Improving Natural Language Understanding via Contrastive Learning Methods

Pengyu Cheng

Department of Electrical and Computer Engineering
Duke University

March 30, 2021

Overview

1. Background
    - Natural Language Understanding
    - Contrastive Learning
2. Ph.D. Research Outline
3. Improving Representation Efficiency
    - Binarized Text Representation
4. Improving Representation Interpretability
    - Disentangled Text Representation
5. Improve Representation Fairness
    - Debiased Text Representation

Background

Background: Natural Language Understanding

Natural Language Processing (NLP):
A branch of artificial intelligence (AI) dealing with interaction between humans and machines via natural language.

Natural Language Understanding (NLU):
An sub-topic of NLP, extracting and understanding the semantic information from raw-text or speech data.

NLU is essential for many NLP applications,
e.g. Sentiment Analysis, Machine Translation, Question Answering.

Background: Text Representation Learning

Earlier work for NLU: **rule-based** or **word-level** methods.
e.g. syntax analysis, regular expression, and bag-of-word models.

Recently, **representation learning** becomes the mainstream for NLU:
Given each sentence as a sequence of word $\boldsymbol{x} = (w_1, \dots, w_L) \in \mathcal{X}$
Learn encoder function $f(\cdot) : \mathcal{X} \to \mathbb{R}^d$
Obtain real-valued representation vector $f(\boldsymbol{x})$, *i.e.*, embedding

With the development of neural networks, **deep** text encoders have achieved significant empirical improvement.
e.g., InferSent [Conneau et al., 2017b], BERT [Devlin et al., 2019], RoBERTa [Liu et al., 2019].

Background: Weakness of Text Representation Learning

Although deep text encoders have shown remarkable NLU performance, many important encoder properties are in lack of exploration:

- **Efficiency**: High-dimensional real-valued embeddings have huge cost of computation and storage, especially for low calculation-ability resources, *e.g.* mobile devices.
- **Interpretability**: How to interpret the learned embeddings? What does each element in the embedding vectors mean? How are they related to original sentences?
- **Fairness**: Data-driven NLU models suffer from the **social bias** problem, which is intrinsically from data, *e.g.*, gender bias in reviews. How to measure and eliminate bias from embeddings?

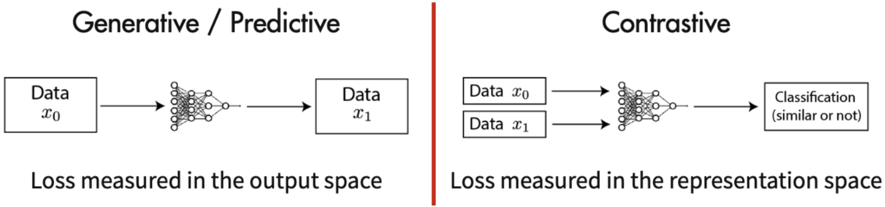
Background: Contrastive Learning

Contrastive learning: a broad class of machine learning strategies, which learn models by enlarging difference between **positive** and **negative** sample pairs.

General formula:

$$\text{score}(f(\boldsymbol{x}), f(\boldsymbol{x}^+)) >> \text{score}(f(\boldsymbol{x}), f(\boldsymbol{x}^-)) \tag{1}$$

Positive:$(\boldsymbol{x}, \boldsymbol{x}^+)$ similar data points; **Negative**:$(\boldsymbol{x}, \boldsymbol{x}^-)$ dissimilar points.



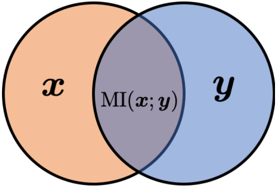
Generative / Predictive
Data x_0
Data x_1
Loss measured in the output space
Contrastive
Data x_0
Data x_1
Classification (similar or not)
Loss measured in the representation space

Background: Mutual Information & Contrastive Learning

Mutual Information (MI) measures the correlation between variables:

$$\mathcal{I}(\boldsymbol{x};\boldsymbol{y}) = \mathbb{E}_{p(\boldsymbol{x};\boldsymbol{y})}[\log \frac{p(\boldsymbol{x},\boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})}]. \tag{2}$$

MI has various applications in machine learning,
but is challenging to estimate when only sample pairs $\{(\boldsymbol{x}_i,\boldsymbol{y}_i)\}_{i=1}^N$ are provided.



Background: Mutual Information & Contrastive Learning

Several contrastive MI estimators are proposed based on positive pair $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ and negative pair $(\boldsymbol{x}_i, \boldsymbol{y}_j)$.

