The Effect of Generalisation on the Inadequacy of the Mode

Bryan Eikema (b.eikema@uva.nl)

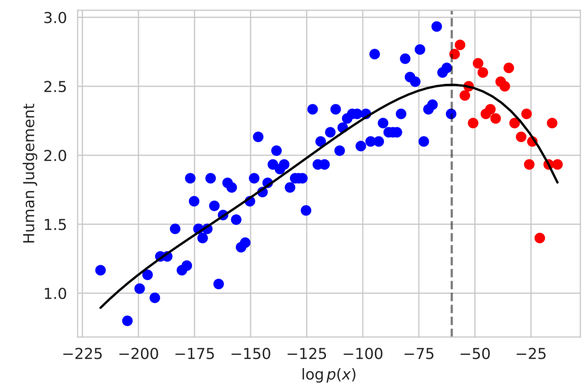
University of Amsterdam

Key Takeaways

- In this short position paper, we consider the potential causes behind the inadequacy of the mode.
- We challenge existing hypotheses and argue they don't paint a full picture.
- We bring light to the role that generalisation may play in causing the inadequacy of the mode.
- We argue that, under current data and modeling limitations, inadequate modes may be an inevitable consequence of our desire to generalise to unseen contexts.

Why do Sequence Generation Models Have Inadequate Modes?

- Natural language generation suffers from the **inadequacy of the mode**: high probability sequences and the mode, i.e. the highest probability sequence, are typically inadequate in some way (e.g. empty sequences, containing excessive repetitions, or being copies of the input).
- **It's unclear why** sequence generation models exhibit inadequate modes.
- **Few general hypotheses exist**. Most hypotheses only focus on particular degeneracies, yet the inadequacy of the mode is observed widely across tasks and models.
- Even **recent large language models still exhibit inadequate modes** *(Yoshida et al., 2023)*.



Sequence quality and model probability only correlate positively until an inflection point, Figure from *(Zhang et al., 2021)*.

Expected Information Hypothesis

- *(Meister et al., 2022)* observe that ground-truth sequences have **surprisal (negative log probability) close to the entropy** of the model.
- They hypothesise that this is no coincidence: **humans communicate to optimise reliability and efficiency**, hence producing sequences with surprisal around the entropy.
- In high entropy distributions the mode has a surprisal far away from the entropy. **Good models may therefore capture this behavior as well**.

Criticisms

- If this is a property of human language, do models *need* to capture this? We typically don't produce inadequate sequences and hence they **should not appear in training data**.
- Sequence models in more constrained tasks are capable of learning sequence distributions with adequate modes. Given sufficient data and modelling capacity, would NLG models not be able to mimic the empirical distributions? **Is this really an inherent property of natural language data?**

Low-Entropy Distractors in Training Data

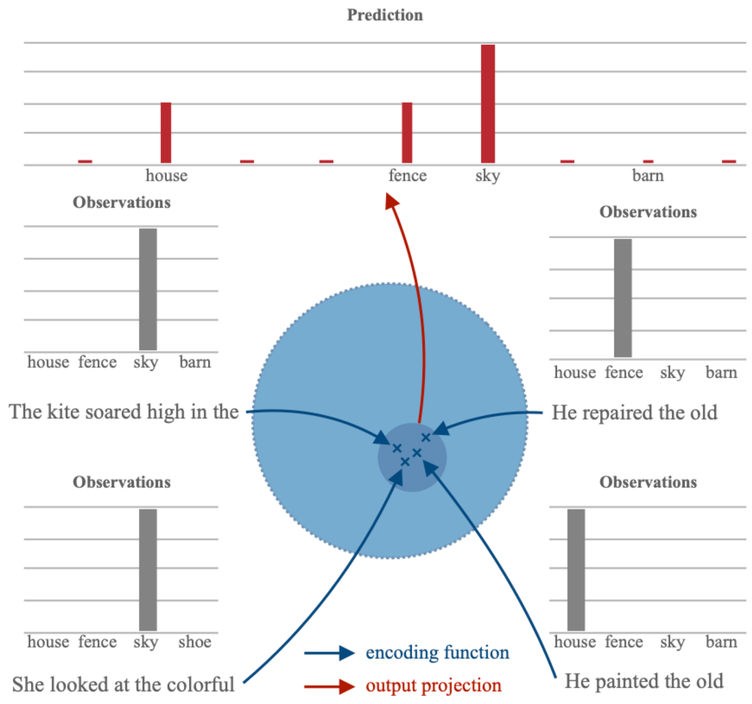
- *(Yoshida et al., 2023)* consider training data noise to be the source of the problem, so-called low-entropy distractors: **consistent low variance noise present in the training data**.
- Conceptually, consider we have many valid sequences collected for any single input, and we place a **uniform distribution over those valid sequences**.
- Any noise that occurs with a **rate higher than the uniform probability becomes the mode** of the empirical distribution.
- Hence, even a model that **perfectly reproduces the empirical distribution has an inadequate mode!**

Criticisms

- Some **degeneracies are still observed even if we filter** them from training data (e.g. the empty sequence).
- This conceptual scenario quite far from reality: we typically **only observe a single valid sequence per input** and wish to generalise to unseen inputs.

The Effect of Generalisation

- We typically **only observe a single or a few observations per context**, even after factorisation: we are dealing with extremely sparse contexts.
- At test-time, we wish to **predict distributions over previously unseen contexts**.
- In order to make meaningful predictions, a neural network will need to map different local contexts to similar representations, such that **their observations jointly contribute to the same next-word distributions**.
- We **hypothesise that this mapping is not always due to actual linguistic equivalence**, in that human continuations may not be identical for all contexts.
- This bundling behavior would **introduce spread**, as maximum likelihood estimation would try to accommodate all observations in the training data.
- As a result, **some contexts may not retain an adequate mode**. Compounded to the sequence-level this can result in inadequate modes in the sequence distributions.
- Therefore, in order to generalise well, it **may be the case that inadequate modes are a necessary by-product of generalisation**.



A conceptual illustration of our argument. To make informed predictions for unseen contexts a neural network may map different contexts to similar representations. As a result, observations for those contexts contribute to the same next-word distributions, introducing spread. We argue that these contexts may not always be linguistically equivalent, i.e. human continuations may distribute differently for those contexts.

References

(Meister et al., 2022) On the probability–quality paradox in language generation in Proceedings of ACL 2022

(Yoshida et al., 2023) Map's not dead yet: Uncovering true language model modes by conditioning away degeneracy on arXiv

(Zhang et al. 2021) Trading Off Diversity and Quality in Natural Language Generation in Proceedings of HumEval 2021