Advanced NLP

Natural Language Inference using InfoBERT

Ashna Dua - 2021101072
Prisha - 2021101075
Vanshita Mahajan - 2021101102

Code Link

Index

Problem Statement

- The Adversarial NLI (ANLI) dataset provides a challenging benchmark, curated with human intervention, to test and improve the robustness of NLI models against adversarial attacks and real-world complexities.

- The task is to implement two regularizers to enhance the performance of existing models such as BERT and InfoBERTa:

  - **Information Bottleneck:** Retains minimal information related to input X while retaining maximal information related to predicting the target Y.

  - **Anchored Feature Regularizer:** Aligning global representations with stable, task-relevant local features.

Related Work

Textual Adversarial Attacks

Early methods focused on word/character-level manipulations (e.g., white-box gradient-based attacks by Ebrahimi et al., 2018).

Defenses Against Attacks

- Adversarial Training: Uses adversarial examples for data augmentation but struggles against unknown attacks.
- Interval Bound Propagation (IBP): Certifies worst-case robustness but has limited adaptability to transformer models.
- Randomized Smoothing: Adds synonym-based noise for robustness guarantees but relies on comprehensive synonym sets.

Related Work

Representation Learning

- Mutual Information (MI) maximization principles (e.g., InfoNCE, van den Oord et al., 2018) used in self-supervised learning.
- InfoBERT extends this by integrating information-theoretic principles for robust training in discrete language inputs.

Dataset Used

Text	Judgments	Hypothesis
A man inspects the uniform of a figure in some East Asian country.	contradiction C C C C C	The man is sleeping
An older and younger man smiling.	neutral N N E N N	Two men are smiling and laughing at the cats playing on the floor.
A black race car starts up in front of a crowd of people.	contradiction C C C C C	A man is driving down a lonely road.
A soccer game with multiple males playing.	entailment E E E E E	Some men are playing a sport.
A smiling costumed woman is holding an umbrella.	neutral N N E C N	A happy woman in a fairy costume holds an umbrella.

SNLI

- **Stanford Natural Language Inference** dataset consists of 570k sentence-pairs manually labeled as entailment, contradiction, and neutral.

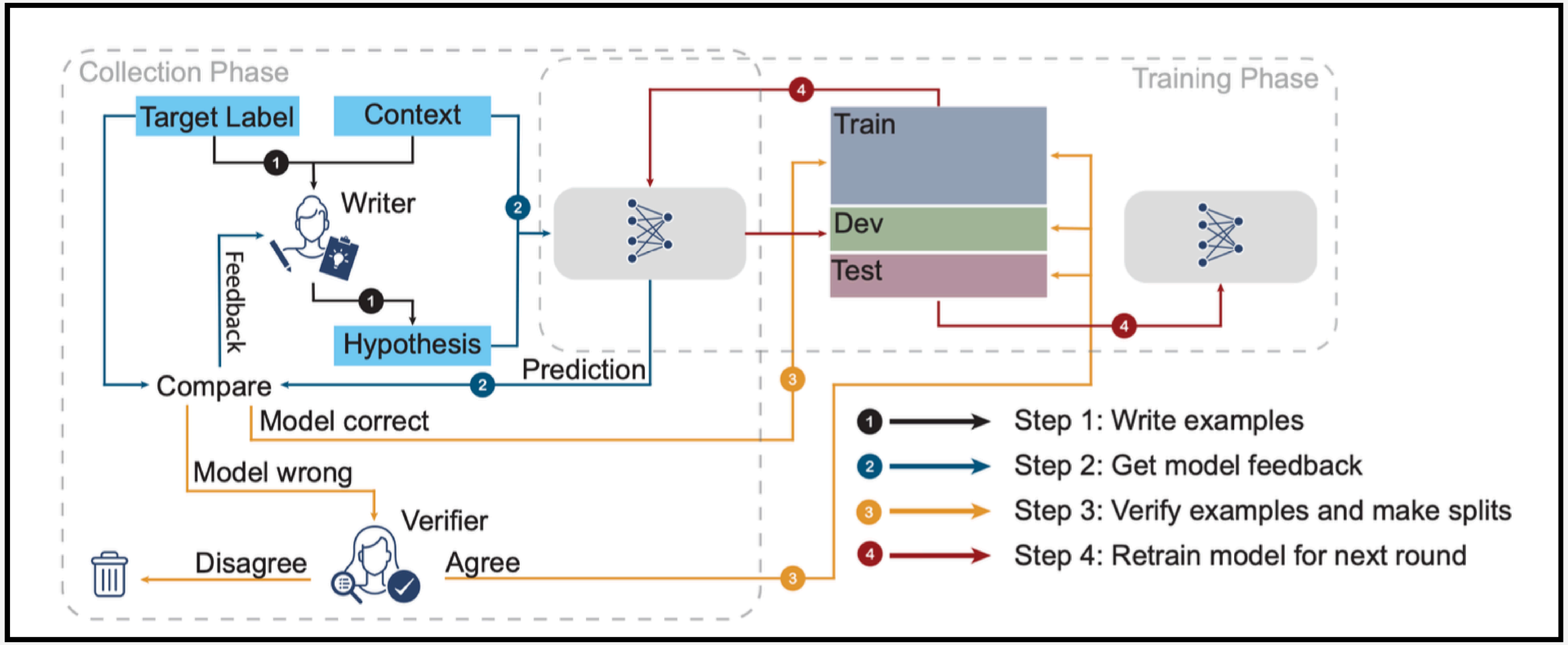
- Premises are image captions from Flickr30k, while hypotheses were generated by crowd-sourced annotators who were shown a premise and asked to generate entailing, contradicting, and neutral sentences.

MNLI

- The Multi-Genre Natural Language Inference (MultiNLI) dataset has 433K sentence pairs. Its size and mode of collection are modeled closely like SNLI.

- MultiNLI offers ten distinct genres (Face-to-face, Telephone, 9/11, Travel, Letters, Oxford University Press, Slate, Verbatim, Goverment and Fiction) of written and spoken English data.

Dataset Used



ANLI

A large-scale NLI benchmark collected through an iterative adversarial human-and-model-in-the-loop process. It consists of three rounds, each introducing progressively more challenging examples, requiring deeper reasoning and understanding from models.