Noise Contrastive Estimation [Oord et al., 2018] (InfoNCE):

$$\mathcal{I}_{\text{NCE}} := \max_f \frac{1}{N} \sum_{i=1}^{N} \left[f(\boldsymbol{x}_i, \boldsymbol{y}_i) - \log\left(\frac{1}{N} \sum_{j=1}^{N} e^{f(\boldsymbol{x}_i, \boldsymbol{y}_j)}\right) \right], \tag{3}$$

Mutual Information Neural Estimation [Belghazi et al., 2018] (MINE):

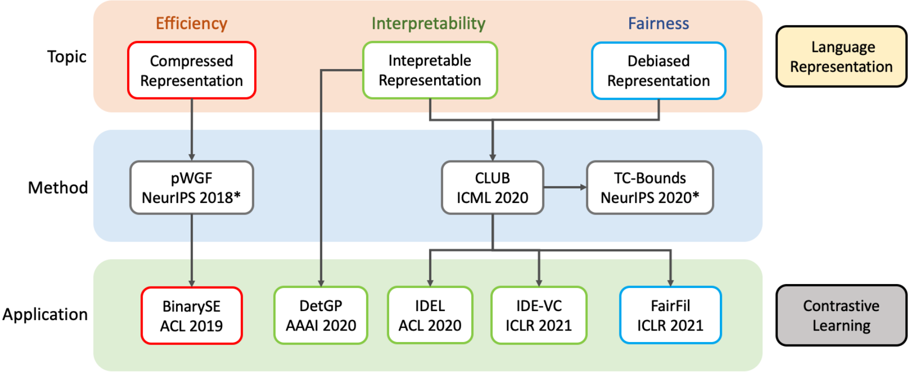
$$\mathcal{I}_{\text{MINE}} := \max_f \left(\frac{1}{N} \sum_{i=1}^{N} f(\boldsymbol{x}_i, \boldsymbol{y}_i) \right) - \log \left(\frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} e^{f(\boldsymbol{x}_i, \boldsymbol{y}_j)} \right), \tag{4}$$

Most previous MI estimators focus on lower-bound estimation.

Ph.D. Research Outline

Research Outline

Improving natural language understanding with contrastive learning.



* Workshop

Improving Representation Efficiency

Learning Compressed Text Representation (ACL 2019)

Sentence encoders output high-dimensional real-valued vectors:
high complexity for computation and storage.

We aim to learn compact and binarized representations from continuous sentence embeddings, and preserve the semantic information.

Binarized embedding: easy for binary storage systems; fast bit operation.

Given a pretrained encoder $f(\cdot) : \mathcal{X} \to \mathbb{R}^d$.
$\boldsymbol{h} = f(\boldsymbol{x})$ is the continuous embedding extracted from sentence $\boldsymbol{x}$.

We plan to learn a transformation $g(\cdot)$ that converts $f(\boldsymbol{x})$ to highly informative binary embedding. *i.e.*, $\boldsymbol{b} = g(\boldsymbol{h}) = g(f(\boldsymbol{x}))$.

Learning Compressed Text Representation

To learn transformation $g(\cdot)$, we consider an auto-encoder architecture.

The transformation is parameterized by

$$\boldsymbol{b}' = \sigma(\boldsymbol{W}\boldsymbol{h} + \boldsymbol{k}), \boldsymbol{b} = \frac{\text{Sign}(\boldsymbol{b}' - \boldsymbol{s}) + 1}{2}, \quad (5)$$

where $\boldsymbol{W}$ and $\boldsymbol{k}$ are learning weights, Sign$(\cdot)$ is the sign function, $\boldsymbol{s}$ is the threshold.

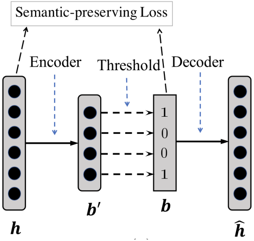


Figure: The encoder-decoder binarization framework.

Learning Compressed Text Representation

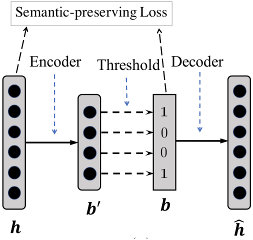
To ensure $\boldsymbol{b}$ including sufficient information from $\boldsymbol{h}$, we use a decoder to reconstruct $\boldsymbol{h}$:

$$\boldsymbol{h}' = \boldsymbol{W}'\boldsymbol{b} + \boldsymbol{k}', \tag{6}$$

where $\boldsymbol{W}'$ and $\boldsymbol{k}'$ are weights of the decoder.

The reconstruction loss $\mathcal{L}_{rec} = \|\boldsymbol{h}' - \boldsymbol{h}\|^2$.

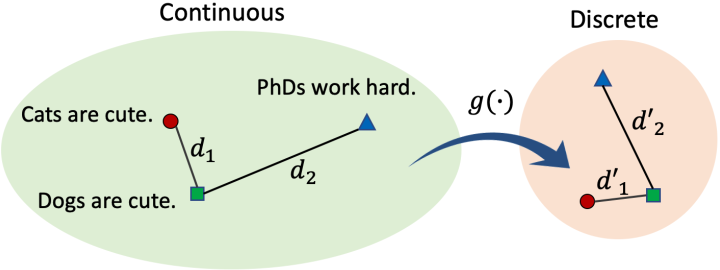
Straight-through (ST) estimator [Hinton, 2012] is utilized to back-propagate gradients gradient though discrete variables.



