Improving the Quality of Adversarial Examples via Contrastive Learning and Pretraining

LYU2106 Final Year Project Term 1 Presentation

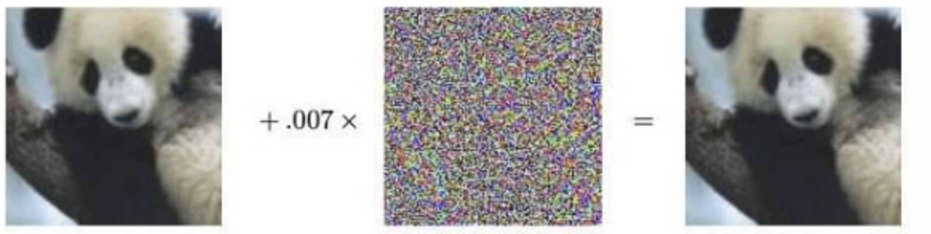
Yung-chieh Huang (1155120711)

Agenda

- Introduction
- Objective
- Contribution
- Methedology
- Model Composition
- Experiments
- Conclusion
- Future Work

Introduction – Adversarial Attack

- Adversarial attack is an approach to test the robustness of machine learning models, by intentionally apply perturbations to make the models misclassify.
- Ensure security in real-life applications.



Perfect performance by the actor	Positive (99%)
Spotless performance by the actor	Negative (100%)

Introduction – Adversarial Attack for Text

- Adversarial examples are generated by attack models, by replacing words in a sentence.
- A well-crafted adversarial example should have minimum perturbations and preserve the structure and charateristics of the original.
- An attack model is composed of:
  - Goal function
  - Transformation
  - Search method
  - Constraints

Objective

- The adversarial examples state-of-the-art attack models generate are of low quality, they contain opposite semantic replacements and irrelevant replacements.

Original sentence	no amount of good intentions is able to overcome the triv-iality of the story	Negative (100%)
Adversarial example	no amount of good intentions is able to overcome the beauty of the story	Positive (99%)

Original sentence	watching spirited away is like watching an eastern imagi-nation explode	Positive (99%)
Adversarial example	watching spirited away is like watching an eastern maga-zine explode	Negative (100%)

