UNSUPERVISED DATA AUGMENTATION FOR CONSISTENCY TRAINING

Qizhe Xie[1,2], **Zihang Dai**[1,2], **Eduard Hovy**[2], **Minh-Thang Luong**[1], **Quoc V. Le**[1]
[1] Google Research, Brain Team, [2] Carnegie Mellon University
{qizhex, dzihang, hovy}@cs.cmu.edu, {thangluong, qvl}@google.com

Abstract

- Consistency training: to constrain model predictions to be **invariant to input noise**

Task	Error rate	# examples	# examples (cf)
IMDb	4.20	20	25000
CIFAR-10	2.7	4000	50000
BERT			
ImageNet			

Introduction

- Consistency training methods simply regularize model predictions to be invariant to small noise applied to either input examples
- A good model should be robust to any small change in an input example or hidden states.
- The authors investigated the role of noise injection in consistency training --> **Unsupervised Data Augmentation**

Contribution

• First, we show that state-of-the-art **data augmentations found in supervised learning** can also serve as a **superior source of noise under the consistency** enforcing semi-supervised framework.
• Second, we show that UDA can match and even outperform purely supervised learning that uses orders of magnitude more labeled data.
• Finally, we show that **UDA combines well with transfer learning**, e.g., when fine-tuning from BERT, and is effective at high-data regime, e.g. on ImageNet.

Supervised Data Augmentation

- Data augmentation aims at creating novel and realistic-looking training data by applying a transformation to an example, without changing its label
- Despite the promising results, data augmentation
  is mostly regarded as the “cherry on the cake” which provides a **steady but limited performance boost**
  - because these augmentations has so far only been applied to a set of labeled examples which is usually of a small size