Textual Adverserial Example

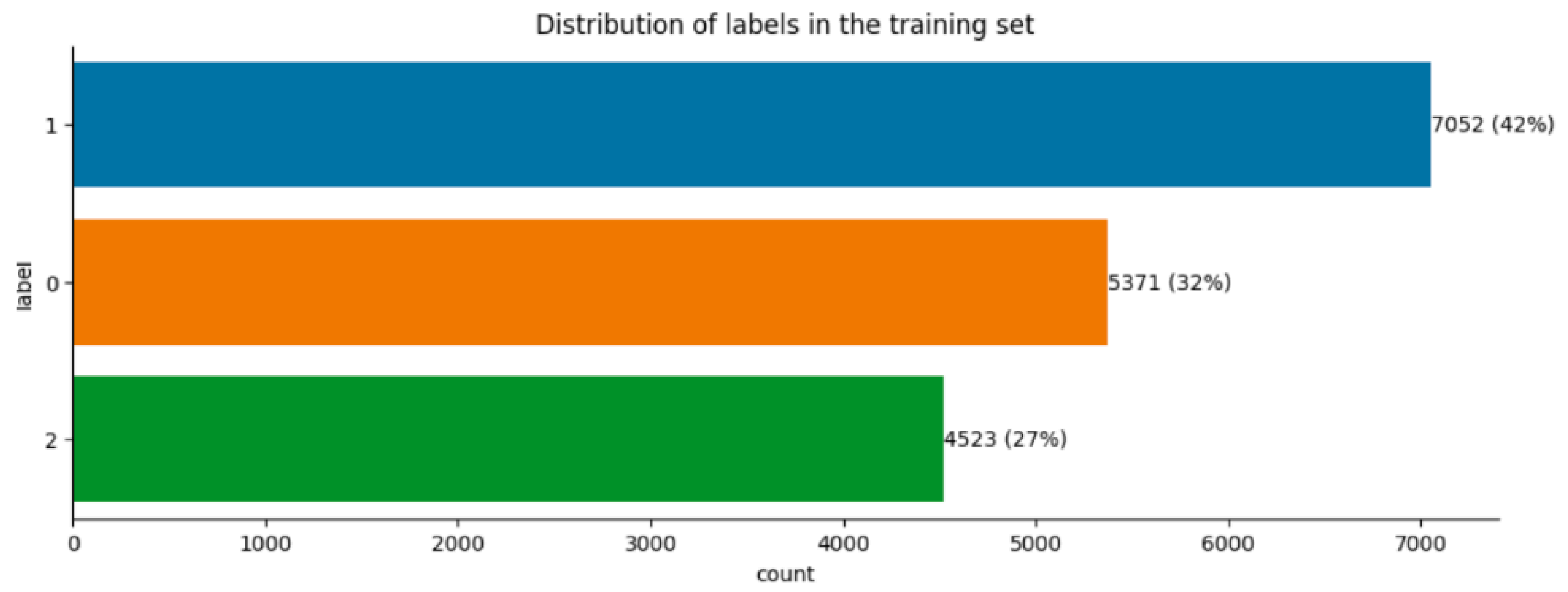
Original sentence: *x = [x1; x2; …; xn]*
Adverserial Sentence: *x′ = [x1′; x2′; …; xn′]*

For a classifier F, this satisfies:

- F(x) = o(x) = o(x′), but F(x′) != o(x′) where o(·) is the oracle (e.g., human decision-maker)
- ||ti − t′i||^2 ⩽ for i = 1, 2, …, n, where n ⩾ 0 and ti is the word embedding of xi

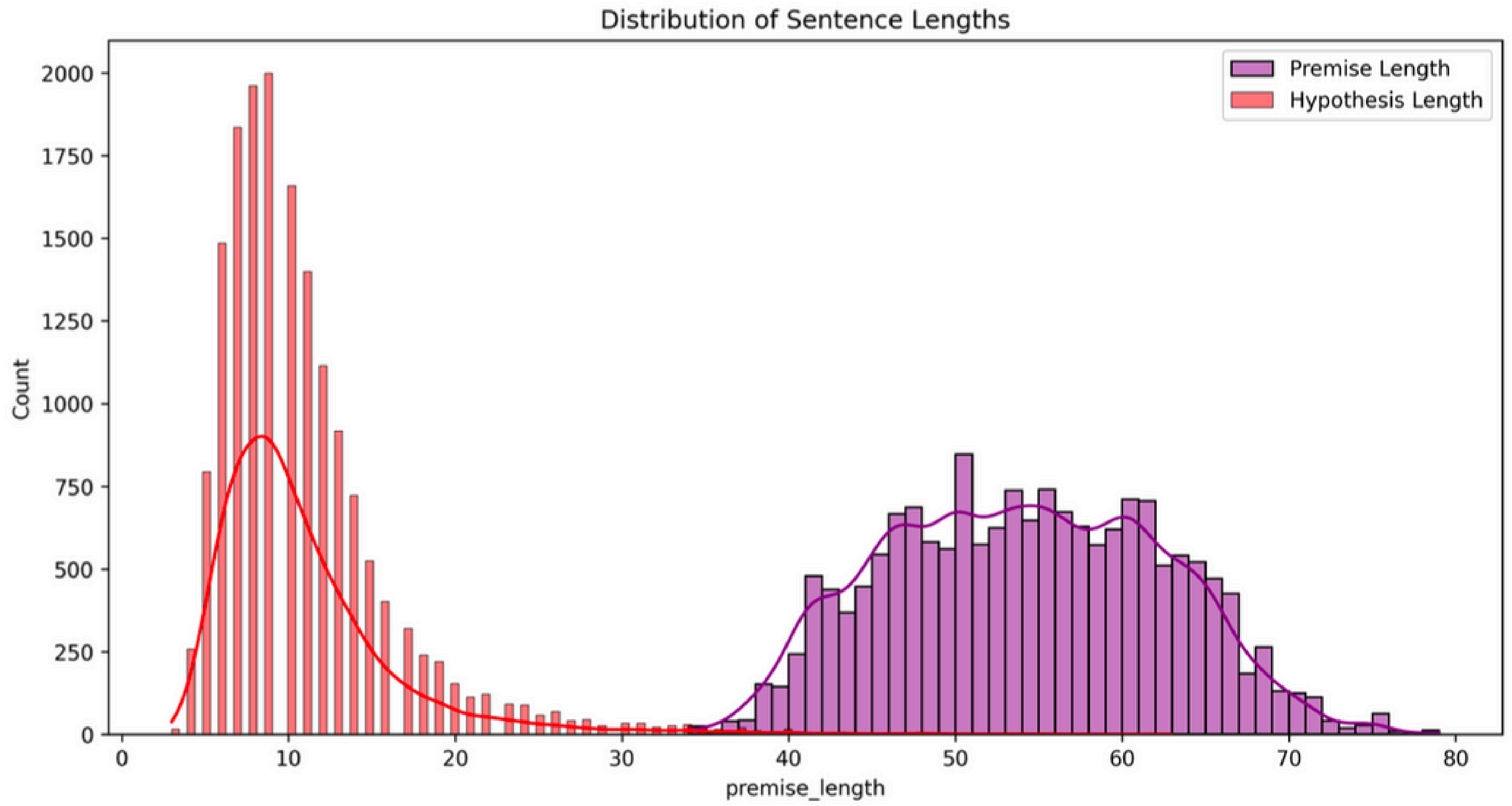
Premise: Two girls are kneeling on the ground
Original Hypothesis: Two girls **stand** around the vending **machines**
Adversarial Hypothesis: Two girls **position** around the vending **machinery**

