Improving Surprise Adequacy with GTSRB Research

Elijah Higgs
eohiggs@andrew.cmu.edu

Background

- **Safety Critical Systems**
  - Autonomous systems like self driving cars rely on model reliability
  - Severe consequences of failures
- **Ensemble Methods**
  - Combining models to increase robustness
  - Building on metrics such as Surprise Adequacy to identify model errors
- **Surprise Adequacy**
  - Inputs significantly different from training data
  - More likely to cause incorrect predictions
  - Identifying faulty model scenarios
- **New Dataset**
  - GTSRB

Previous Setup (CIFAR-10)

- Dataset
  - 10 different training and validation sets to train diverse models
  - 10,000 test samples
  - 45,000 training samples
  - 5,000 validation samples

- VGG10 Architecture
  - Train models on random splits
  - Extract activation traces and calculate MDSA scores
  - Aggregate predictions using max voting

Previous Findings

- 1.3% increase compared to average accuracy
- True Positive (TP): 8631 - Incorrect classifications marked as Surprising
- False Negative (FN): 205 - Incorrect classifications marked as Unsurprising
- False Positive (FP): 144 - Correct classifications marked as Surprising
- True Negative (TN): 1020 - Correct classifications marked as Unsurprising
- Total Classifications: 10000

```
Surprise accuracy of model 0: 0.9538
Surprise accuracy of model 1: 0.9565
Surprise accuracy of model 2: 0.9565
Surprise accuracy of model 3: 0.9767
Surprise accuracy of model 4: 0.9638
Surprise accuracy of model 5: 0.9718
Surprise accuracy of model 6: 0.9736
Surprise accuracy of model 7: 0.9699
Surprise accuracy of model 8: 0.984
Surprise accuracy of model 9: 0.9798
Average surprise accuracy of individual models: 0.96864
Surprise accuracy of the ensemble model: 0.982
```

New Setup

- German Traffic Sign Recognition Benchmark
  - More than 40 classes
  - More than 50,000 images

- Process
  - Train 3 models on random splits (odd number for majority)
  - Modify final dense layer to use standalone softmax layer
  - Aggregate using various surprise thresholds
  - VGG10 model

Results

Individual and Ensemble Surprise Results

- Model 0 - TP: 45, FN: 0, FP: 522, TN: 7274
  - Classification accuracy of Model 0: 99.43%
- Model 1 - TP: 23, FN: 0, FP: 536, TN: 7282
  - Classification accuracy of Model 1: 99.71%
- Model 2 - TP: 16, FN: 0, FP: 553, TN: 7272
  - Classification accuracy of Model 2: 99.80%
- Ensemble - TP: 9, FN: 0, FP: 1028, TN: 6804
  - Classification accuracy of the ensemble model: 99.89%

TP: Surprising, misclassified
FP: Surprising, correctly classified
FN: Not surprising, misclassified
TN: Not surprising, correctly classified

Surprise Thresholds

Threshold: 70%

- Model 0: TP: 45, FN: 0, FP: 2307, TN: 5489
- Model 1: TP: 23, FN: 0, FP: 2329, TN: 5489
- Model 2: TP: 16, FN: 0, FP: 2336, TN: 5489
- Ensemble: TP: 9, FN: 0, FP: 2343, TN: 5489

Threshold: 80%

- Model 0: TP: 45, FN: 0, FP: 1523, TN: 6273
- Model 1: TP: 23, FN: 0, FP: 1545, TN: 6273
- Model 2: TP: 16, FN: 0, FP: 1552, TN: 6273
- Ensemble: TP: 9, FN: 0, FP: 1559, TN: 6273

Threshold: 90%

- Model 0: TP: 45, FN: 0, FP: 739, TN: 7057
- Model 1: TP: 23, FN: 0, FP: 761, TN: 7057
- Model 2: TP: 16, FN: 0, FP: 768, TN: 7057
- Ensemble: TP: 9, FN: 0, FP: 775, TN: 7057

Threshold: 95%

- Model 0: TP: 45, FN: 0, FP: 347, TN: 7449
- Model 1: TP: 23, FN: 0, FP: 369, TN: 7449
- Model 2: TP: 16, FN: 0, FP: 376, TN: 7449
- Ensemble: TP: 9, FN: 0, FP: 383, TN: 7449

Threshold: 99%

- Model 0: TP: 40, FN: 5, FP: 39, TN: 7757
- Model 1: TP: 22, FN: 1, FP: 57, TN: 7761
- Model 2: TP: 16, FN: 0, FP: 63, TN: 7762
- Ensemble: TP: 9, FN: 0, FP: 70, TN: 7762

New Findings

- Ensemble not outperforming individual models
  - Increased complexity
  - Less decisive aggregation
- Impact of thresholds
  - Higher thresholds reduce FP but also TP
- Wider range of surprise scores

Future Work

- Explore ensemble techniques that better handle variance
- Trying alternative SA calculations
- Different model architecture

Thanks!

Do you have any questions?

youremail@freepik.com
+34 654 321 432
yourwebsite.com