Unsupervised Data Augmentation

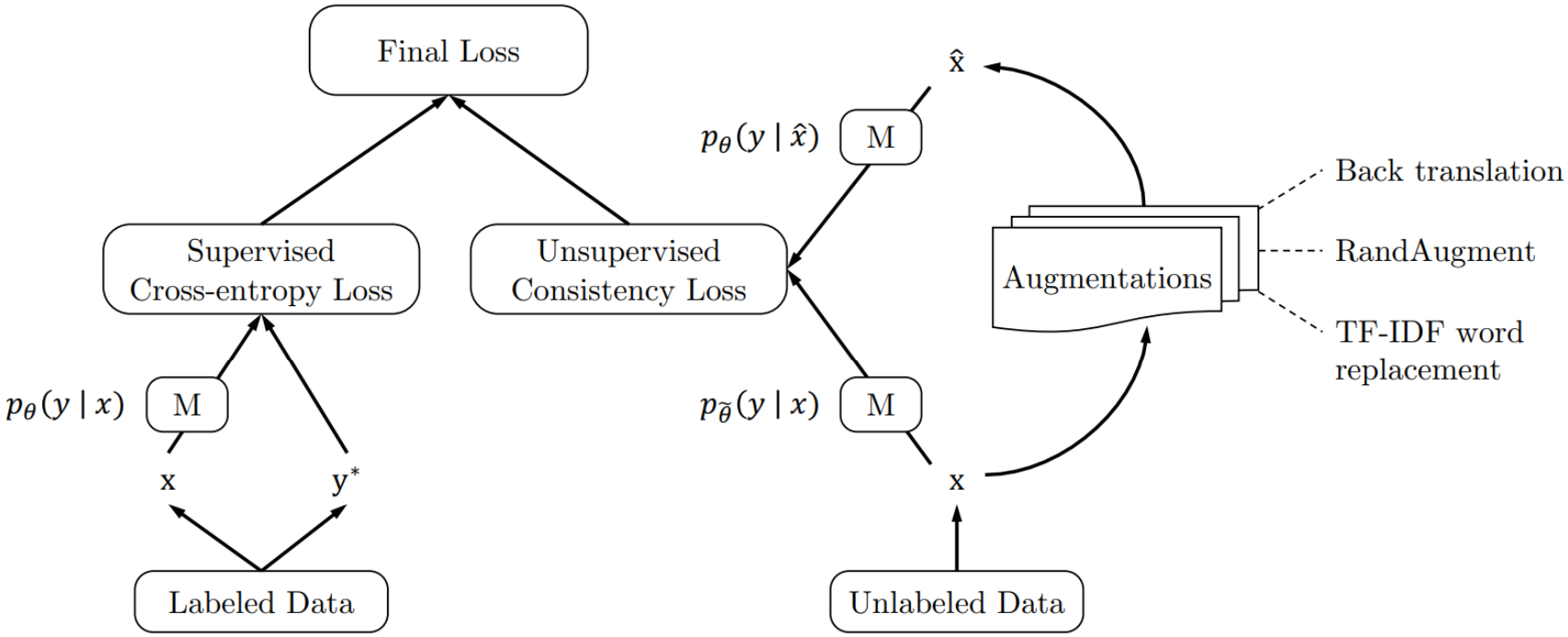
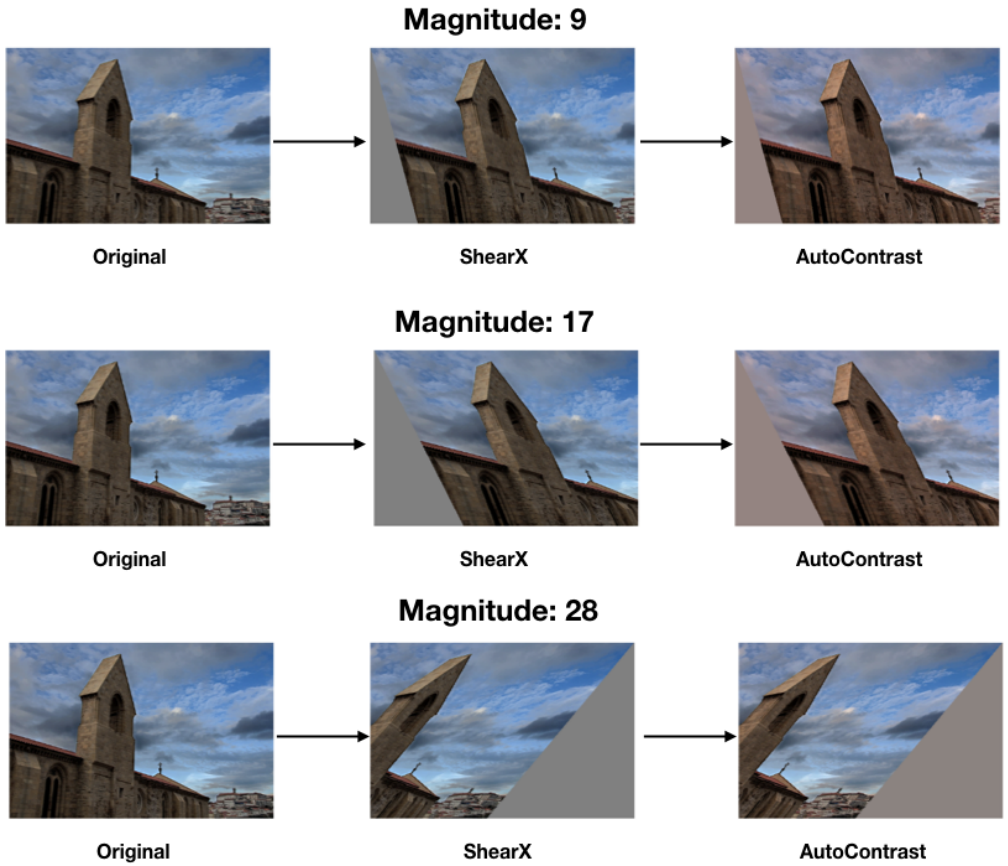


Figure 1: Training objective for UDA, where M is a model that predicts a distribution of y given x.

$$\min_\theta \; \mathcal{J}(\theta) = \mathbb{E}_{x,y^* \in L}\left[-\log p_\theta(y^* \mid x)\right] + \lambda \mathbb{E}_{x \in U} \mathbb{E}_{\hat{x} \sim q(\hat{x}|x)}\left[\mathcal{D}_{\mathrm{KL}}\left(p_{\tilde{\theta}}(y \mid x) \,\middle\|\, p_\theta(y \mid \hat{x})\right)\right].$$