Exploratory Data Analysis



Label 0 (Entailment): 5371 samples (32%)
Label 1 (Neutral): The dataset has 7052 samples (42%)
Label 2 (Contradiction): 4523 samples (27%)

The dataset is imbalanced, with neutral being the dominant category. Any model trained on this data might need adjustments (e.g., weighted loss) to handle the imbalance effectively.



Premise Length: Shown in purple, with the majority of sentences clustering around 40-50 tokens. A smaller number of premises are shorter (<30 tokens).

Hypothesis Length: Shown in pink, with a distinct peak at shorter lengths (<15 tokens), indicating that hypotheses are generally more concise compared to premises.

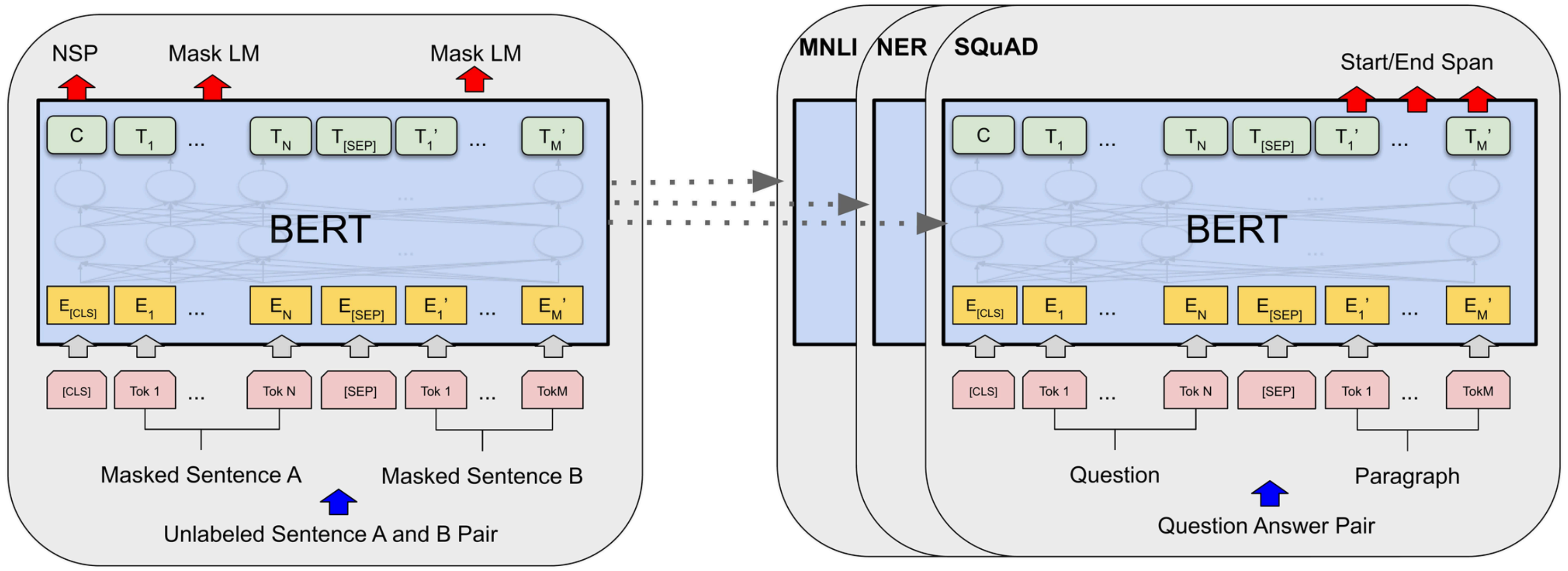
Token truncation or padding will be necessary to standardize variable input lengths.

Testing with Baselines - BERT

BERT (Bidirectional Encoder Representations from Transformers) is a language model that improves natural language processing (NLP) tasks.

We implemented the BERT-Uncased-Base model as the baseline for our experiments. This model has 12 layers, 768 hidden dimensions, and 12 attention heads.

We fine-tuned the BERT-Uncased-Base model on the ANLI dataset for a complete epoch.During this process, the model learned to classify the relationship between pairs of sentences as entailment, contradiction, or neutral across all three rounds of the ANLI dataset.



