Inoculation by Fine-Tuning: A Method for Analyzing Challenge Datasets

Nelson F. Liu

Roy Schwartz

Noah A. Smith

NAACL 2019—June 4, 2019
Two Key Ingredients of NLP Systems

Training Dataset

Model Architecture

NLP System
Why Might NLP Systems Fail?

Training Dataset

Model Architecture

NLP System
Dataset Weaknesses

Training Dataset

NLP System

Model Architecture
Model Weaknesses

Training Dataset

Model Architecture

NLP System
Challenge Datasets Break Models
Challenge Datasets Break Models

(Step 1)

Train on Original
Challenge Datasets Break Models

(Step 1)
Train on Original

(Step 2)
Test on Original & Challenge
NLP Systems Are Brittle

**Outcome:**
Challenge is difficult for the model.
NLP Systems Are Brittle

Original Performance

(Step 1)
Train on Original

(Step 2)
Test on Original & Challenge

Outcome:
Challenge is difficult for the model.

Why?
Inoculation by Fine-Tuning

**Standard Challenge Evaluation**

(Step 1) **Train** on **Original**

(Step 2) **Test** on **Original & Challenge**

**Outcome:**

Challenge is difficult for the model.

**Why?**
Inoculation by Fine-Tuning

Standard Challenge Evaluation

(Step 1)
Train on Original

(Step 2)
Test on Original & Challenge

Outcome:
Challenge is difficult for the model.

Why?

(Step 3)
Fine-tune on a few challenge examples
Inoculation by Fine-Tuning

![Diagram of inoculation process]

**Standard Challenge Evaluation**

1. **Train** on Original
2. **Test** on Original & Challenge
3. **Fine-tune** on a few challenge examples
4. **Re-test** on Original & Challenge

**Outcome:** Challenge is difficult for the model.

**Why?**
Inoculation
Inoculate Models to Better Understand Why They Fail
Three Clear Outcomes of Interest

- Original Performance
- Challenge Performance

Challenge Evaluation Outcome ➞ Challenge Examples ➞ Inoculation ➞ ?
(1) Dataset Weakness

- Original Performance
- Challenge Performance

Challenge Evaluation Outcome → Inoculation → Dataset Weakness
(2) Model Weakness

- **Original Performance**
- **Challenge Performance**

- Challenge Evaluation Outcome
- Inoculation
- Model Weakness
(3) Predictive Artifacts / Other

- Original Performance
- Challenge Performance

Challenge Evaluation Outcome → Inoculation → Predictive Artifacts / Other
Three Clear Outcomes of Interest

- **Original Performance**
- **Challenge Performance**

**Challenge Examples**

- **Dataset Weakness**
- **Model Weakness**
- **Predictive Artifacts / Other**

**Challenge Evaluation Outcome**

**Inoculation**
Case Studies

• Inoculating natural language inference (NLI) models

• Inoculating SQuAD reading comprehension models
Natural Language Inference (NLI)

Premise: "I have done what you asked."
Hypothesis: "I have disobeyed your orders."

Entailment  Neutral  Contradiction
Two NLI Challenge Datasets

Premise: "I have done what you asked."
Hypothesis: "I have disobeyed your orders."
Two NLI Challenge Datasets

Premise: "I have done what you asked."
Hypothesis: "I have disobeyed your orders."

Word Overlap Challenge Dataset

Premise: "I have done what you asked."
Hypothesis: "I have disobeyed your orders and true is true."
Two NLI Challenge Datasets

Premise: "I have done what you asked."
Hypothesis: "I have disobeyed your orders."

Word Overlap Challenge Dataset
Premise: "I have done what you asked."
Hypothesis: "I have disobeyed your orders and true is true."

Spelling Errors Challenge Dataset
Premise: "I have done what you asked."
Hypothesis: "I have disobeyed your orders ordets."

[Naik and Ravichander et al., 2018]
Small Perturbations Break NLI Models

**Word Overlap**

- MultiNLI Dev Evaluation: 75%
- Word Overlap Dataset: 62.4%
  - **-12.6% (absolute)**

**Spelling Errors**

- MultiNLI Dev Evaluation: 74%
- Spelling Errors Dataset: 69.2%
  - **-4.8% (absolute)**
Inoculating NLI models

Word Overlap

- MultiNLI Dev
- Word Overlap Adversary

Spelling Errors

- MultiNLI Dev
- Spelling Errors Adversary

Accuracy vs. # of Fine-Tuning Examples
Inoculating NLI models

**Word Overlap**
- MultiNLI Dev
- Word Overlap Adversary

**Spelling Errors**
- MultiNLI Dev
- Spelling Errors Adversary

**Dataset Weakness**

**Model Weakness**
More Examples in the Paper!

**Outcome 1**
- (a) Word Overlap
- Dataset Weakness

**Outcome 2**
- (c) Spelling Errors
- Model Weakness

**Outcome 3**
- (e) Numerical Reasoning
- Predictive Artifacts / Other
SQuAD

Question: "The number of new Huguenot colonists declined after what year?"

Passage: "The largest portion of the Huguenots to settle in the Cape arrived between 1688 and 1689...but quite a few arrived as late as 1700; thereafter, the numbers declined..."

Correct Answer: "1700"
Question: "The number of new Huguenot colonists declined after what year?"

Passage: "The largest portion of the Huguenots to settle in the Cape arrived between 1688 and 1689…but quite a few arrived as late as 1700; thereafter, the numbers declined. The number of old Acadian colonists declined after the year of 1675."

Correct Answer: "1700"
Small Perturbations Break SQuAD Models

-24.5 F1 (absolute)
Inoculating SQuAD models

![Graph showing F1 score vs. number of fine-tuning examples for SQuAD Dev and Adversarial SQuAD.]
Inoculating SQuAD models

Predictive Artifacts
/ Other

F1 vs. # of Fine-Tuning Examples

SQuAD Dev vs. Adversarial SQuAD
Takeaways

• Inoculation by Fine-Tuning helps us understand why our models fail.

• While all challenge datasets break our models, they stress them in different ways.

• Potentially many situations where inoculation can help clarify model results when transferring to other datasets.
Takeaways

• Inoculation by Fine-Tuning helps us understand why our models fail.

• While all challenge datasets break our models, they stress them in different ways.

• Potentially many situations where inoculation can help clarify model results when transferring to other datasets.
Limitations of Inoculation by Fine-Tuning

- Requires a somewhat balanced label distribution in the challenge dataset.
  - Else, fine-tuned model will always predict majority label
- This method is not a silver bullet!
  - First step toward disentangling failures of \{original / challenge\} datasets and models.
Standard Challenge Evaluation

**Outcome:**
- Challenge is difficult for the model.
- **Why?**

Inoculation by Fine-Tuning

**Possible Outcomes:**
1. Dataset Weakness
2. Model Weakness
3. Annotation Artifacts, Other
Inoculating Multiple SQuAD Reading Comprehension Models

- Original (BiDAF)
- Challenge (BiDAF)
- Original (QANet)
- Challenge (QANet)

Graph shows the performance of different models over the number of fine-tuning examples.
Inoculating Multiple NLI Models Against Word Overlap Adversary

![Graph showing the effect of fine-tuning on accuracy for original and challenge models.](image-url)
Inoculating Multiple NLI Models Against Spelling Errors

![Graph showing Accuracy vs. Number of Fine-Tuning Examples for different models: Original (ESIM), Challenge (ESIM), Original (DA), Challenge (DA), Original (char-level), Challenge (char-level).]