Contrastive Semantic-preserving Regularizer

Only the auto-encoder framework can not guarantee to preserve relative similarity information.

i.e. If two sentences have higher similarity in the continuous embedding space, they should also have higher similarity in the binary space.

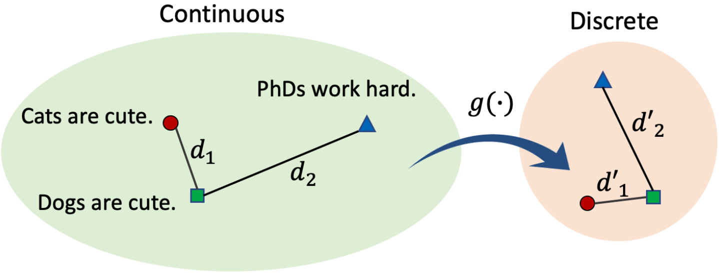


Contrastive Semantic-preserving Regularizer

Consider a triple group of sentences $(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3)$
continuous embeddings: $(\boldsymbol{h}_1, \boldsymbol{h}_2, \boldsymbol{h}_3)$; binarized embeddings:$(\boldsymbol{b}_1, \boldsymbol{b}_2, \boldsymbol{b}_3)$.

contrastive semantic-preserving regularizer:

$$\mathcal{L}_{sp} = \mathrm{ReLU}\left(\mathbf{1}_{d(\boldsymbol{h}_1,\boldsymbol{h}_3)>d(\boldsymbol{h}_1,\boldsymbol{h}_2)}\left(d'(\boldsymbol{b}_1, \boldsymbol{b}_2) - d'(\boldsymbol{b}_1, \boldsymbol{b}_3)\right)\right).$$



Experiments

Continuous embeddings:
InferSent [Conneau et al., 2017a] outputs, dimension 4096.

Downstream task performance:
Our binarized embedding achieves competitive results
only ∼ 2% performance loss; reducing ∼ 98% storage.

With the semantic-preserving regularizer (AE-binary-SP), the performance of binarized embedding further improved.

Model	Dim	MR	CR	SUBJ	MPQA	SST	STS14	STSB	SICK-R	MRPC
Continuous *(dense) sentence embeddings*										
fastText-BoV	300	78.2	80.2	91.8	88.0	82.3	.65/.63	58.1/59.0	0.698	67.9/74.3
SkipThought	4800	76.5	80.1	93.6	87.1	82.0	.29/.35	41.0/41.7	0.595	57.9/66.6
SkipThought-LN	4800	79.4	83.1	**93.7**	89.3	82.9	.44/.45	-	-	-
InferSent-FF	4096	79.7	84.2	92.7	89.4	84.3	.68/.66	55.6/56.2	0.612	**67.9/73.8**
InferSent-G	4096	**81.1**	**86.3**	92.4	**90.2**	**84.6**	**.68/.65**	**70.0/68.0**	**0.719**	67.4/73.2
Binary *(compact) sentence embeddings*										
AE-binary	2048	78.7	**84.9**	90.6	89.6	82.1	.68/.66	71.7/69.7	0.673	65.8/70.8
AE-binary-SP	2048	**79.1**	84.6	**90.8**	**90.0**	**82.7**	**.69/.67**	**73.2/70.6**	**0.705**	**67.2**/72.0

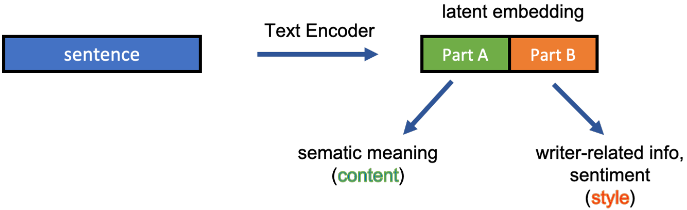
Improving Representation Interpretability

Learning Disentangled Text Representation (ACL 2020)

Disentangled representation learning is an important approach to improve the interpretability of embeddings.

Specifically, disentangled representation maps different data attributes into different latent embedding parts.

We aim to disentangle the **style** and **content** information of sentences.

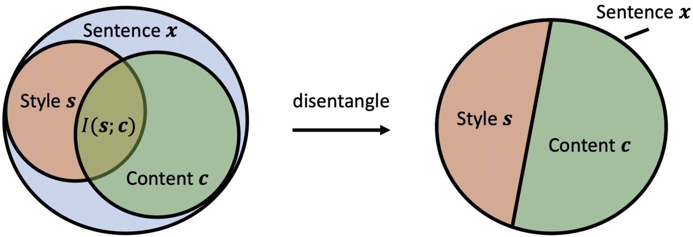


Learning Interpretable Text Representation

Both style and content embeddings should:
(1) be representative; (2) not reveal information from each other.