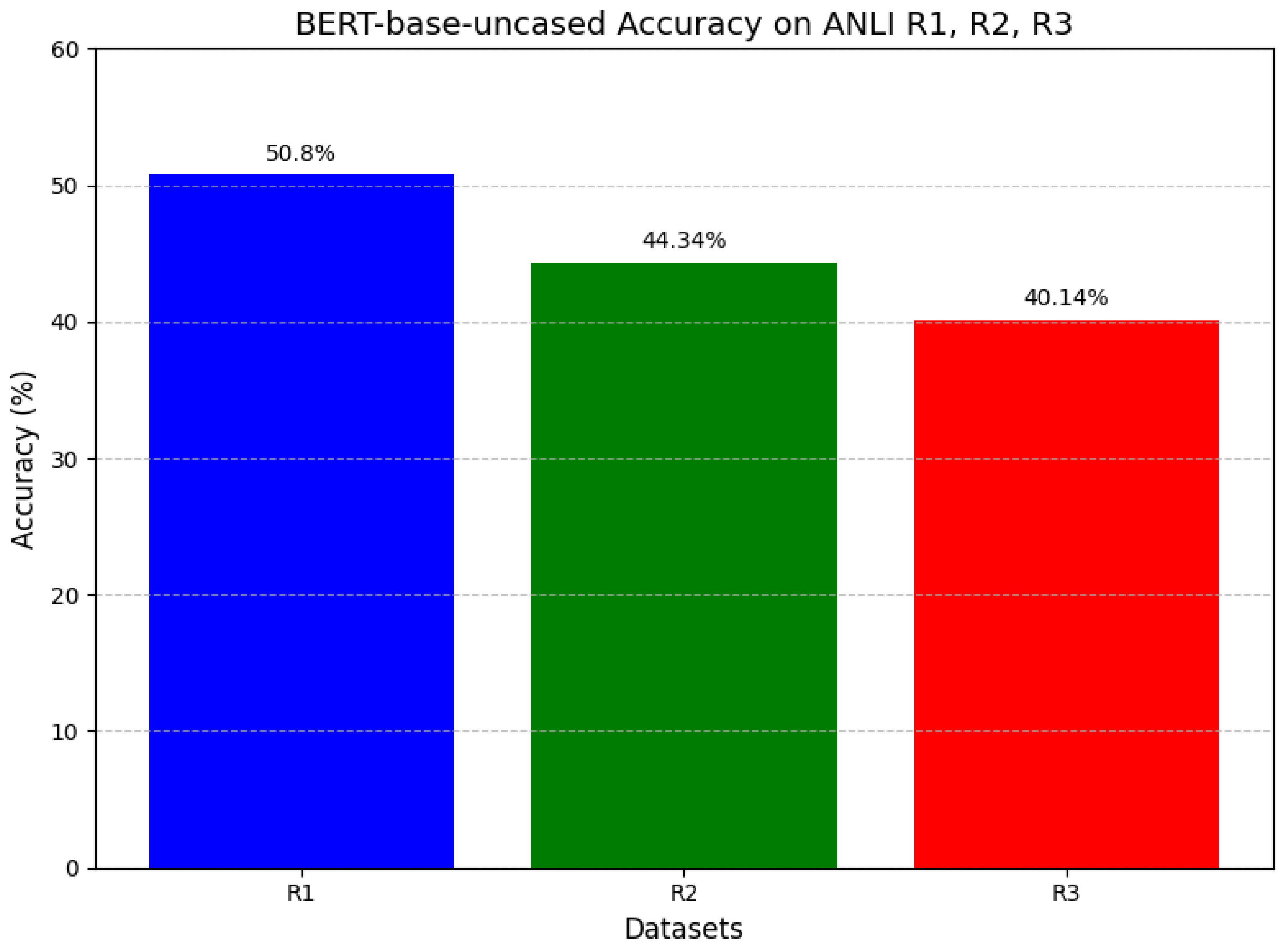
Pre-training

Fine-Tuning

Testing with Baselines - BERT

We used BERT base uncased to analyze the performance on the datasets , particularly R1,R2,R3, following results were obtained after running 1 epoch (65,000 iterations)

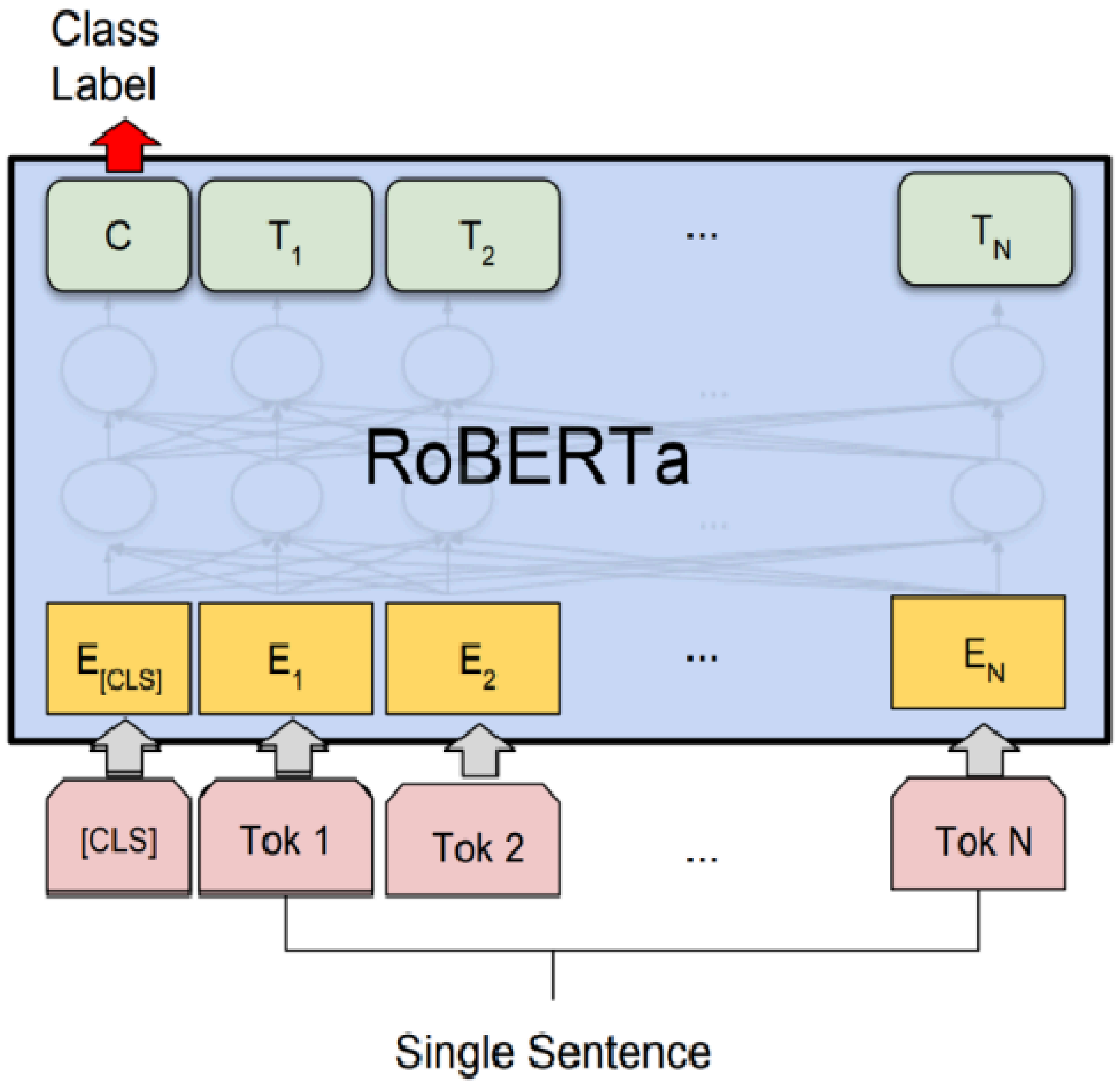
Round	Accuracy
R1	0.508 (50.8%)
R2	0.4434 (44.34%)
R3	0.4014 (40.92%)