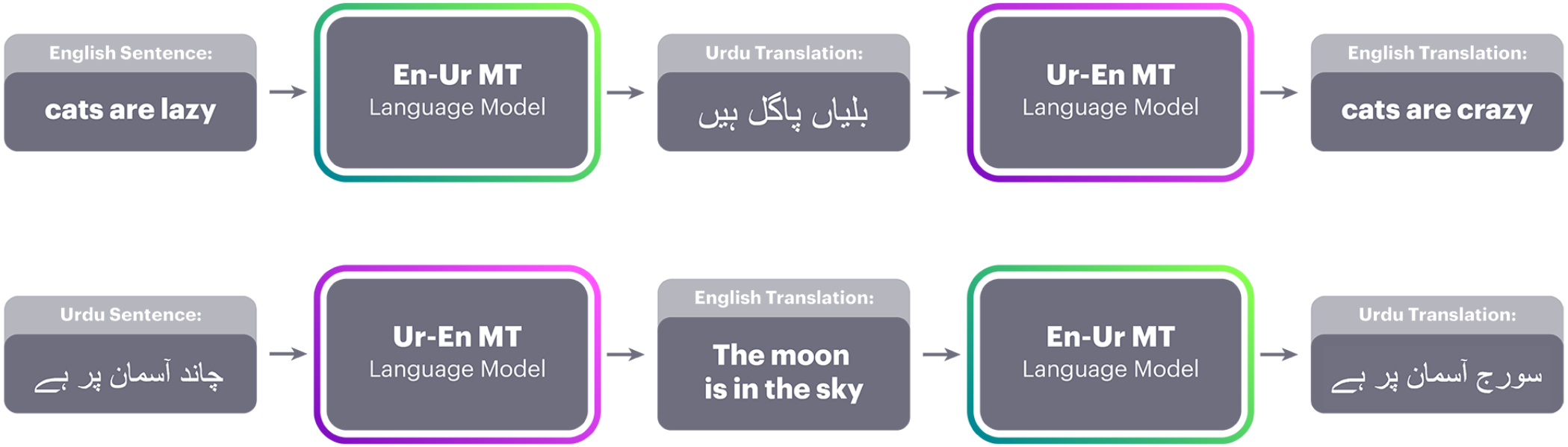
Augmentation Strategies

- RandAugment for Image Classification



Augmentation Strategies

- Back-translation for Text Classification



Augmentation Strategies

- Word replacing with TF-IDF for Text Classification
  - This method replaces uninformative words with low TF-IDF scores while **keeping those with high TF-IDF values**

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

N = total number of documents

Training Signal Annealing (TSA) for Low-data Regime

- the model often quickly overfits the limited amount of labeled data while still underfitting the unlabeled data
- TSA utilize a labeled example if the model's confidence on that example is **lower than a predefined threshold** which increases according to a schedule

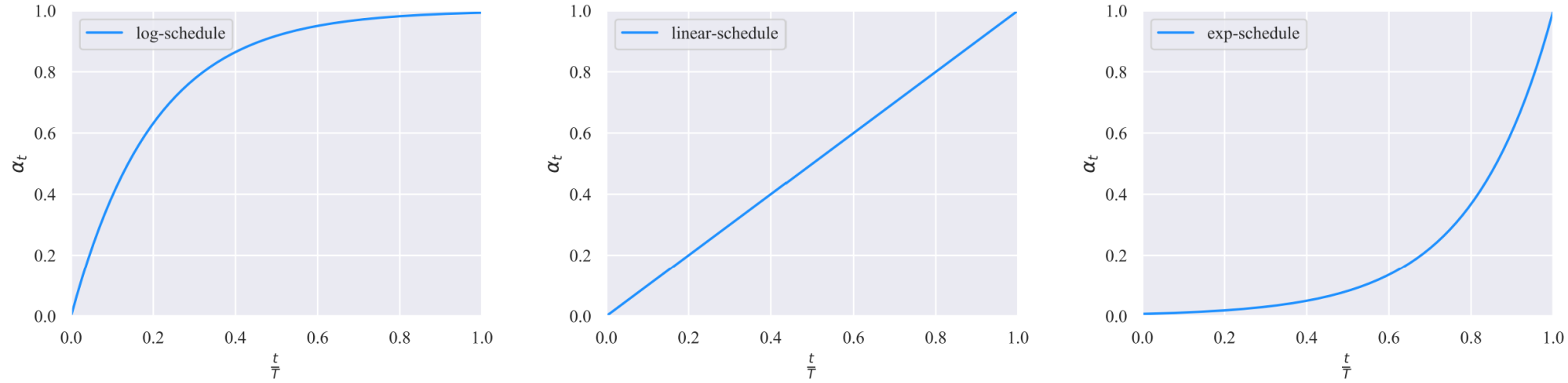


Figure 3: Three schedules of TSA. We set $\eta_t = \alpha_t * (1 - \frac{1}{K}) + \frac{1}{K}$. α_t is set to $1 - \exp(-\frac{t}{T} * 5)$, $\frac{t}{T}$ and $\exp((\frac{t}{T} - 1) * 5)$ for the log, linear and exp schedules.

