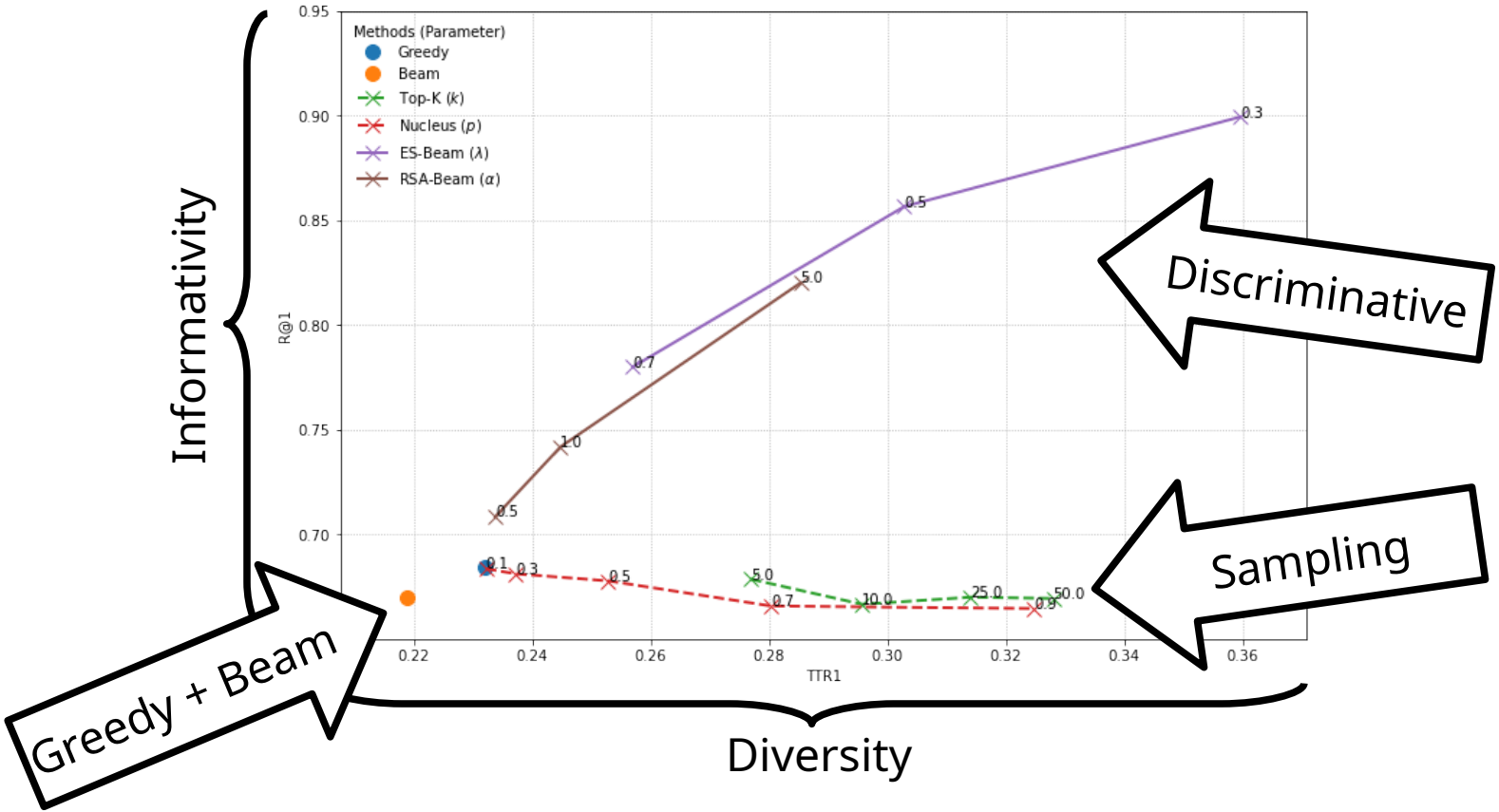


Method	BLEU$_4$	CIDEr	SPICE	TTR1	TTR2	% nov.	Types	% cov.	avg. rank Types	avg. rank Tokens
Greedy										
Beam										
Top-K$_{k10-t0.7}$										
Top-K$_{k10-t1.0}$										
Top-K$_{k25-t0.7}$										
Top-K$_{k25-t1.0}$										
Nucleus$_{p0.7-t0.7}$										
Nucleus$_{p0.7-t1.0}$										
Nucleus$_{p0.9-t0.7}$										
Nucleus$_{p0.9-t1.0}$										
ES-Beam$_{\lambda 0.7}$										
ES-Beam$_{\lambda 0.5}$										
ES-Beam$_{\lambda 0.3}$										
RSA-Beam$_{\alpha 0.5}$										
RSA-Beam$_{\alpha 1.0}$										
RSA-Beam$_{\alpha 5.0}$										
Human	-	-	-							

Results: Informativity

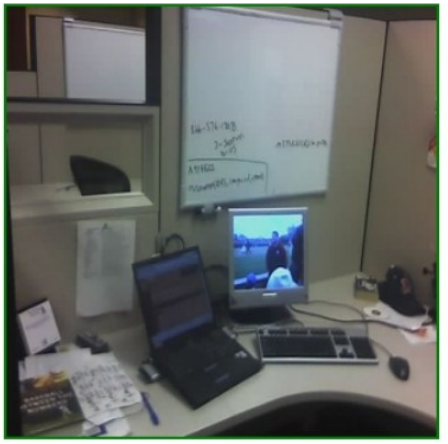


Results: Informativity



Results: Qualitative Examples

Greedy	a clock tower with a clock on top of it
$Nucleus_{p0.7-t1.0}$	a building that has a clock tower in the center
$ES - Beam_{\lambda 0.5}$	a tall clock tower with trees in the background
$RSA - Beam_{\alpha 1.0}$	a view of a tall clock tower with trees in the background



Greedy	a desk with a laptop and a desktop computer
$Nucleus_{p0.7-t1.0}$	a desktop computer sitting on top of a desk
$ES - Beam_{\lambda 0.5}$	a cluttered cubicle with multiple computers and monitors
$RSA - Beam_{\alpha 1.0}$	an office cubicle with multiple computers and monitors

Results: Specificity / Modifiers

- Lexical Specificity (Distance from WordNet root)
  - More specific for higher rationalities
- Descriptive Modifiers (Distribution of POS Tags)
  - Higher ratio of adjectives for higher rationalities

  (should be taken with some caution)

Discussion & Conclusion

- Main Results:
  - Discriminative decoding increases diversity & informativity, although not aiming at diversity itself
  - Indications for linguistically plausible variation
- Conclusions:
  - Diversity arises from (and can be modeled through) general principles of language use

Discussion & Conclusion

- Future work:
  - other kinds of data (e.g. Referring Expression datasets)
  - other dialogue tasks
  - other sources of variation

Citations

- Reuben Cohn-Gordon, Noah Goodman, and Christopher Potts. 2018. Pragmatically informative image captioning with character-level inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 439–443.
- Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. VSE++: improving visual-semantic embeddings with hard negatives. In *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*, page 12. BMVA Press.
- 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.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.
- Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 375–383.
- 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.
- Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. 2017. Context-aware captions from context-agnostic supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 251–260.