Testing with Baselines - RoBERTa

RoBERTa is a transformer-based language model that uses self-attention to process input sequences and generate contextualized word representations.

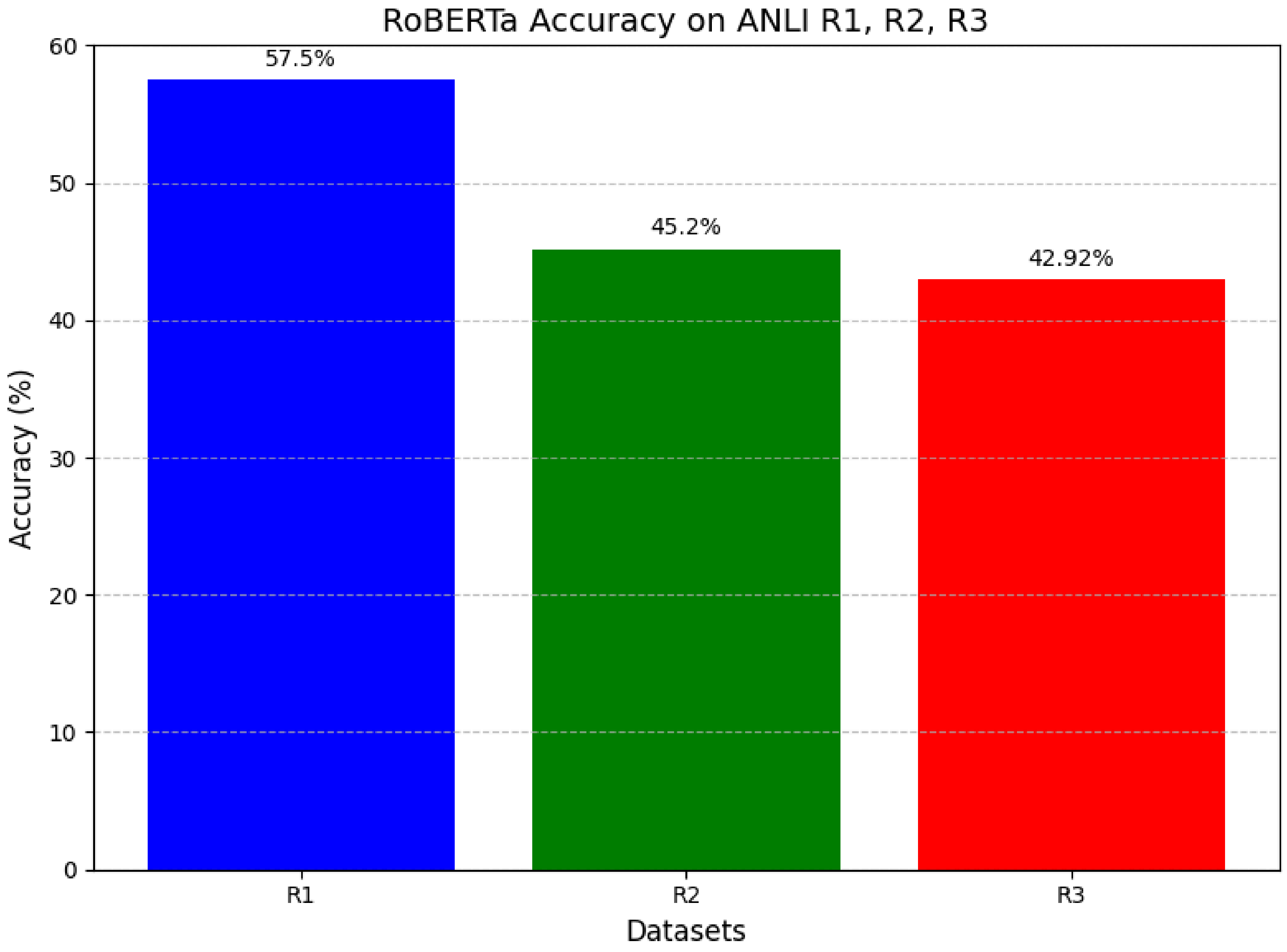
Compared to BERT, RoBERTa removes the Next Sentence Prediction task and optimizes training with larger batches, more data, and dynamic masking, achieving improved performance across NLP tasks.



Testing with Baselines - RoBERTa

We used Roberta base uncased to analyze the performance on the datasets, particularly ANLI R1, R2, R3 as mentioned above. The following results were obtained after running 1 epoch (approximately 65,000iterations):

Round	Accuracy
R1	0.575 (57.5%)
R2	0.452 (45.2%)
R3	0.4292 (42.92%)



Information Bottleneck Regularizer

Goal: Maximize the information relevance of model features for the task while minimizing irrelevant noise.

The mechanism enforces a information thoretic trade-off between compressing input representations and retaining task-relevant information.

The following loss function is optimized:

$$\mathcal{L}_{\mathrm{IB}} = I(Y;T) - \beta I(X;T),$$

- $I(X;T)$: Mutual information between the input X and the representation T (compression).
- $I(Y;T)$: Mutual information between the representation T and the target Y (relevance).
- β: Trade-off parameter controlling the balance between compression and relevance.

Information Bottleneck **Regularizer**

Goal: IB principle formulates the goal of deep learning as an information-theoretic trade off between representation compression and predictive power.

The mechanism enforces a information thoretic trade-off between compressing input representations and retaining task-relevant information.

