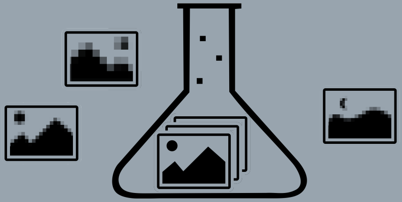
Reliable JPEG Forensics via Model Uncertainty

Detecting the training-test mismatch with Bayesian logistic regression

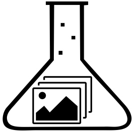
Benedikt Lorch, Anatol Maier, Christian Riess
IT Security Infrastructures Lab
Friedrich-Alexander University Erlangen-Nuremberg, Germany
December 9, 2020



The training-test mismatch in JPEG forensics

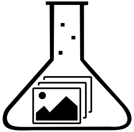
The training-test mismatch in JPEG forensics

Train in controlled lab environment



The training-test mismatch in JPEG forensics

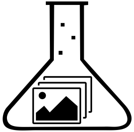
Train in controlled lab environment



Test accuracy: 99%

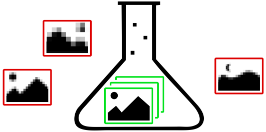
The training-test mismatch in JPEG forensics

Train in controlled lab environment



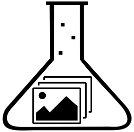
Test accuracy: 99%

Test on images of unknown quality



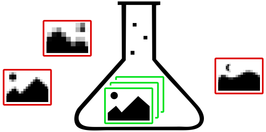
The training-test mismatch in JPEG forensics

Train in controlled lab environment



Test accuracy: 99%

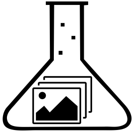
Test on images of unknown quality



Test accuracy: $\sim$ random guessing

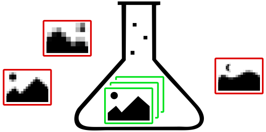
The training-test mismatch in JPEG forensics

Train in controlled lab environment



Test accuracy: 99%

Test on images of unknown quality

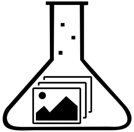


Test accuracy: $\sim$ random guessing

- Detectors do not naturally generalize to unseen JPEG settings
- ... and fail silently.

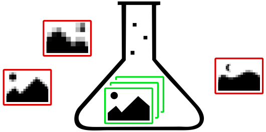
The training-test mismatch in JPEG forensics

Train in controlled lab environment



Test accuracy: 99%

Test on images of unknown quality



Test accuracy: $\sim$ random guessing

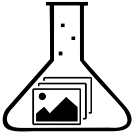
- Detectors do not naturally generalize to unseen JPEG settings
- ... and fail silently.

Current approaches to mitigating the training-test mismatch

1. Create more robust detectors with broad applicability (open challenge)

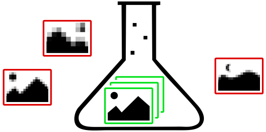
The training-test mismatch in JPEG forensics

Train in controlled lab environment



Test accuracy: 99%

Test on images of unknown quality



Test accuracy: $\sim$ random guessing

- Detectors do not naturally generalize to unseen JPEG settings
- ... and fail silently.

Current approaches to mitigating the training-test mismatch

1. Create more robust detectors with broad applicability (open challenge)
2. Create several detectors specialized to a narrow range of JPEG settings (not fool-proof)

Contribution: Detect training-test mismatch with Bayesian detector

Our proposal: **Create reliable detectors that express uncertainty in unfamiliar situations**

$\Rightarrow$ Quantify when to trust the model's predictions

Contribution: Detect training-test mismatch with Bayesian detector

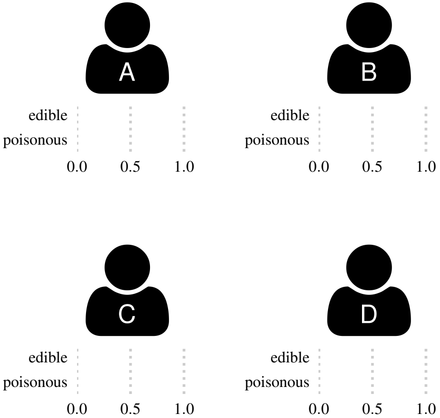
Our proposal: **Create reliable detectors that express uncertainty in unfamiliar situations**

