8. Holistic Evaluation

Evaluation beyond accuracy

Most basic characterization: **Accuracy**

Linguists, cognitive scientists: **interpretability**

Practitioners: **efficiency, robustness**

Product managers: **user interaction, calibration, explainability**

Policymakers: **fairness, privacy**

Robustness

Our standard setting assumes that the training and test examples are independent and identically distributed (iid). However, this is almost never true in practice.

Different types of robustness:

- Robustness to **adversarial examples** that are designed to fool the model
- Robustness to **perturbation** of iid examples

Adversarial robustness

- Find minimal Δx that maximizes $L(x + \Delta x, y)$
- Solve an optimization problem
- Challenge in NLP: optimizing in discrete space

Adversarial examples for reading comprehension [Jia et al., 2017]

- Goal: perturb paragraph+question to change the model's prediction but not the groundtruth
- Perturbation needs to be minimal: add a distractor sentence to the paragraph
- The distractor sentence needs to change the prediction (make it similar to the answer sentence)

Text perturbations: small edits to the input text

Label-perserving perturbations: can often be automated

- Typos: the table is sturdy → the tabel is sturdy
- Capitalization: the table is sturdy → The table is sturdy
- Synonym substitution: the table is sturdy → the table is solid

Behaviorial testing of NLP models

Checklist [Ribeiro et al., 2020]

- Inspired by unit tests in software engineering

- Minimum functionality test: simple test cases focus on a capability
- Invariance test: label-perserving edits (e.g., change entities in sentiment tasks)
- Directional expectation test: label-changing edits

Calibration

In high-stake settings (e.g., healthcare), we want to know how **uncertain** the model prediction is.

Definition: A **perfectly-calibrated** model should output confidence scores that are equal to the probability that the prediction is correct.

$\mathbb{P}(\text{prediction} = \text{groundtruth} | \text{confidence} = p) = p, \forall p \in [0, 1]$

Measuring calibration error: **expected calibration error** [Naeini et al., 2015]

Main idea: discretize confidence score; partition predictions into equally-spaced bins $B_1, ..., B_M$

$\text{ECE} = \sum_{m=1}^{M} \frac{|B_M|}{n} |\text{accuracy}(B_m) - \text{confidence}(B_m)|$

Modern neural networks are poorly calibrated [Gao et al., 2017]

How can we use the confidence score?

- Abstain (not predicting) on examples with low confidence
- Optionally ask for human help
- Accuracy-coverage trade-off: accuracy can be improved by raising the confidence threshold, but coverage (fraction of examples where we make a prediction) is reduced

Fairness and bias

Amplification of bias through the model:

- Cooking is about 33% more likely to involve females than males
- But the model predicts woman 68% more likely than man

Fairness and bias metrics

- Counterfactual fairness: the model should produce the same prediction when the related social group is changed in the data (all else being equal).
    - Gender substitution from “he” to “she” → invariant prediction
- Performance disparities: the model should have similar performance across different groups