The following loss function is optimized:

$$\mathcal{L}_{\text{IB}} = I(Y;T) - \beta I(X;T),$$

- $I(X; T)$: Mutual information between the input X and the representation T (compression).
- $I(Y; T)$: Mutual information between the representation T and the target Y (relevance).
- β: Trade-off parameter controlling the balance between compression and relevance.

Anchored Feature Regularizer

Goal:

The goal of the local anchored feature extraction is to find features that carry useful and stable information for downstream tasks.

A feasible plan is to choose the words whose perturbation is neither too large (nonrobust features) nor too small (unuseful features).

Algorithm 1 - Local Anchored Feature Extraction. This algorithm takes in the word local features and returns the index of local anchored features.

1: **Input:** Word local features t, upper and lower threshold c_h and c_l
2: $\delta \leftarrow 0$ *// Initialize the perturbation vector* δ
3: $g(\delta) = \nabla_\delta \ell_{\text{task}}(q_\psi(t+\delta), y)$ *// Perform adversarial attack on the embedding space*
4: Sort the magnitude of the gradient of the perturbation vector from $||g(\delta)_1||_2, ||g(\delta)_2||_2, ..., ||g(\delta)_n||_2$ into $||g(\delta)_{k_1}||_2, ||g(\delta)_{k_2}||_2, ..., ||g(\delta)_{k_n}||_2$ in ascending order, where z_i corresponds to its original index.
5: **Return:** $k_i, k_{i+1}, ..., k_j$, where $c_l \leq \frac{i}{n} \leq \frac{j}{n} \leq c_h$.

Anchored Feature Regularizer

The integrated loss function with both regularizers:

$$\mathcal{L}_{\mathrm{AFE}} = I(Y;T) - n\beta \sum_{i=1}^{n} I(X_i;T_i) + \alpha \sum_{j=1}^{M} I(T_{k_j};Z),$$

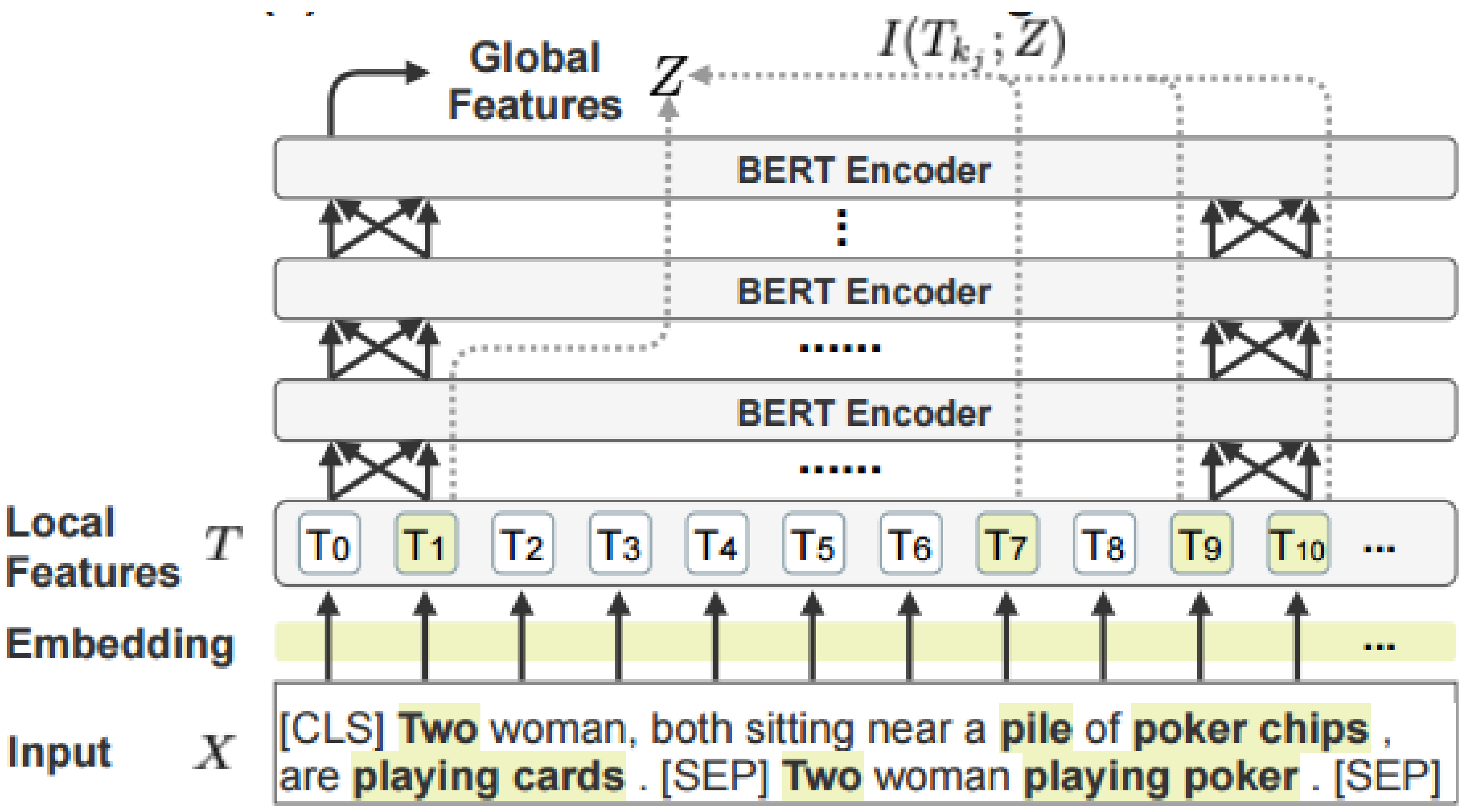
- M is the number of anchored features
- α is a trade-off parameter controlling the importance of features.
- *Tkj* are the local anchored features selected by the algorithm .
- *Z* is the global feature representation.

Premise: **Two** woman, both sitting near a pile of **poker chips**, are **playing cards.**

Hypothesis: **Two** woman **playing poker**.

Here bold words are local stable words for local anchored features

Anchored Feature Regularizer



Local Features (T): The BERT encoder layers process the word embeddings. Each layer creates a representation for each word; T0, T1, T2... T10 represent the word-level features (local features) extracted from the input.