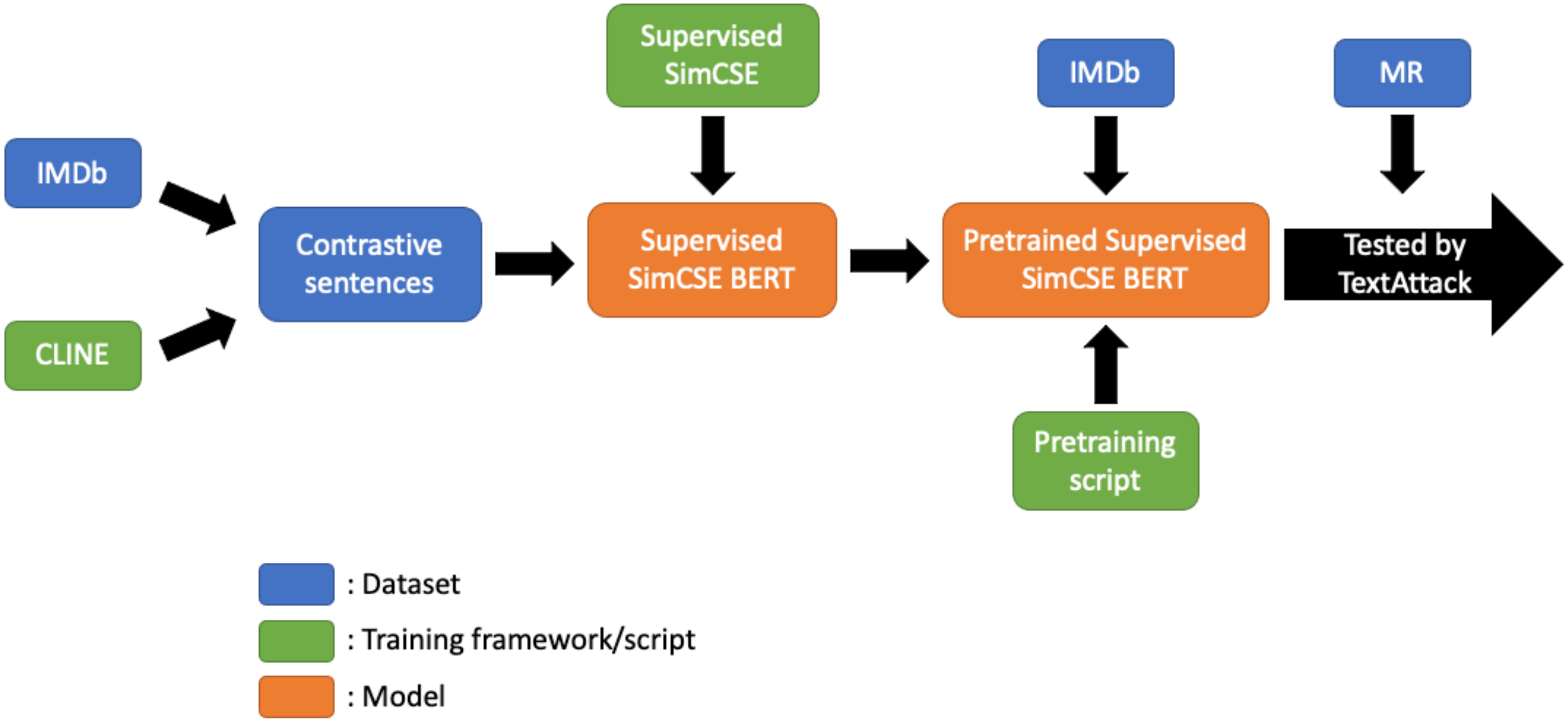
Objective

- Overcome the flaws in previous works and generate high quality adversarial examples.
- Free from opposite semantic or out-of-context replacements while maintaining fluency.
- Higher successful attack rate and lower perturbation.

Contribution

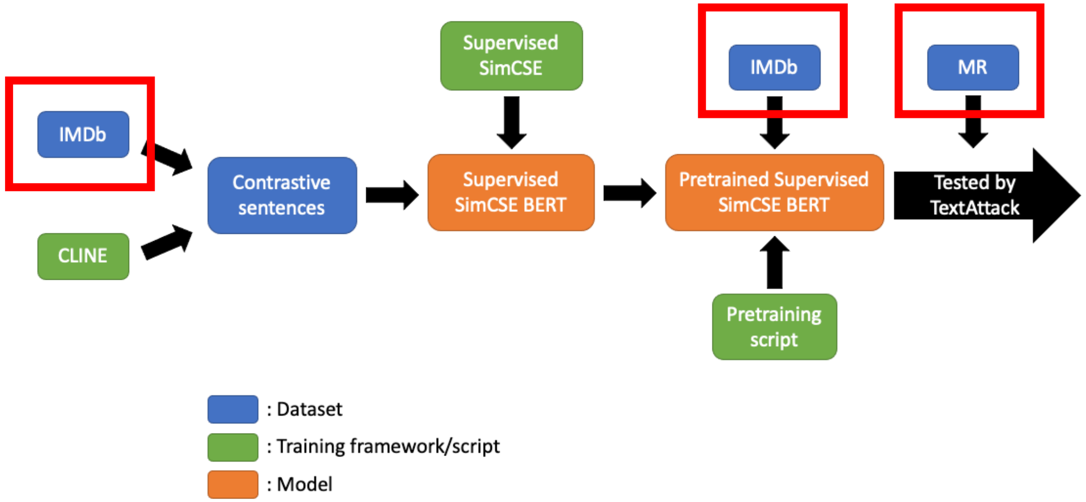
- Opposite semantic replacements are caused by the embedding space of language models. With contrastive learning, our attack model is capable of separating synonyms and antonyms.
- Out-of-context replacements exist because attack models are too general. We make our attack model domain-specific (movie reviews) through a second-phase pretraining.
- We are the first to generate adversarial examples via a combination of contrastive learning and pretraining.