Correlation between Supervised and Semi-supervised Performances

- There is a positive correlation of data augmentation's effectiveness in supervised learning and semi-supervised learning

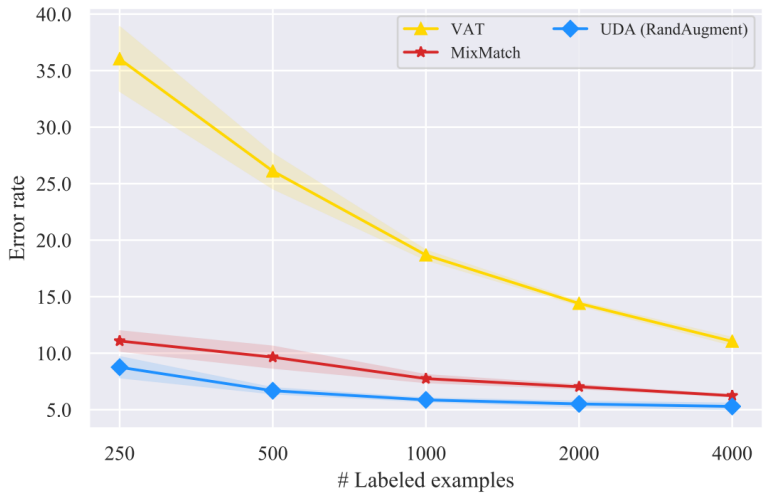
Augmentation (# Sup examples)	Sup (50k)	Semi-Sup (4k)
Crop & flip	5.36	16.17
Cutout	4.42	6.42
RandAugment	**4.23**	**5.29**

Table 1: Error rates on CIFAR-10.

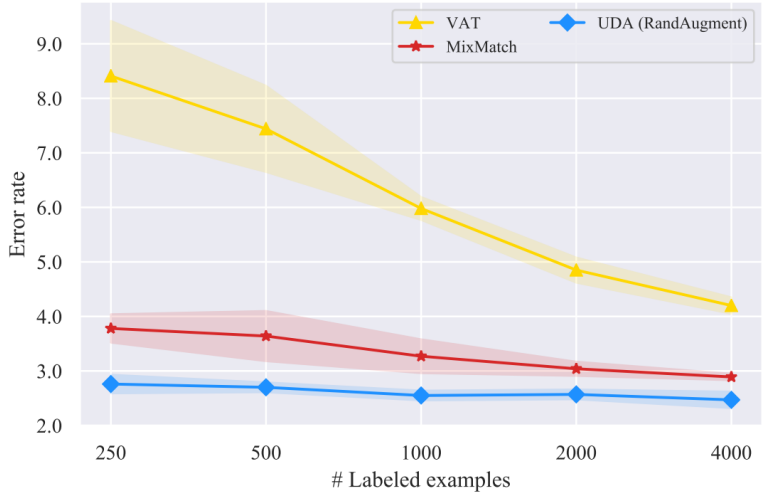
Augmentation (# Sup examples)	Sup (650k)	Semi-sup (2.5k)
✗	38.36	50.80
Switchout	37.24	43.38
Back-translation	**36.71**	**41.35**

Table 2: Error rate on Yelp-5.

Algorithm Comparison on Vision Semi-supervised Learning Benchmarks



(a) CIFAR-10



(b) SVHN

Figure 4: Comparison with two semi-supervised learning methods on CIFAR-10 and SVHN with varied number of labeled examples.

- Virtual adversarial training (VAT): an algorithm that generates adversarial Gaussian noise on input
- MixMatch: combines previous advancements in semi-supervised learning