Global Features (Z): This typically refers to the output of a special token ([CLS]) in the BERT architecture which is often treated as a sentence-level representation. The [CLS] embeddings encode information about the entire input sequence and forms a higher-level representation.

Ablation Study

To understand how two regularizers contribute to the improvement of robustness separetely, we apply two regularizers individually to adversarial training.

The following are the three variants we analyzed:

- **INFOBERT**: The full model with all components.
- **INFOBERT without MI**: A model variant excluding the mutual information component.
- **INFOBERT without IB**: A model variant excluding the information bottleneck component.

1. Using only IB Regularizer

Isolating the effect of the IB regularizer. Allows us to assess the individual contribution of the IB regularizer in improving model performance, especially in terms of adversarial robustness and generalization capabilities.

Total (train and inference) time per epoch: 3 hours 42 minutes

Time Analysis

- IB **suppresses noisy mutual information** between the input and the feature representation while learning the **most efficient information features**, or approximate **minimal sufficient statistics**, for downstream tasks.

- Therefore, IB **accelerates convergence**, leading to shorter training times.

2. Using only MI Regularizer

Only retaining the MI regularizer. This setup tests the effect of excluding the IB regularizer, which is designed to mitigate overfitting by constraining the mutual information between
the learned representations and the model's inputs.

Total (train and inference) time per epoch: 7 hours 39 minutes

Time Analysis

- MI imposes an **additional constraint** to minimize irrelevant information and pertubations, and increasing the mutual information between local stable features and global features adding computational overhead.

3. Infobert

Total (train and inference) time per epoch: 6 hours 16 minutes

Why Training Time Is Less Than MI Alone:

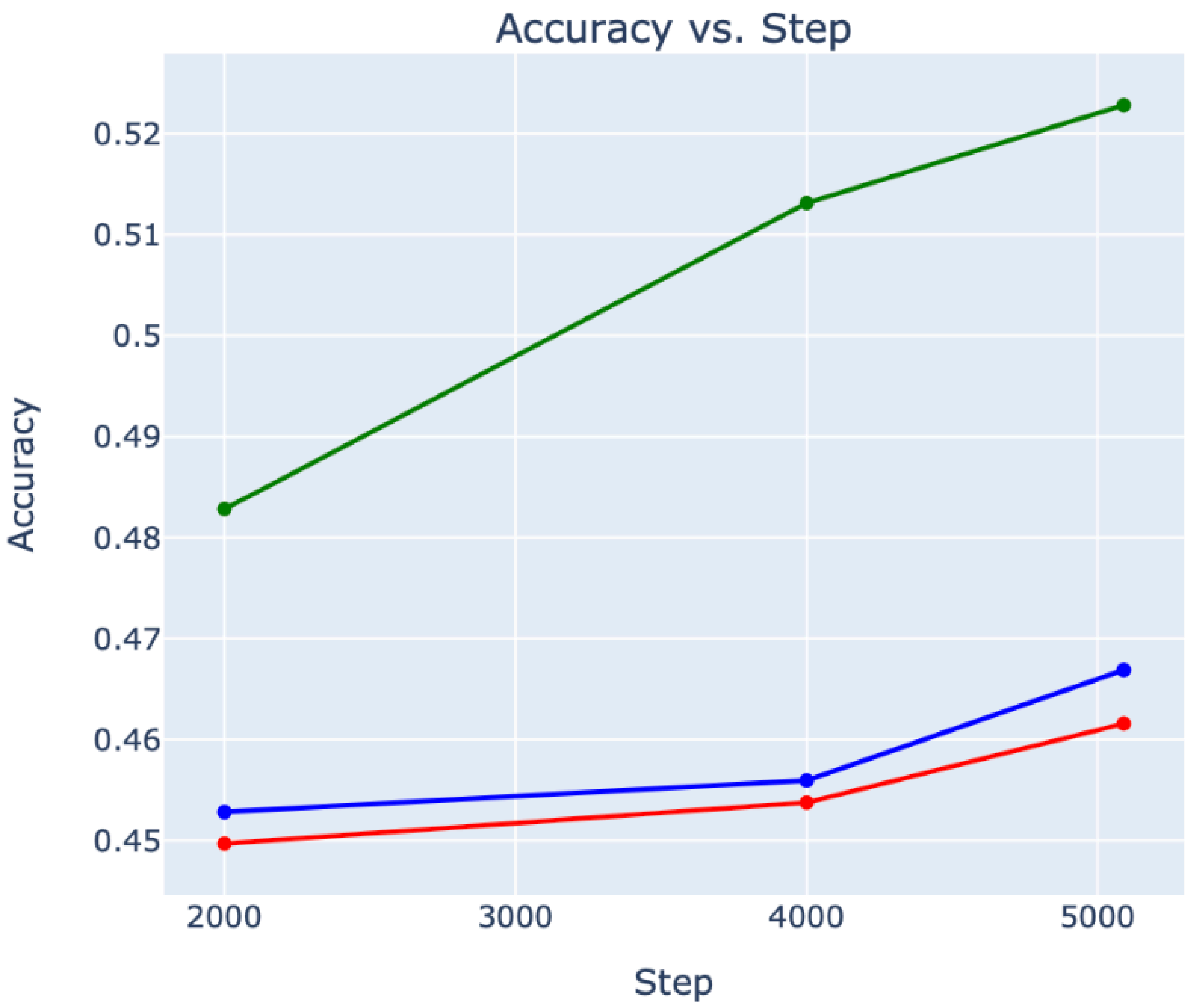
- MI introduces additional computations for focusing on relevant features, slightly increasing the overall workload