Methedology



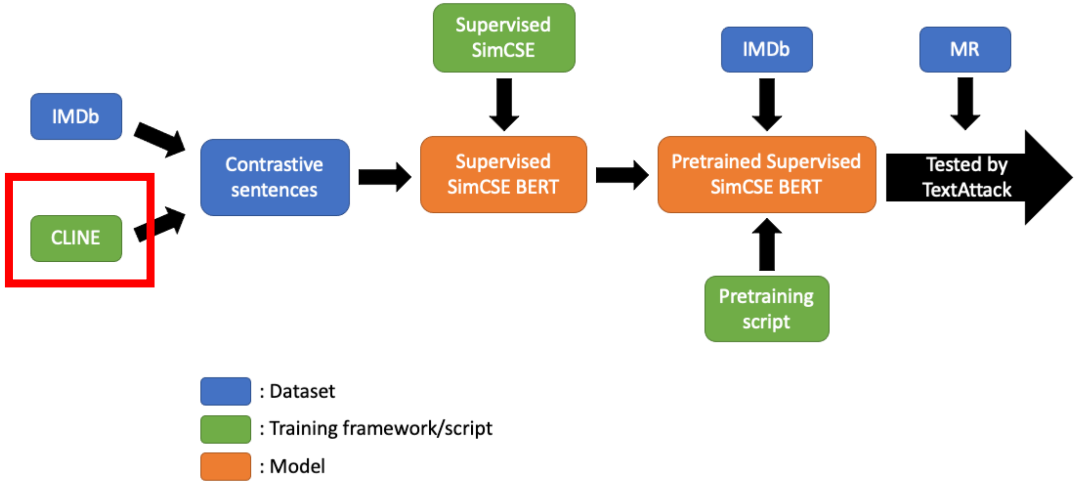
Methedology- Datasets

- IMDb (Mass et al. 2011): 25,000 highly polar movie reviews for training, 25,000 for testing, and additional 50,000 unlabeled data.
- MR (Pang and L. Lee 2005): 5,331 positive and 5,331 negative reviews from Rotten Tomatoes.

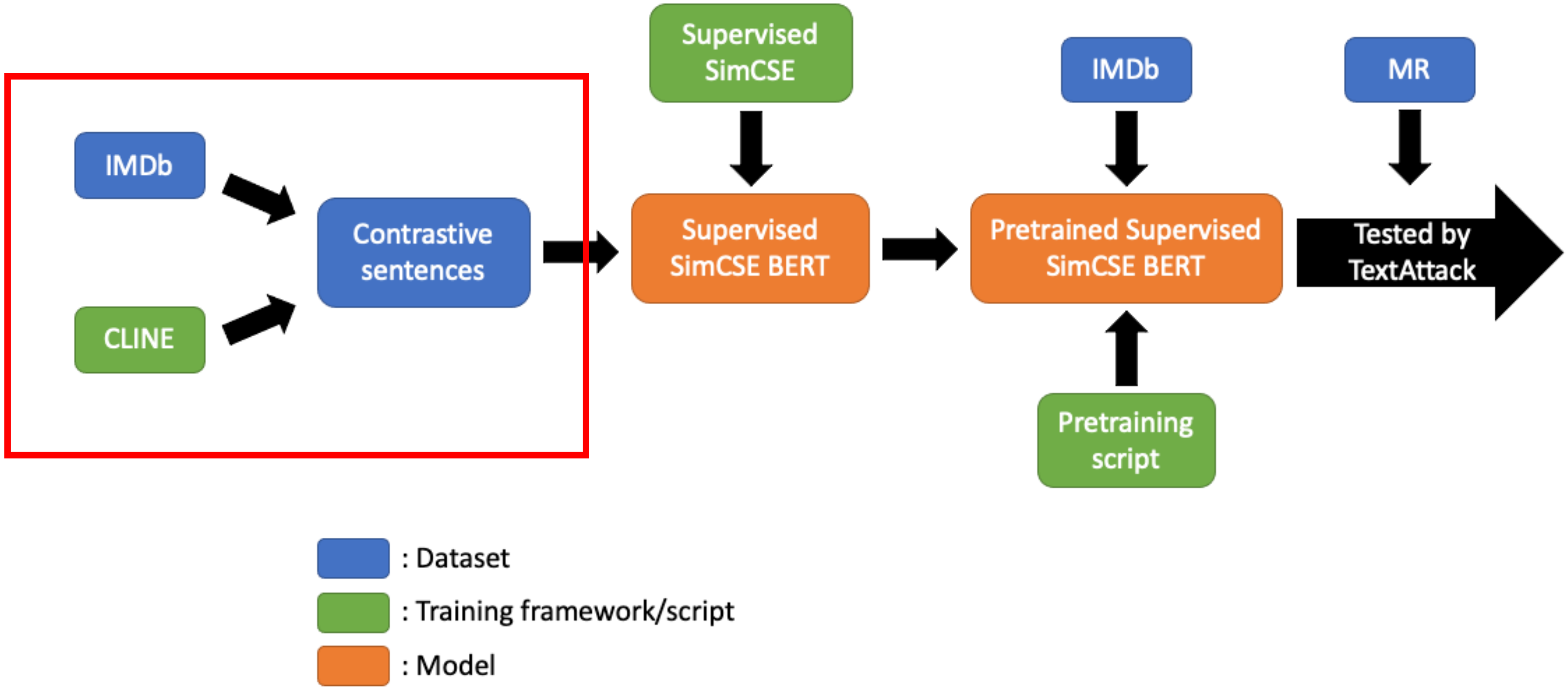


Methedology- CLINE

- Generate positive sentences by replacing words with synonyms.
- Generate negative sentences by replacing words with antonyms or random words.