$\Rightarrow$ Quantify when to trust the model's predictions

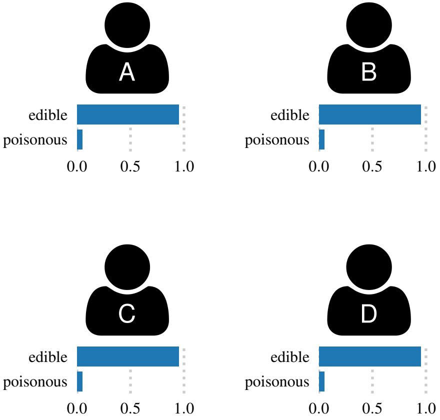
Experiments

- Detect JPEG double compression based on first-digit features
- Uncertainty measure allows anticipating misclassifications when test image is not aligned with the training data
  - Mismatch in JPEG quality factors
  - Mismatch in quantization tables $\leftarrow$ this talk
  - Mismatch in DCT implementation

Data and model uncertainty

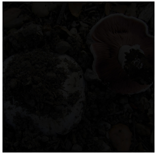


Data and model uncertainty

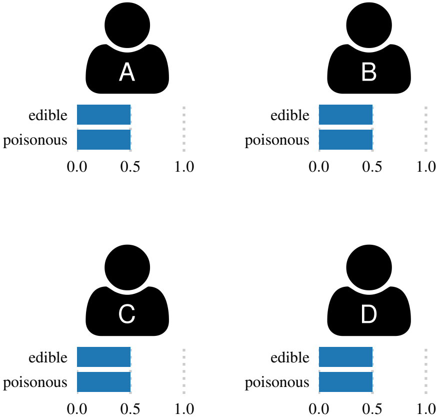


1. No uncertainty: All experts agree

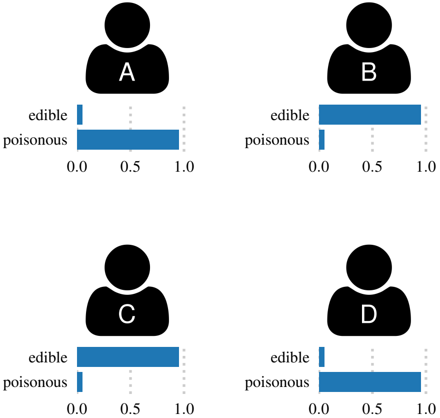
Data and model uncertainty



1. No uncertainty: All experts agree
2. Data uncertainty: All experts are uncertain



Data and model uncertainty



1. No uncertainty: All experts agree
2. Data uncertainty: All experts are uncertain
3. **Model uncertainty: Experts have different opinions**

Bayesian logistic regression

Bayesian inference of predictive distribution

- Express uncertainty about decision boundary by modeling weights as probability distributions

Bayesian inference of predictive distribution

- Express uncertainty about decision boundary by modeling weights as probability distributions
- Goal: Obtain **predictive distribution** over possible outcomes instead of single estimate

Bayesian inference of predictive distribution

- Express uncertainty about decision boundary by modeling weights as probability distributions
- Goal: Obtain **predictive distribution** over possible outcomes instead of single estimate
- Mean of predictive distribution gives prediction, variance indicates uncertainty

Bayesian inference of predictive distribution

- Express uncertainty about decision boundary by modeling weights as probability distributions
- Goal: Obtain **predictive distribution** over possible outcomes instead of single estimate
- Mean of predictive distribution gives prediction, variance indicates uncertainty

$$p(\boldsymbol{y}^* \mid \boldsymbol{x}^*, \boldsymbol{x}_{\text{train}}, \boldsymbol{y}_{\text{train}}) = \int p(\boldsymbol{y}^* \mid \boldsymbol{x}^*, \boldsymbol{w})\, p(\boldsymbol{w} \mid \boldsymbol{x}_{\text{train}}, \boldsymbol{y}_{\text{train}})\, d\boldsymbol{w} \tag{1}$$

Bayesian inference of predictive distribution

- Express uncertainty about decision boundary by modeling weights as probability distributions
- Goal: Obtain **predictive distribution** over possible outcomes instead of single estimate
- Mean of predictive distribution gives prediction, variance indicates uncertainty

$$p(\boldsymbol{y}^* \mid \boldsymbol{x}^*, \boldsymbol{x}_{\text{train}}, \boldsymbol{y}_{\text{train}}) = \int p(\boldsymbol{y}^* \mid \boldsymbol{x}^*, \boldsymbol{w})\, p(\boldsymbol{w} \mid \boldsymbol{x}_{\text{train}}, \boldsymbol{y}_{\text{train}})\, d\boldsymbol{w} \tag{1}$$

with

- $p(\boldsymbol{y}^* \mid \boldsymbol{x}^*, \boldsymbol{w})$ - prediction of classifier with weights $\boldsymbol{w}$