An information-theoretic perspective of disentangling:

Intuitively, we could: $\min \mathcal{I}(\boldsymbol{s};\boldsymbol{c}) - \mathcal{I}(\boldsymbol{s};\boldsymbol{x}) - \mathcal{I}(\boldsymbol{c};\boldsymbol{x})$.



Mutual Information Estimation

Data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$: sentence $\boldsymbol{x}_i$, style label y_i .
Corresponding style embedding $\boldsymbol{s}_i$; content embedding $\boldsymbol{c}_i$.

To maximize $\mathcal{I}(\boldsymbol{x};\boldsymbol{c})$ and $\mathcal{I}(\boldsymbol{x};\boldsymbol{s})$, introduce $q_\phi(\boldsymbol{x}|\boldsymbol{c})$ and $q_\psi(y|\boldsymbol{s})$ a variational lower bound from Chen et al. [2016]:

$$\mathcal{I}(\boldsymbol{x};\boldsymbol{c}) \geq \mathcal{H}(\boldsymbol{x}) + \mathbb{E}_{p(\boldsymbol{x};\boldsymbol{c})}[\log q_\phi(\boldsymbol{x}|\boldsymbol{c})].$$

$$\mathcal{I}(\boldsymbol{x};\boldsymbol{s}) \geq \mathcal{I}(y;\boldsymbol{s}) \geq \mathcal{H}(y) + \mathbb{E}_{p(y,\boldsymbol{s})}[\log q_\psi(y|\boldsymbol{s})]$$

Both entropy terms $\mathcal{H}(\boldsymbol{x})$ and $\mathcal{H}(y)$ are constants from the data.
Only need to minimize:

$$\bar{\mathcal{L}}_{\text{Dis}} = \mathcal{I}(\boldsymbol{s};\boldsymbol{c}) - \frac{1}{N}\sum_{i=1}^{N} \log q_\phi(\boldsymbol{x}_i|\boldsymbol{c}_i) - \frac{1}{N}\sum_{i=1}^{N} \log q_\psi(y_i|\boldsymbol{s}_i). \tag{7}$$

MI Sample-based Contrastive Upper Bound

To estimate $\mathcal{I}(\boldsymbol{s}; \boldsymbol{c})$, we propose a novel sample-based upper bound.

Theorem (Contrastive Log-ratio Upper Bound (CLUB))

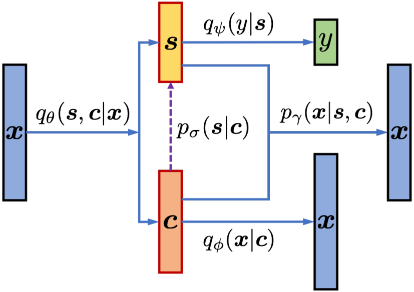
If $\{(\boldsymbol{s}_j, \boldsymbol{c}_j)\}_{j=1}^N \sim p(\boldsymbol{s}, \boldsymbol{c})$, *then*

$$\mathcal{I}(\boldsymbol{s}; \boldsymbol{c}) \leq \mathbb{E}\Big[\frac{1}{N}\sum_{j=1}^{N}\Big[\log p(\boldsymbol{s}_j|\boldsymbol{c}_j) - \frac{1}{N}\sum_{k=1}^{N}\log p(\boldsymbol{s}_j|\boldsymbol{c}_k)\Big]\Big] =: \mathcal{I}_{CLUB}(\boldsymbol{s}; \boldsymbol{c}).$$

The calculation of $\mathcal{I}_{\text{CLUB}}$ requires the conditional distribution $p(\boldsymbol{s}|\boldsymbol{c})$.

We use a variational network $p_\sigma(\boldsymbol{s}|\boldsymbol{c})$ to approximate $p(\boldsymbol{s}|\boldsymbol{c})$ by maximizing log-likelihood $\mathcal{L}(\sigma) = \frac{1}{N}\sum_{j=1}^{N}\log p_\sigma(\boldsymbol{s}_j|\boldsymbol{c}_j)$.

Framework


$\boldsymbol{x}$
$q_\theta(\boldsymbol{s}, \boldsymbol{c}|\boldsymbol{x})$
$\boldsymbol{s}$
$q_\psi(y|\boldsymbol{s})$
y
$p_\sigma(\boldsymbol{s}|\boldsymbol{c})$
$p_\gamma(\boldsymbol{x}|\boldsymbol{s}, \boldsymbol{c})$
$\boldsymbol{x}$
$\boldsymbol{c}$
$\boldsymbol{x}$
$q_\phi(\boldsymbol{x}|\boldsymbol{c})$

The style embedding $\boldsymbol{s}$ goes through a classifier $q_\psi(y|\boldsymbol{s})$ to predict the style label y; the content embedding $\boldsymbol{c}$ is used to reconstruct $\boldsymbol{x}$.