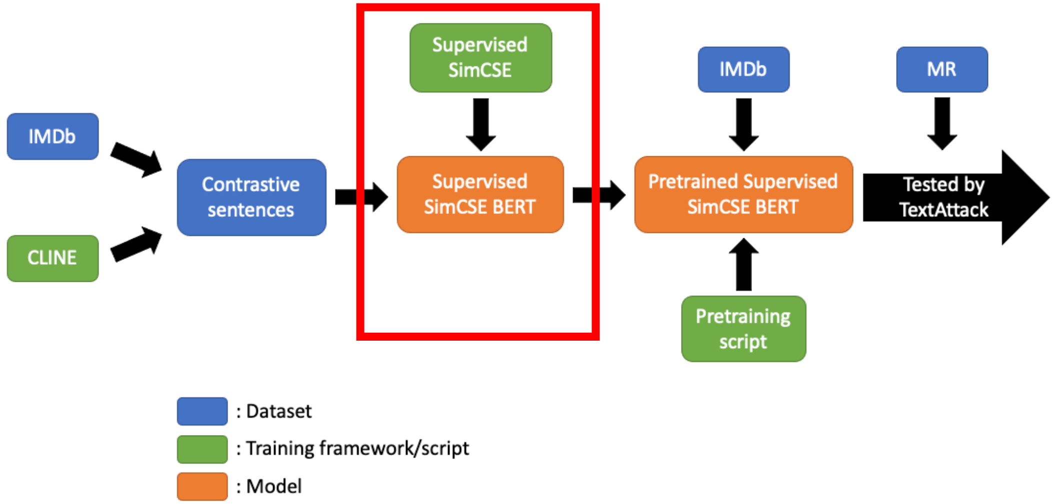
Methedology



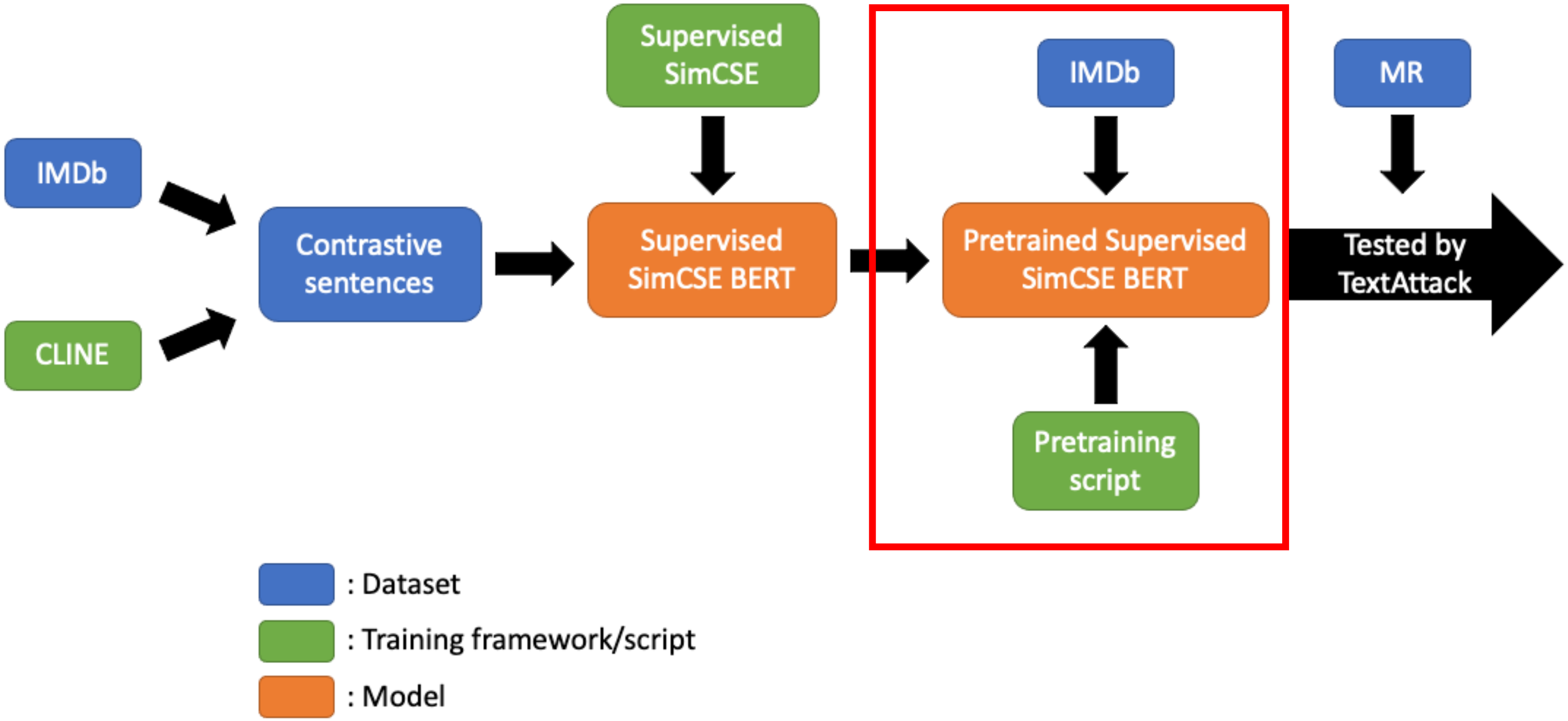
Methedology- SimCSE

- Pulling semantically close neighbors together and pushing apart non-neighbors.
- The training objective is defined by:

$$-\log \frac{e^{\mathrm{sim}(\mathbf{h}_i,\mathbf{h}_i^+)/\tau}}{\sum_{j=1}^{N}\left(e^{\mathrm{sim}(\mathbf{h}_i,\mathbf{h}_j^+)/\tau} + e^{\mathrm{sim}(\mathbf{h}_i,\mathbf{h}_j^-)/\tau}\right)}$$

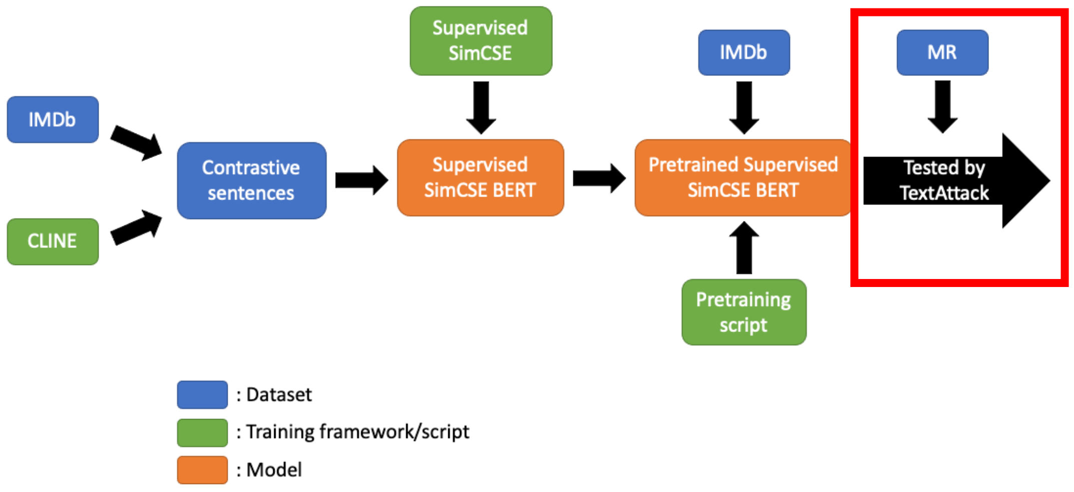


Methedology



Methedology- TextAttack

- A framework to evaluate different NLP attacks.
- Generate adversarial examples from a given dataset using an attack recipe and attack a victim model.



Methedology – Baseline

- We use BAE (Garg and Ramakrishnan 2020) as our baseline attack model.
- BAE uses BERT to predict masked tokens and apply constraints to ensure fluency.

Model Composition

- Goal function: untargeted classification.
- Transformation: our own pretrained supervised SimCSE BERT.
- Search method: greedy word swap, importance order.
- Constraints: Part of Speech, Universal Sentence Encoder.