Bayesian inference of predictive distribution

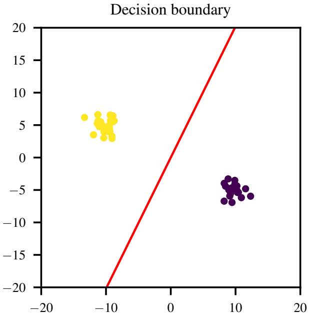
- Express uncertainty about decision boundary by modeling weights as probability distributions
- Goal: Obtain **predictive distribution** over possible outcomes instead of single estimate
- Mean of predictive distribution gives prediction, variance indicates uncertainty

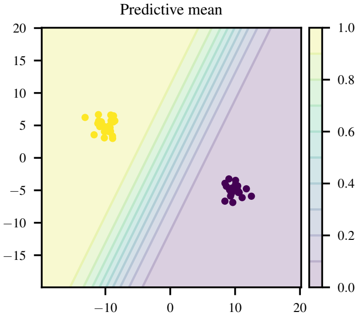
$$p(\boldsymbol{y}^* \mid \boldsymbol{x}^*, \boldsymbol{x}_{\text{train}}, \boldsymbol{y}_{\text{train}}) = \int p(\boldsymbol{y}^* \mid \boldsymbol{x}^*, \boldsymbol{w})\, p(\boldsymbol{w} \mid \boldsymbol{x}_{\text{train}}, \boldsymbol{y}_{\text{train}})\, d\boldsymbol{w} \tag{1}$$

with

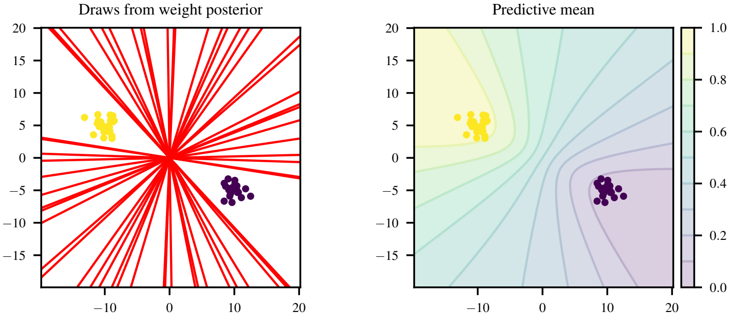
- $p(\boldsymbol{y}^* \mid \boldsymbol{x}^*, \boldsymbol{w})$ - prediction of classifier with weights $\boldsymbol{w}$
- $p(\boldsymbol{w} \mid \boldsymbol{x}_{\text{train}}, \boldsymbol{y}_{\text{train}})$ - posterior distribution over the weights after training data is seen

Toy example: Standard logistic regression

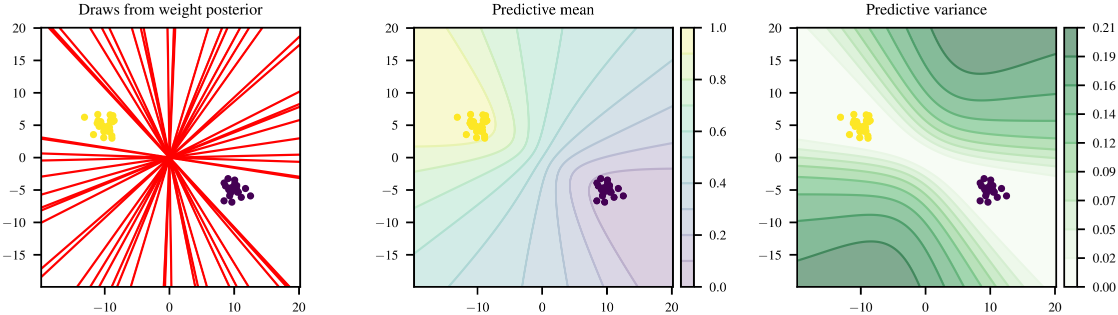




Toy example: Bayesian logistic regression



Toy example: Bayesian logistic regression



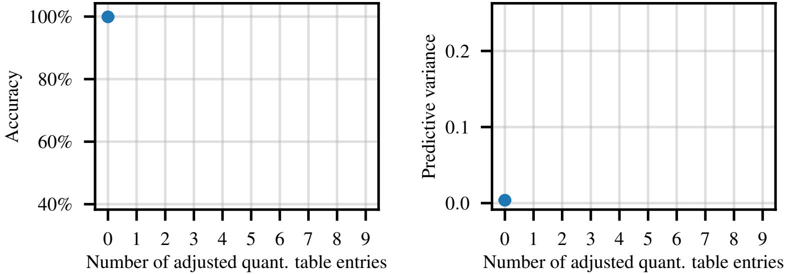
Experiments & Results

Application scenario: Mismatch in JPEG quantization tables

- Minor discrepancy between training and test quantization tables cause misclassifications