Algorithm Comparison on Vision Semi-supervised Learning Benchmarks

- Comparisons with published results

Method	Model	# Param	CIFAR-10 (4k)	SVHN (1k)
Π-Model (Laine & Aila, 2016)	Conv-Large	3.1M	12.36 ± 0.31	4.82 ± 0.17
Mean Teacher (Tarvainen & Valpola, 2017)	Conv-Large	3.1M	12.31 ± 0.28	3.95 ± 0.19
VAT + EntMin (Miyato et al., 2018)	Conv-Large	3.1M	10.55 ± 0.05	3.86 ± 0.11
SNTG (Luo et al., 2018)	Conv-Large	3.1M	10.93 ± 0.14	3.86 ± 0.27
VAdD (Park et al., 2018)	Conv-Large	3.1M	11.32 ± 0.11	4.16 ± 0.08
Fast-SWA (Athiwaratkun et al., 2018)	Conv-Large	3.1M	9.05	-
ICT (Verma et al., 2019)	Conv-Large	3.1M	7.29 ± 0.02	3.89 ± 0.04
Pseudo-Label (Lee, 2013)	WRN-28-2	1.5M	16.21 ± 0.11	7.62 ± 0.29
LGA + VAT (Jackson & Schulman, 2019)	WRN-28-2	1.5M	12.06 ± 0.19	6.58 ± 0.36
mixmixup (Hataya & Nakayama, 2019)	WRN-28-2	1.5M	10	-
ICT (Verma et al., 2019)	WRN-28-2	1.5M	7.66 ± 0.17	3.53 ± 0.07
MixMatch (Berthelot et al., 2019)	WRN-28-2	1.5M	6.24 ± 0.06	2.89 ± 0.06
Mean Teacher (Tarvainen & Valpola, 2017)	Shake-Shake	26M	6.28 ± 0.15	-
Fast-SWA (Athiwaratkun et al., 2018)	Shake-Shake	26M	5.0	-
MixMatch (Berthelot et al., 2019)	WRN	26M	4.95 ± 0.08	-
UDA (RandAugment)	WRN-28-2	1.5M	5.29 ± 0.25	**2.55 ± 0.09**
UDA (RandAugment)	Shake-Shake	26M	3.7	-
UDA (RandAugment)	PyramidNet	26M	**2.7**	-

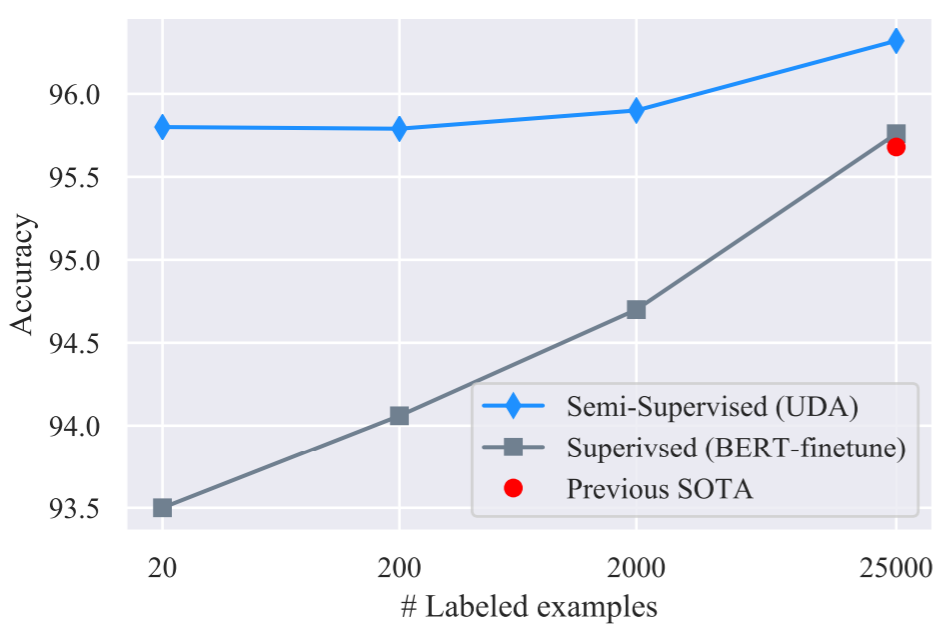
Evaluation on Text Classification Datasets

Fully supervised baseline						
Datasets (# Sup examples)		IMDb (25k)	Yelp-2 (560k)	Yelp-5 (650k)	Amazon-2 (3.6m)	Amazon-5 (3m)	DBpedia (560k)
Pre-BERT SOTA		*4.32*	2.16	29.98	3.32	34.81	0.70
$\text{BERT}_{\text{LARGE}}$		4.51	*1.89*	*29.32*	*2.63*	*34.17*	*0.64*
Semi-supervised setting							
Initialization	**UDA**	IMDb (20)	Yelp-2 (20)	Yelp-5 (2.5k)	Amazon-2 (20)	Amazon-5 (2.5k)	DBpedia (140)
Random	✗	43.27	40.25	50.80	45.39	55.70	41.14
	✓	25.23	8.33	41.35	16.16	44.19	7.24
$\text{BERT}_{\text{BASE}}$	✗	18.40	13.60	41.00	26.75	44.09	2.58
	✓	5.45	2.61	33.80	3.96	38.40	1.33
$\text{BERT}_{\text{LARGE}}$	✗	11.72	10.55	38.90	15.54	42.30	1.68
	✓	4.78	2.50	33.54	3.93	37.80	1.09
$\text{BERT}_{\text{FINETUNE}}$	✗	6.50	2.94	32.39	12.17	37.32	-
	✓	**4.20**	**2.05**	**32.08**	**3.50**	**37.12**	-