$p_\sigma(\boldsymbol{s}|\boldsymbol{c})$ helps disentangle the style and content embeddings.

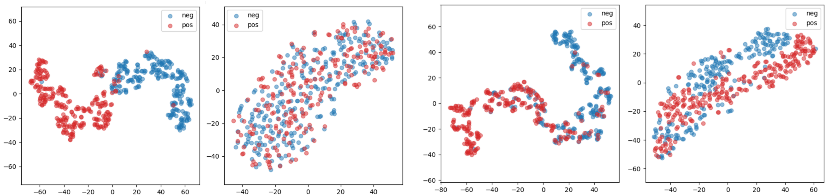
The decoder $p_\gamma(\boldsymbol{x}|\boldsymbol{s}, \boldsymbol{c})$ generates sentences based on $\boldsymbol{s}$ and $\boldsymbol{c}$.

Experiments: Embedding Disentanglement Quality

We call this final framework Information-theoretic Disentangled text Embedding Learning (IDEL).

We first analyze the disentanglement of learned embeddings on Yelp review dataset (including positive and negative reviews).

Ablation Study: IDEL$^-$ without minimizing $\mathcal{I}(\boldsymbol{s}; \boldsymbol{c})$.



(a) Latent spaces t-SNE plots of IDEL on Yelp.

(b) t-SNE plots of IDEL$^-$ without $\hat{\mathcal{I}}(\boldsymbol{s}; \boldsymbol{c})$.

Experiments: Embedding Representation Quality

To show the representation ability of IDEL, we conduct experiments on two text-generation tasks: style transfer and conditional generation.

We test the following metrics: (1) **Style Preservation:** pre-train a style classifier and use it to test whether a generated sentence can be categorized into the correct target style class.

(2) **Content Preservation:** The self-BLEU score [Lample et al., 2019] is calculated between one original sentence and its style-transferred sentence.

(3) **Generation Quality:** calculate the corpus-level BLEU score [Papineni et al., 2002] between a generated sentence and the testing data corpus.

(4) **Geometric Mean:** use the geometric mean (GM) of the above metrics as an overall evaluation.

Experiments: Embedding Representation Quality

	Yelp Dataset							**Personality Captioning Dataset**						
	Conditional Generation			Style Transfer				Conditional Generation			Style Transfer			
	ACC	**BLEU**	**GM**	**ACC**	**BLEU**	**S-BLEU**	**GM**	**ACC**	**BLEU**	**GM**	**ACC**	**BLEU**	**S-BLEU**	**GM**
CtrlGen	82.5	20.8	41.4	83.4	19.4	31.4	37.0	73.6	18.9	37.0	73.3	18.9	30.0	34.6
CAAE	78.9	19.7	39.4	79.3	18.5	28.2	34.6	72.2	19.5	37.5	72.1	18.3	27.4	33.1
ARAE	78.3	**23.1**	42.4	78.5	21.3	32.5	37.9	72.8	**22.5**	40.4	71.5	20.4	31.6	35.8
BT	81.4	20.2	40.5	**86.3**	24.1	**35.6**	**41.9**	74.1	21.0	39.4	**75.9**	23.1	34.2	39.1
DRLST	83.7	22.8	43.7	85.0	23.9	34.9	41.4	74.9	22.0	40.5	75.7	21.9	33.8	38.3
IDEL^{-}	78.1	20.3	39.8	79.1	20.1	27.5	35.1	72.0	19.7	37.7	72.4	19.7	27.1	33.8
IDEL	**83.9**	23.0	**43.9**	85.7	**24.3**	35.2	**41.9**	**75.1**	22.3	**40.9**	75.6	**23.3**	**34.6**	**39.4**

Our IDEL learns more representative and balanced representations in trade-off between style and content preservation.

Comparison between IDEL and IDEL^{-} support the effectiveness of proposed MI upper bound.

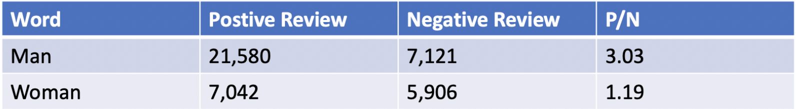
Improve Representation Fairness

Social Bias in Text Representation

Data-driven models suffer from the bias of the training data.

Many human-language datasets contain **social bias** in terms of gender, religion, race, *etc.*

Example: Yelp review dataset:



Word	Postive Review	Negative Review	P/N
Man	21,580	7,121	3.03
Woman	7,042	5,906	1.19

A clear gap between the numbers of positive reviews to word ”man” and word ”woman”.

