Attacking Compressed NLP

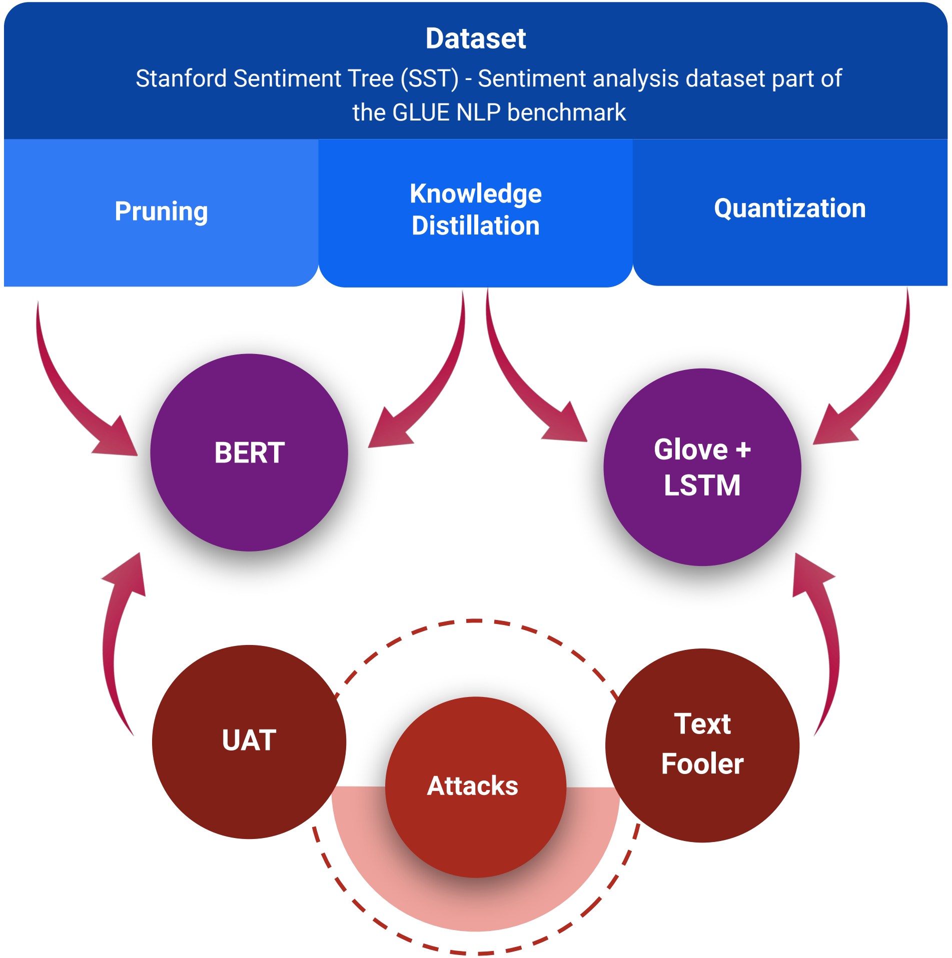
Swapnil Parekh (sp6646) Anant Singh (as14229)

Exec summary

- NLP models are increasingly getting embedded into industrial systems but their memory and power requirements makes deploying them to edge devices a challenging task.
- Model compression techniques are now widely used to deploy models on edge devices as they decreases the resource requirements and make model inference very fast and efficient.
- Their reliability and robustness from a security perspective is another major issue in safety-critical applications.
- Adversarial attacks are like optical illusions for machines and such samples can severely impact the accuracy and reliability of models.
- **Novel Contribution**: Investigate the transferability of adversarial samples across the SOTA NLP models and their compressed versions and infer the effects different compression techniques have on adversarial attacks

Approach

1. Effects of compression were tested on two models **BERT** and **Glove + LSTM**
2. **Knowledge Distillation** and **Pruning** was tested on BERT model and 8-bit **Quantization** and **Knowledge Distillation** on Glove + LSTM model
3. We used 2 types of attacks: White box attack- **Universal Adversarial Triggers** on BERT model and Black box attack- **TextFooler** on Glove + LSTM model
4. We report compression speedups and averaged accuracy before and after multiple universal attacks and textfooler attacks



Dataset
Stanford Sentiment Tree (SST) - Sentiment analysis dataset part of the GLUE NLP benchmark
Pruning
Knowledge Distillation
Quantization
BERT
Glove + LSTM
UAT
Attacks
Text Fooler