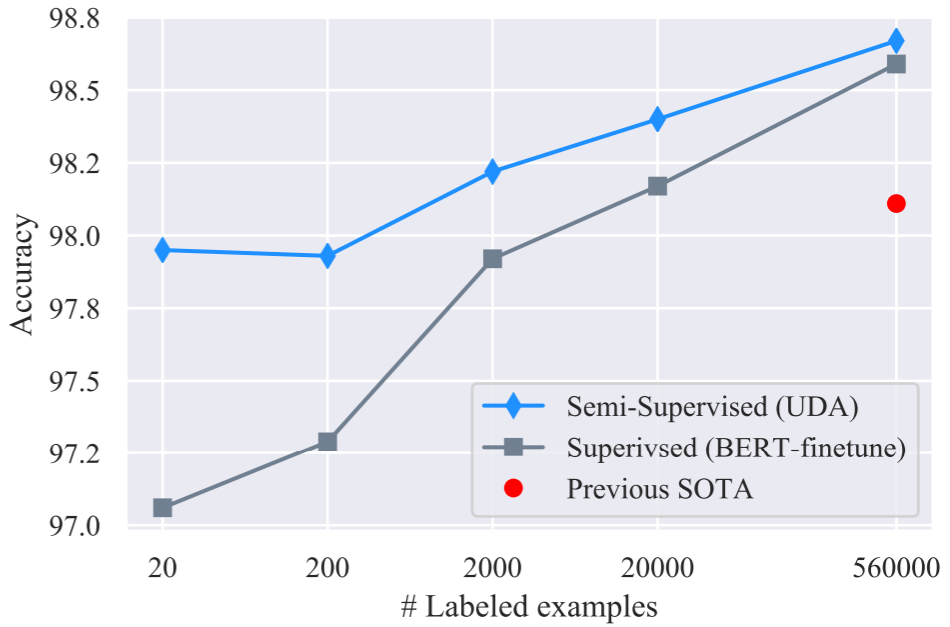
Evaluation on Text Classification Datasets

- Even with very few labeled examples, UDA can offer decent or even competitive performances compared to the SOTA model trained with full supervised data. Particularly, on binary sentiment analysis tasks, with only 20 supervised examples, UDA outperforms the previous SOTA
  trained with full supervised data on IMDb and is competitive on Yelp-2 and Amazon-2.
- UDA is complementary to transfer learning / representation learning. As we can see, when initialized with BERT and further finetuned on in-domain data, UDA can still significantly reduce the error rate from 6.50 to 4.20 on IMDb.
- We also note that for five-category sentiment classification tasks, there still exists a clear gap between UDA with 500 labeled examples per class and BERT trained on the entire supervised set. Intuitively, five-category sentiment classifications are much more difficult than their binary counterparts. This suggests a room for further improvement in the future.

Evaluation on Text Classification Datasets



(a) IMDb



(b) Yelp-2

Figure 5: Accuracy on IMDb and Yelp-2 with different number of labeled examples. In the large-data regime, with the full training set of IMDb, UDA also provides robust gains.

Scalability Test on the ImageNet Dataset

Methods	SSL	10%	100%
ResNet-50	✗	55.09 / 77.26	77.28 / 93.73
w. RandAugment		58.84 / 80.56	78.43 / 94.37
UDA (RandAugment)	✓	**68.78 / 88.80**	**79.05 / 94.49**

Table 5: Top-1 / top-5 accuracy on ImageNet with 10% and 100% of the labeled set. We use image size 224 and 331 for the 10% and 100% experiments respectively.

Ablation Studies for TSA

TSA schedule	Yelp-5	CIFAR-10
✗	50.81	5.67
log-schedule	49.06	5.67
linear-schedule	45.41	**5.29**
exp-schedule	**41.35**	7.81

Table 6: Ablation study for Training Signal Annealing (TSA) on Yelp-5 and CIFAR-10. The shown numbers are error rates.