May et al. [2019] proposed an embedding fairness measurement and pointed the existence of social bias in pretrained sentence encoders.

Learning Fair Text Representation (ICLR2021)

Previous works mainly focus on word-level debiasing; Lack exploration to sentence-level debiasing

I developed the first neural debiasing method for pretrained text encoders.

Suppose $E(\cdot)$ is a pretrained sentence encoder:
encodes sentence $\boldsymbol{x}$ into embedding $\boldsymbol{z} = E(\boldsymbol{x})$.

Aim to learn a fair filter network $f(\cdot)$ on the top of $E(\cdot)$,
such that the output embedding $\boldsymbol{d} = f(\boldsymbol{z})$ can be debiased.

Learning Fair Text Representation

Our method includes three parts:

- Augmentation for biased data
- Contrastive Learning Framework
- Debiasing Regularizer

Sentence Augmentation for Biased Datasets

-**Social sensitive topic**: $\mathcal{T} = \{\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_K\}$,
$\mathcal{D}_k$ is a **potential bias direction** under the topic.

Example: $\mathcal{T} =$ *"gender"* , $\{\mathcal{D}_1, \mathcal{D}_2\} = \{$ *"male"*, *"female"*$\}$.

-**Sensitive attribute word**: a word $\boldsymbol{w}$ related to bias direction $\mathcal{D}_k$

Example: "he" $\in$ *"male"*; "she" $\in$ *"female"*.

For each sensitive word $\boldsymbol{w} \in \mathcal{D}_k$ in $\boldsymbol{x}$,
replace it with $\boldsymbol{w}' \in \mathcal{D}_j$ in another bias direction.

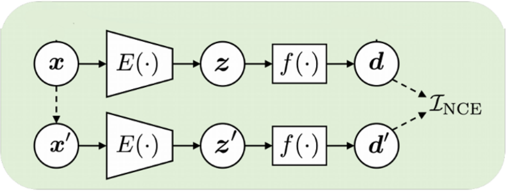
	Bias direction	Sensitive Attribute words	Text content
Original	male	he, his	{He} is good at playing {his} basketball.
Augmentation	female	she, her	{She} is good at playing {her} basketball.

Contrastive Learning Framework

For each sentence $\boldsymbol{x}$, generate an augmented sentence $\boldsymbol{x}'$:
the same semantic meaning; different potential bias direction.

The debiased embedding $(\boldsymbol{d}, \boldsymbol{d}')$ should share same semantic information.

- (1) encode $(\boldsymbol{z}, \boldsymbol{z}') = (E(\boldsymbol{x}), E(\boldsymbol{x}'))$ with pretrained encoder $E(\cdot)$;
- (2) obtain debiased embedding $(\boldsymbol{d}, \boldsymbol{d}') = (f(\boldsymbol{z}), f(\boldsymbol{z}'))$ via filter $f(\cdot)$;
- (3) maximize the mutual information $\mathcal{I}(\boldsymbol{d}; \boldsymbol{d}')$.



Contrastive Learning Framework

Given sentence pairs $\{(\boldsymbol{x}_i, \boldsymbol{x}_i')\}_{i=1}^N$, obtain $(\boldsymbol{d}_i, \boldsymbol{d}_i') = (f(E(\boldsymbol{x}_i)), f(E(\boldsymbol{x}_i')))$.

We use the InfoNCE [Oord et al., 2018] mutual information estimator

$$\mathcal{I}_{\text{NCE}} = \frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(g(\boldsymbol{d}_i, \boldsymbol{d}_i'))}{\frac{1}{N} \sum_{j=1}^{N} \exp(g(\boldsymbol{d}_i, \boldsymbol{d}_j'))}. \tag{8}$$

By maximizing $\mathcal{I}(\boldsymbol{d}; \boldsymbol{d}')$ we encourage $\boldsymbol{d}$ sharing more semantic information with $\boldsymbol{d}'$.

Debiasing Regularizer

Assumption: potential bias comes from the sensitive attribute words in $\boldsymbol{x}$.

Eliminate bias from $\boldsymbol{d}$ by reducing $\mathcal{I}(\boldsymbol{d}; \boldsymbol{w}^p)$.

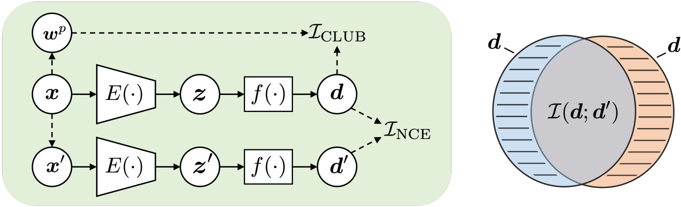
Let $\boldsymbol{w}^p$ be the embedding of a sensitive attribute word w^p in sentence $\boldsymbol{x}$.
Embedding $\boldsymbol{w}^p$ can be obtained from pretrained text encoders.