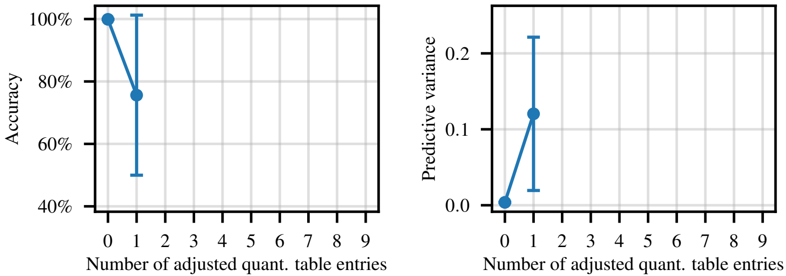
Application scenario: Mismatch in JPEG quantization tables

- Minor discrepancy between training and test quantization tables cause misclassifications



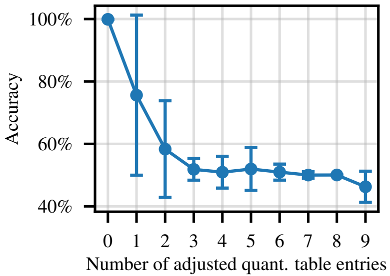
Application scenario: Mismatch in JPEG quantization tables

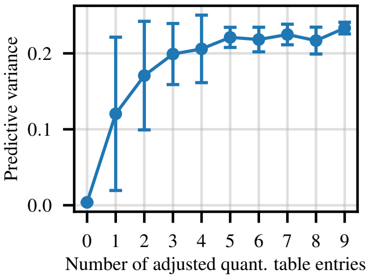
- Minor discrepancy between training and test quantization tables cause misclassifications
- Experiment: Randomly select *i* quantization table entries, adjust quantization factor by ± 1



Application scenario: Mismatch in JPEG quantization tables

- Minor discrepancy between training and test quantization tables cause misclassifications
- Experiment: Randomly select *i* quantization table entries, adjust quantization factor by ± 1

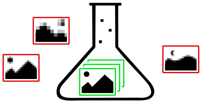




$\Rightarrow$ Bayesian detector anticipates misclassifications from quantization table mismatch

Conclusion

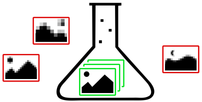
Conclusion: Reliable detectors from model uncertainty



- Machine learning models are sensitive to training-test mismatches
- Forensic methods are often faced with data from unknown origins

  $\Rightarrow$ Forensic methods must take care of training-test mismatch (instead of failing silently)

Conclusion: Reliable detectors from model uncertainty



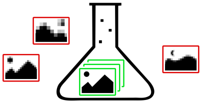
- Machine learning models are sensitive to training-test mismatches
- Forensic methods are often faced with data from unknown origins

  $\Rightarrow$ Forensic methods must take care of training-test mismatch (instead of failing silently)

Proposal: Bayesian detector indicates training-test mismatch via model uncertainty

- Quantify when to trust in the model's prediction
- Avoid misclassifications on unseen compression settings
- Applicable to neural networks but requires restrictive approximations

Conclusion: Reliable detectors from model uncertainty



- Machine learning models are sensitive to training-test mismatches
- Forensic methods are often faced with data from unknown origins

  $\Rightarrow$ Forensic methods must take care of training-test mismatch (instead of failing silently)

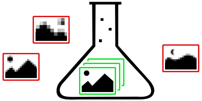
Proposal: Bayesian detector indicates training-test mismatch via model uncertainty

- Quantify when to trust in the model's prediction
- Avoid misclassifications on unseen compression settings
- Applicable to neural networks but requires restrictive approximations

Long term goal

- Foster research on reliable, trustworthy learning-based methods

Conclusion: Reliable detectors from model uncertainty



- Machine learning models are sensitive to training-test mismatches
- Forensic methods are often faced with data from unknown origins

$\Rightarrow$ Forensic methods must take care of training-test mismatch (instead of failing silently)

Proposal: Bayesian detector indicates training-test mismatch via model uncertainty

- Quantify when to trust in the model's prediction
- Avoid misclassifications on unseen compression settings
- Applicable to neural networks but requires restrictive approximations

Long term goal

- Foster research on reliable, trustworthy learning-based methods

Thank you

References

- Tube icon adapted from environmental science icon
- Mushroom photos from Wikipedia [1, 2]