Underspecification Presents Challenges for Credibility in Modern Machine Learning

Paper by D'Amour et al. (2020)
Presenter: Matias Alvo

February 13, 2025

Key Question

- Why do ML models often fail unexpectedly when deployed in real-world settings?
- Main Answer: Underspecification
  - Many different models can achieve similar performance during training
  - But behave very differently in deployment

The Structural-Conflict View

- Common explanation: fundamental conflict between iid performance and encoding credible inductive biases
- Example: Disease Prediction Model
  - Training data only from US hospitals with advanced equipment
  - Model learns to rely on high-resolution test results
  - But deployment in developing countries has different quality
- Example: Training and deploying on populations of different geographic areas
- If this view was enough
  - Models trained under same data would have similar OOD performance
  - Underspecification provides another explanation

Paper Contributions

- Two main claims:
  1. Underspecification is a key obstacle to reliable ML deployment
     - Even if good solutions exist, pipeline might not find them
     - Arbitrary choices affect real-world behavior
  2. Underspecification is ubiquitous in modern ML
     - Affects computer vision, NLP, medical imaging, etc.
     - Impacts robustness, fairness, and causal understanding

- Suggestions:
  - Need explicit testing beyond iid evaluation
  - Develop methods to constrain models toward desired behaviors

What is Underspecification?

- A problem is underspecified when multiple distinct solutions solve it equally well
- Example: Underdetermined system of linear equations
- In ML: Many models achieve similar training performance but behave differently in deployment

Shortcuts and Failure Modes

- Structural Failure:
  - Example: Skin cancer detection using surgical markings
  - Model *must* use spurious features to achieve optimal training performance
- Underspecified Failure:
  - Example: Image classification with sufficient information in relevant features
  - Model *could* use proper features, but might learn shortcuts
  - Different training runs can learn different shortcuts

Stress Tests and Credibility

Three types of stress tests:

1. Stratified Performance Evaluations
   - Test across different subgroups
   - Example: Face recognition across skin types
2. Shifted Performance Evaluations
   - Test under specific distribution changes
   - Example: ImageNet-C (corrupted images)
3. Contrastive Evaluations
   - Test on matched sets of modified inputs
   - Example: Testing gender bias by changing pronouns

Random Feature Model Analysis

- Regression task with linear target ($y = \beta^T x$)
- Simple model to study overparameterization:
  - First layer with fixed random weights
  - Trained second layer
- Key findings:
  - Different random initializations $\rightarrow$ same training performance
  - But predictors nearly orthogonal to each other
  - Very different behaviors under distribution shift
  - Some models vulnerable to specific shifts, others robust

Key Insights

- Underspecification is ubiquitous in modern ML
- Problems:
  - Models with same test performance behave differently in deployment
  - Standard validation doesn't capture these differences
  - Random choices in training can lead to very different models
- Need new ways to:
  - Test for required behaviors beyond standard validation
  - Constrain models to have desired properties

Approach: Testing Underspecification in Deep Learning

- Goal: Show underspecification exists in real deep learning systems
- Method:
  - Take state-of-the-art models in different domains
  - Create ensemble by e.g., perturbing random seed
  - All models achieve similar training/validation performance
  - Test behavior on application-specific stress tests
- Domains tested:
  - Computer vision (including medical imaging)
  - Natural language processing
  - Clinical predictions from health records

Evaluating Variation in Model Behavior

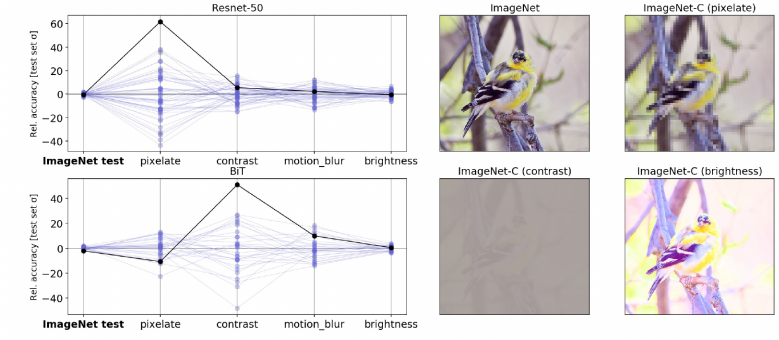
Three key properties to establish underspecification:

1. Magnitude of Variation
   - How much do models differ on stress tests?
   - Example: Some ImageNet models 10x more sensitive to image corruptions than others
2. Unpredictability from iid Performance
   - Does good validation performance predict good stress test performance?
   - Example: Model accuracy on clean images doesn't predict robustness to corruptions
3. Systematic Differences
   - Are differences random or do they reflect meaningful patterns?
   - Example: Some models consistently more robust to specific types of image corruptions

Why These Properties Matter

- Large Magnitude $\rightarrow$ Choices matter
  - Random seed can lead to drastically different deployment behavior
  - Even when validation performance is identical
- Unpredictability $\rightarrow$ Can't select good models using validation
  - Need explicit testing for desired properties
  - Can't rely on standard metrics
- Systematic Differences $\rightarrow$ Not just noise
  - Models learn genuinely different strategies
  - Different inductive biases emerge from random choices

Case Study: Computer Vision



- Models tested:
  - ResNet-50 (standard)
  - BiT (pretrain + finetune)
- Created ensembles:
  - 50 ResNet-50s (change random seed), 30 BiTs (random seed + init)
- Variability of performance in stress tests » ImageNet
- Weak correlation of performance across datasets

Case Study: Computer Vision

Dataset	ImageNet	pixelate	contrast	motion blur	brightness	ObjectNet
ResNet-50	0.759 (0.001)	0.197 (0.024)	0.091 (0.008)	0.100 (0.007)	0.607 (0.003)	0.259 (0.002)
BiT	0.862 (0.001)	0.555 (0.008)	0.462 (0.019)	0.515 (0.008)	0.723 (0.002)	0.520 (0.005)

Table 1: **Accuracies of ensemble members on stress tests.** Ensemble mean (standard deviations) of accuracy proportions on ResNet-50 and BiT models.

Dataset	ImageNet	ImageNet (subset)	ObjectNet
ResNet-50	0.160 (0.001)	0.245 (0.005)	0.509 (0.003)
BiT	0.064 (0.004)	0.094 (0.006)	0.253 (0.012)

Table 2: **Ensemble disagreement proportions for ImageNet vs ObjectNet models.** Average disagreement between pairs of predictors in the ResNet and BiT ensembles. The "subset" test set only includes classes that also appear in the ObjectNet test set. Models show substantially more disagreement on the ObjectNet test set.

Some Additional Case Studies

- **Medical Imaging - Ophthalmology:**
  - Model for diabetic retinopathy detection from retinal fundus images
  - When testing on new camera types not seen in training, models showed large performance variations
  - Models with identical training but different random seeds showed systematically different calibration curves
- **NLP - Gender Bias in BERT:**
  - Examined how identical BERT models handle gender bias differently
  - Models varied significantly in gender associations despite same training
  - On tasks like sentence similarity and pronoun resolution, some models showed strong gender biases while others showed much weaker biases
  - Demonstrates how underspecification leads to unpredictable bias behavior

Implications and conclusions

- **Underspecification is Ubiquitous:** Arbitrary choices (random seeds, initialization, hyperparameters) can significantly impact model behavior
- **Need for Robust Testing:**
  - Develop application-specific stress tests
  - Ensure performance stability across different domains
- **Prescriptions/research directions:**
  - Systematically map set of risk minimizers to quantify uncertainty
  - Test models on application-specific tasks
    - Design stress-tests that provide coverage of failure modes
  - Design criteria to better select predictor among the risk-minimizers
    - Likely needs to be application-specific