Experiments – Pretraining only

- Pretrain a regular BERT-base on IMDb for 50,000 steps.

Dataset: MR		
	BAE	Ours
Number of successful attacks	473	**475**
Number of failed attacks	365	**363**
Number of skipped attacks	162	162
Original accuracy	83.8%	83.8%
Accuracy under attack	36.5%	**36.3%**
Attack success rate	56.44%	**56.68%**
Average perturbed word %	13.91%	**13.37%**
Average number of words per input	18.64	18.64
Average number of queries	63.49	**63.19**

Experiments – Pretraining only

- The replacements are more related to movies. However, there are still a considerable amount of opposite semantic and out-of-context replacements.

Original sentence	the movie is a little tired; maybe the original inspiration has run its course	Negative (100%)
BAE	the mind is a little tired; yet the original memory has continued its course	Positive (100%)
Ours	the beginning is a little tired; maybe the original tale has improved its course	Positive (88%)
Original sentence	one of the funnier movie in town	Positive (94%)
BAE	one of the funnier locations in town	Negative (97%)
Ours	one of the funnier scenes in town	Negative (99%)

Experiments – Contrastive Learning and Pretraining

- Instead of pretraining a regular BERT-base, now we pretrain supervised SimCSE BERT-base on IMDb for different number of steps.
- The one trained for 2,500 steps have the best overall performance.

Dataset: MR					
	BAE	Ours (50,000)	Ours (25,000)	Ours (5,000)	Ours (2,500)
Number of successful attacks	473	471	473	487	**501**
Number of failed attacks	365	367	365	351	**337**
Number of skipped attacks	162	162	162	162	162
Original accuracy	83.8%	83.8%	83.8%	83.8%	83.8%
Accuracy under attack	36.5%	36.7%	36.5%	35.1%	**33.7%**
Attack success rate	56.44%	56.21%	56.44%	58.11%	**59.79%**
Average perturbed word %	13.91%	13.19%	**13.13%**	13.58%	13.17%
Average number of words per input	18.64	18.64	18.64	18.64	18.64
Average number of queries	63.49	64.27	64.05	64.01	**62.96**

Experiments – Contrastive Learning and Pretraining

Original sentence	fans of the modern day hong kong action film finally have the worthy successor to a better tomorrow and the killer which they have been patiently waiting for	Positive (100%)	
BAE	fans of the modern day hong kong action film finally have the only successor to a better tomorrow and the killer which they have been helplessly waiting for	Negative (99%)	Low quality
Ours (50,000)	fans of the modern day hong kong action film finally have the disappointing successor to a better tomorrow and the killer which they have been patiently waiting for	Negative (51%)	Low quality
Ours (25,000)		Failed	Unsuccessful
Ours (5,000)		Failed	Unsuccessful
Ours (2,500)	fans of the modern day hong kong action movie now have the usual successor to a better tomorrow and the killer which they have been already waiting for	Negative (83%)	High quality and successful attack

Experiments – Using CLINE to Create Contrastive Sentences

- We create our own contrastive sentences using IMDb. We refer to the word replace script by CLINE.
- Then we train a supervised SimCSE BERT with the contrastive sentences.
- Finally, we pretrain the supervised SimCSE BERT on IMDb for 2,500 steps.

Experiments – Using CLINE to Create Contrastive Sentences

Dataset: MR			
	BAE	Ours (pre-training only)	Ours (IMDb contrastive sentences)
Number of successful attacks	473	475	**495**
Number of failed attacks	365	363	**343**
Number of skipped attacks	162	162	162
Original accuracy	83.8%	83.8%	83.8%
Accuracy under attack	36.5%	36.3%	**34.3%**
Attack success rate	56.44%	56.68%	**59.07%**
Average perturbed word %	13.91%	**13.37%**	13.5%
Average number of words per input	18.64	18.64	18.64
Average number of queries	63.49	**63.19**	63.77

Experiments – Using CLINE to Create Contrastive Sentences

- Don’t have enough contrastive sentences.
- The training strategy SimCSE uses is not suitable for our goal.

Conclusion

- Pretraining and contrastive learning have positive effects on generating high quality examples.
- Alter the embedding space by contrastive learning.
- Make our attack model domain-specific by a second-phase pretraining.
- Our attack model has better results than the baseline model.

Future Work

- Better method to combine contrastive learning and pretraining.
- Conduct larger scale experiment.
- Involve human evaluation to demonstrate the effectiveness.

Thank you