Given a batch of embedding pairs $\{(\boldsymbol{d}_i, \boldsymbol{w}^p)\}_{i=1}^N$, debiasing regularizer is:

$$\mathcal{I}_{\text{CLUB}} = \frac{1}{N} \sum_{i=1}^{N} \Big[\log q_\theta(\boldsymbol{w}_i^p | \boldsymbol{d}_i) - \frac{1}{N} \sum_{j=1}^{N} \log q_\theta(\boldsymbol{w}_j^p | \boldsymbol{d}_i) \Big], \tag{9}$$

where q_θ is a variational approximation $p(\boldsymbol{w}|\boldsymbol{d})$.

Overall Framework



(a)

(b)

By maximizing $\mathcal{I}_{\text{NCE}}$ term, we enlarge the overlapped area of $\boldsymbol{d}$ and $\boldsymbol{d}'$;
By minimizing $\mathcal{I}_{\text{CLUB}}$ term, we shrink the biased shadow parts.

Overall training objective: $\mathcal{I}_{\text{NCE}} + \beta \mathcal{I}_{\text{CLUB}}$.

Experiments: Evaluation Metric

Sentence Embedding Association Test (SEAT) [May et al., 2019]:

- Bias degree of embedding $\boldsymbol{t}$ to two attributes $\mathcal{A}$ and $\mathcal{B}$:

$$s(\boldsymbol{t}, \mathcal{A}, \mathcal{B}) = \text{mean}_{\boldsymbol{a} \in \mathcal{A}} \cos(\boldsymbol{t}, \boldsymbol{a}) - \text{mean}_{\boldsymbol{b} \in \mathcal{B}} \cos(\boldsymbol{t}, \boldsymbol{b}),$$

- Overall regularized bias degree of two sets $\mathcal{X}, \mathcal{Y}$ to two attributes $\mathcal{A}, \mathcal{B}$:

$$d_{\text{WEAT}} = \frac{\text{mean}_{\boldsymbol{x} \in \mathcal{X}} s(\boldsymbol{x}, \mathcal{A}, \mathcal{B}) - \text{mean}_{\boldsymbol{y} \in \mathcal{Y}} s(\boldsymbol{y}, \mathcal{A}, \mathcal{B})}{\text{std}_{\boldsymbol{t} \in \mathcal{X} \cup \mathcal{Y}} s(\boldsymbol{t}, \mathcal{A}, \mathcal{B})}. \qquad (10)$$

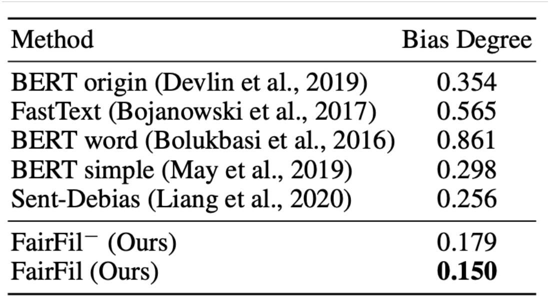
- Each embedding in SEAT is encoded from a pre-designed sentence template, *e.g.*, “this is <word>.”

Experiments Setups

- Pretrained Encoders: (1) Pretrained BERT; (2) BERT post tasks.
- Downstream Tasks: SST-2; CoLA; QNLI.
- Ablation Study:
  FairFil: the whole model;
  FairFil^{-}: without the debiasing regularizer;

Experimental Results: Pretrained BERT

Debiasing performance on Pretrained BERT:



Method	Bias Degree
BERT origin (Devlin et al., 2019)	0.354
FastText (Bojanowski et al., 2017)	0.565
BERT word (Bolukbasi et al., 2016)	0.861
BERT simple (May et al., 2019)	0.298
Sent-Debias (Liang et al., 2020)	0.256
FairFil^{-} (Ours)	0.179
FairFil (Ours)	**0.150**