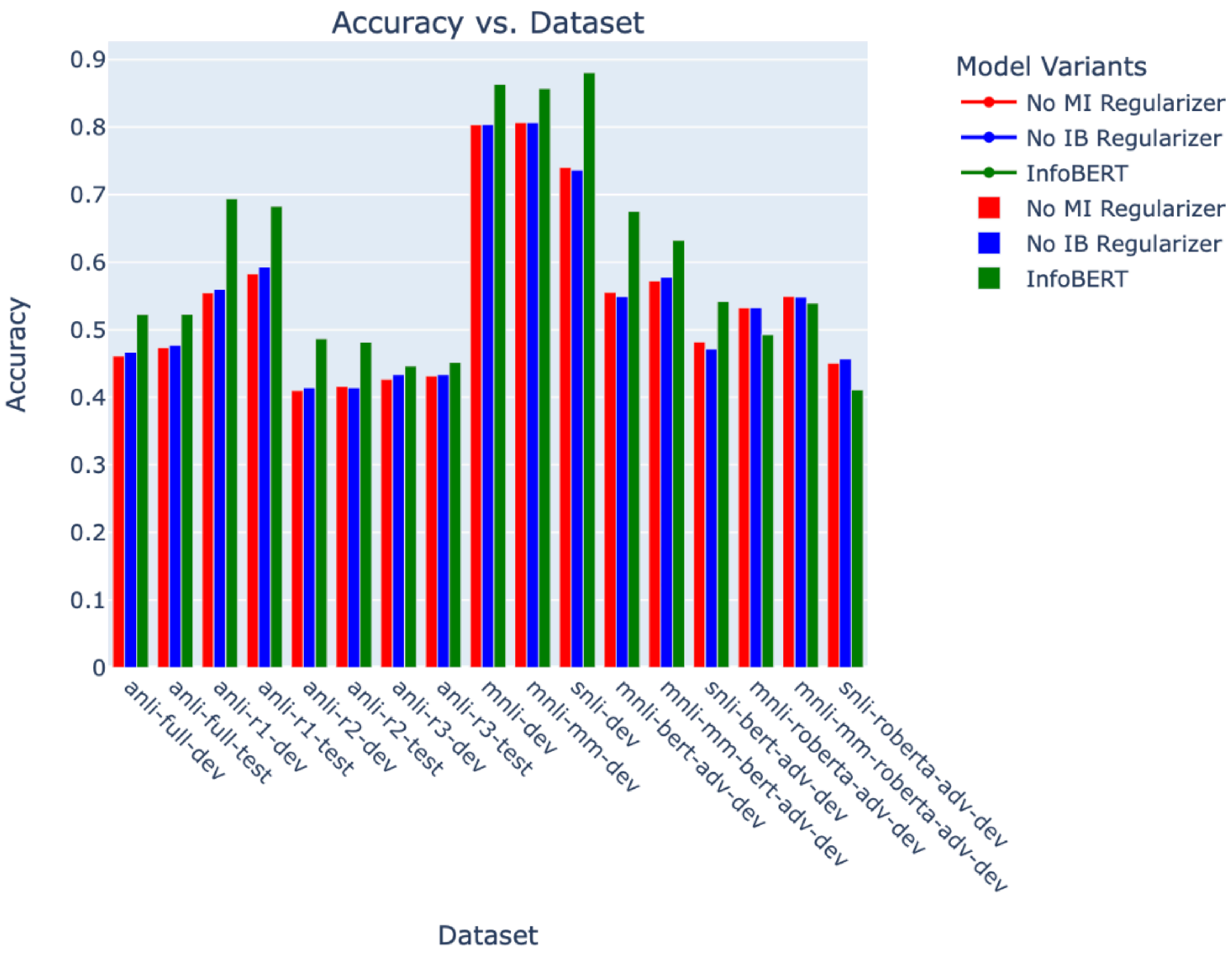
Why Training Time Is More Than IB Alone:

- IB helps **constraint irrelevant information**, leading to more efficient use of computational resources and better optimization.
- However, since MI is also being used, it **offsets** the performance of IB alone, leading to slower training time.

MI and IB hence exhibit a complementary interaction, with IB partially balancing MI's time-intensive nature, highlighting their **non-linear influence** on computational efficiency

Comparison of Variants





Dataset	InfoBERT	No MI Regularizer	No IB Regularizer
ANLI Full Dev	0.5228	0.4616	0.4669
ANLI Full Test	0.5231	0.4741	0.4772
ANLI R1 Dev	0.6940	0.5550	0.5600
ANLI R1 Test	0.6830	0.5830	0.5930
ANLI R2 Dev	0.4870	0.4100	0.4140
ANLI R2 Test	0.4820	0.4160	0.4140
ANLI R3 Dev	0.4467	0.4267	0.4333
ANLI R3 Test	0.4517	0.4317	0.4333
MNLI Dev	0.8635	0.8035	0.8036
MNLI MM Dev	0.8571	0.8071	0.8067
SNLI Dev	0.8805	0.7405	0.7364
MNLI BERT Adv Dev	0.6757	0.5557	0.5492
MNLI MM BERT Adv Dev	0.6324	0.5724	0.5777
SNLI BERT Adv Dev	0.5423	0.4823	0.4717
MNLI RoBERTa Adv Dev	0.4929	0.5329	0.5329
MNLI MM RoBERTa Adv Dev	0.5397	0.5497	0.5484
SNLI RoBERTa Adv Dev	0.4110	0.4510	0.4569

Note: These are R3 accuracies

Bibliography

Alzantot, Moustafa et al. (Oct. 2018). “Generating Natural Language Adversarial Examples”. In:ProceedingsofThe2018ConferenceonEmpiricalMethodsinNaturalLanguageProcessing.Ed.by EllenRiloffetal.Brussels,Belgium:AssociationforComputationalLinguistics,pp.2890–2896. doi:10.18653/v1/D18-1316.url: https://aclanthology.org/D18-1316.

Cohen,JeremyM,ElanRosenfeld,andJ.ZicoKolter(2019).CertifiedAdversarialRobustnessvia RandomizedSmoothing. arXiv: 1902.02918[cs.LG]. url: https://arxiv.org/abs/1902.02918.

Gan, Zhe et al. (2020). Large-Scale Adversarial Training for Vision-and-Language Representation Learning. arXiv: 2006.06195[cs.CV].url:https://arxiv.org/abs/2006.06195.

Jia, Robin and Percy Liang (Sept. 2017). “Adversarial Examples for Evaluating Reading Comprehension Systems”.In: Proceedings ofthe 2017 Conference onEmpirical Methodsin Natural Language Processing. Ed. by MarthaPalmer, Rebecca Hwa, and Sebastian Riedel. Copenhagen, Denmark: Association for Computational Linguistics, pp. 2021–2031. doi: 10.18653/v1/D17-1215. url: https://aclanthology.org/D17-1215.

Jin,Dietal.(2020).IsBERTReallyRobust?AStrongBaselineforNaturalLanguageAttackonText Classification and Entailment. arXiv: 1907.11932 [cs.CL]. url: https://arxiv.org/abs/1907.11932.

THANK YOU!