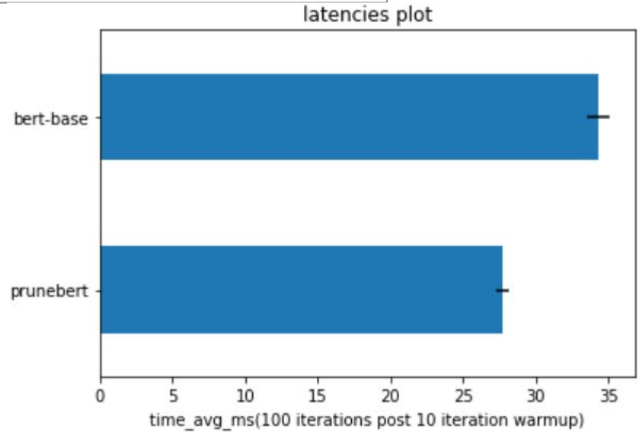
Results - BERT

BERT Distillation: White Box Attack (Targeted)

Attack Created / Tested	Base model	Distil-bert finetune(2x)	Distilled-bert(3x)
Base model	0.918, 0.113	0.901, 0.202	0.882, 0.162
Distil-bert finetune(2x)	0.918, 0.213	0.901, 0.054	NA
Distilled-bert(3x)	0.918, 0.305	NA	0.882, 0.0033

BERT Pruning: White Box Attack (Targeted)

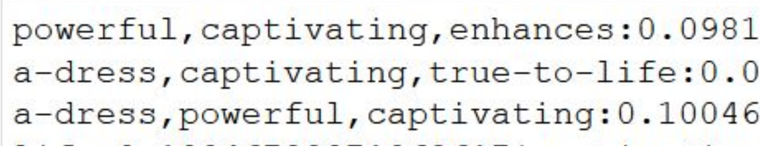
Attack Created / Tested	Base Model	Pruned Model(~2x)
Base model	0.918, 0.113	0.887,0.204
Pruned Model	0.918,0.284	0.887,0.0204



Results - Word2Vec + LSTM Model

Distillation (Born Again network) - White Box Attack (Targeted)

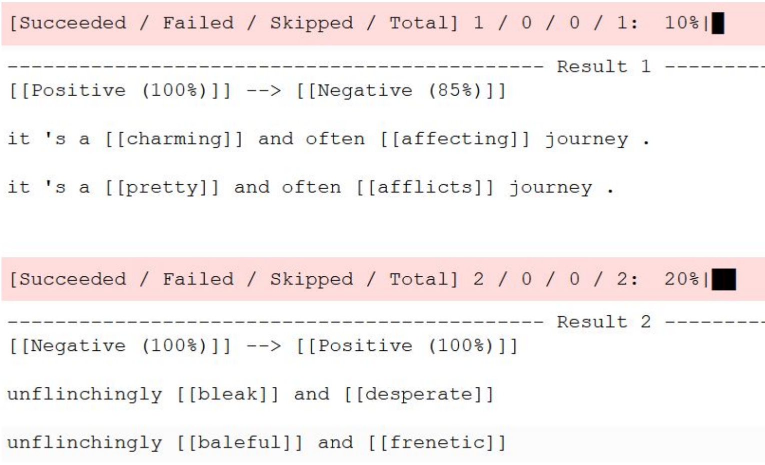
Attack Created / Tested	Base Model	Distilled Model
Base Model	0.827,0.0787	0.834,0.132
Distilled Model	0.827,0.0971	0.834,0.101



```
powerful,captivating,enhances:0.0981
a-dress,captivating,true-to-life:0.0
a-dress,powerful,captivating:0.10046
```

Quantization(8bit) - Black Box Attack(Misclass)

Attack Created / Tested	Base Model	Quantized Model
Base Model	0.856, 0.262	0.854, 0.230
Quantized Model	0.856, 0.320	0.854, 0.256



```
[Succeeded / Failed / Skipped / Total] 1 / 0 / 0 / 1:  10%|█
---------------------------------------------- Result 1 ---------
[[Positive (100%)]] --> [[Negative (85%)]]

it 's a [[charming]] and often [[affecting]] journey .

it 's a [[pretty]] and often [[afflicts]] journey .


[Succeeded / Failed / Skipped / Total] 2 / 0 / 0 / 2:  20%|██
---------------------------------------------- Result 2 ---------
[[Negative (100%)]] --> [[Positive (100%)]]

unflinchingly [[bleak]] and [[desperate]]

unflinchingly [[baleful]] and [[frenetic]]
```

Conclusions

- Compressed models are often more vulnerable to white/black box attacks
- In pruning: adversarial samples generated from base models are less effective on compressed models
- In Quantization: adversarial samples are transferable between compressed and uncompressed models, but are less effective on transferring attacks from the quantized to the base model.
- Distillation effects vary but are correlated to the performance of the compressed models.
- Separately finetuned DistilBERT model are less affected by attacks created on BERT rather than distilling the finetuned BERT itself.
- While compressed models may provide performance benefits, they do not provide much in way of security.
- https://github.com/95anantsingh/NYU-Attacks-on-Compressed-NLP
- Demo:https://drive.google.com/file/d/1jsp36A_Q_o9ySBLrUIo_T86kVbhqHMwW/view?usp=sharing