Experimental Results: BERT post tasks

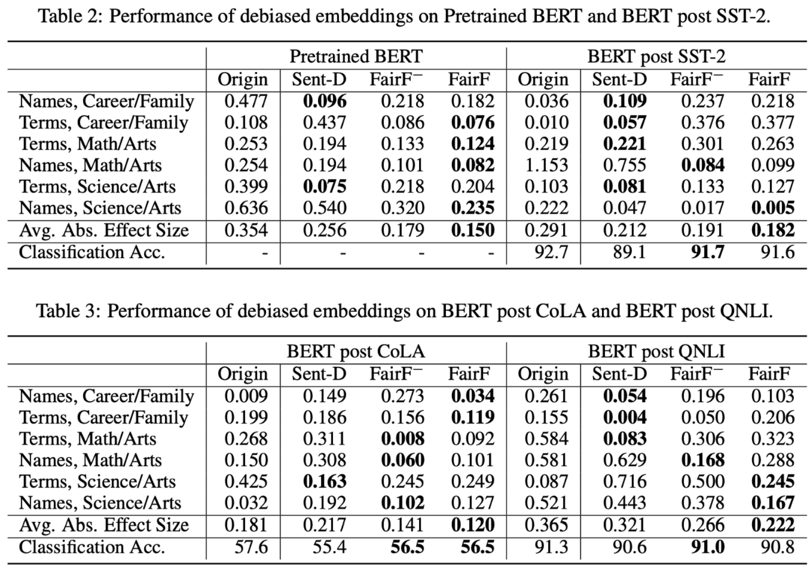


Table 2: Performance of debiased embeddings on Pretrained BERT and BERT post SST-2.

	Pretrained BERT				BERT post SST-2			
	Origin	Sent-D	FairF$^-$	FairF	Origin	Sent-D	FairF$^-$	FairF
Names, Career/Family	0.477	**0.096**	0.218	0.182	0.036	**0.109**	0.237	0.218
Terms, Career/Family	0.108	0.437	0.086	**0.076**	0.010	**0.057**	0.376	0.377
Terms, Math/Arts	0.253	0.194	0.133	**0.124**	0.219	**0.221**	0.301	0.263
Names, Math/Arts	0.254	0.194	0.101	**0.082**	1.153	0.755	**0.084**	0.099
Terms, Science/Arts	0.399	**0.075**	0.218	0.204	0.103	**0.081**	0.133	0.127
Names, Science/Arts	0.636	0.540	0.320	**0.235**	0.222	0.047	0.017	**0.005**
Avg. Abs. Effect Size	0.354	0.256	0.179	**0.150**	0.291	0.212	0.191	**0.182**
Classification Acc.	-	-	-	-	92.7	89.1	**91.7**	91.6

Table 3: Performance of debiased embeddings on BERT post CoLA and BERT post QNLI.

	BERT post CoLA				BERT post QNLI			
	Origin	Sent-D	FairF$^-$	FairF	Origin	Sent-D	FairF$^-$	FairF
Names, Career/Family	0.009	0.149	0.273	**0.034**	0.261	**0.054**	0.196	0.103
Terms, Career/Family	0.199	0.186	0.156	**0.119**	0.155	**0.004**	0.050	0.206
Terms, Math/Arts	0.268	0.311	**0.008**	0.092	0.584	**0.083**	0.306	0.323
Names, Math/Arts	0.150	0.308	**0.060**	0.101	0.581	0.629	**0.168**	0.288
Terms, Science/Arts	0.425	**0.163**	0.245	0.249	0.087	0.716	0.500	**0.245**
Names, Science/Arts	0.032	0.192	**0.102**	0.127	0.521	0.443	0.378	**0.167**
Avg. Abs. Effect Size	0.181	0.217	0.141	**0.120**	0.365	0.321	0.266	**0.222**
Classification Acc.	57.6	55.4	**56.5**	**56.5**	91.3	90.6	**91.0**	90.8

Experiments: Visualization

T-SNE plot: plot sentence embedding mean of each words contextualized in sentence templates.

After debiasing step, word “he” and “she” have more equal distance to other objectives.

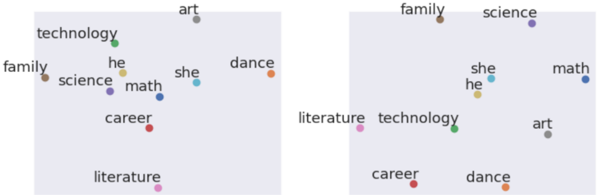


Figure 3: T-SNE plots of sentence embedding mean of each words contextualized in templates. The left-hand side is from the original pretrained BERT; the right-hand side is from our FairFil.

Acknowledgement

Advisor: Lawrence Carin

Acknowledgement

Yiran Chen

Rong Ge

Ricardo Henao

Vahid Tarokh

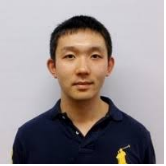
Committee Members

Acknowledgement

Renqiang Min

Jingjing Liu

Zhe Gan



Yu Cheng

Host during my internship at NEC Labs and Microsoft.

And many of my smart and diligent group mates!

Thank you!

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Devon Hjelm, and Aaron Courville. Mutual information neural estimation. In *International Conference on Machine Learning*, pages 530–539, 2018.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, pages 2172–2180, 2016.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data. In *EMNLP*, 2017a.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark, September 2017b. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186, 2019.

G Hinton. Neural networks for machine learning. coursera,[video lectures], 2012.

Guillaume Lample, Sandeep Subramanian, Eric Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. Multiple-attribute text rewriting. In *International Conference on Learning Representations*, 2019.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, 2019.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.