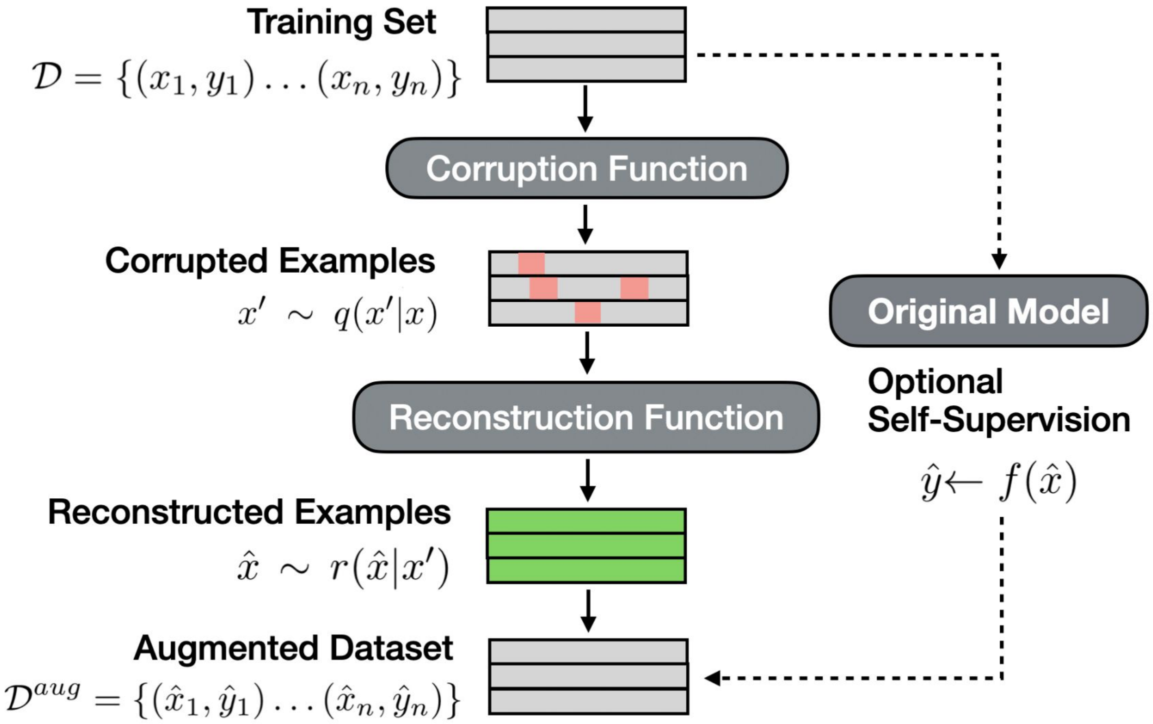
SSMBA: Self-Supervised Manifold Based Data Augmentation for Improving Out-of-Domain Robustness

Nathan Ng, Kyunghyun Cho, Marzyeh Ghassemi
University of Toronto
New York University
EMNLP 2020

Motivation

- Out-of-domain problem
  - Bias in data collection
  - Distribution shift over time
- Data augmentation
  - Synthetically generate training examples by pertubing the input
  - In NLP, it is difficult because of shifting in semantic after the perturbation
- This paper
  - Propose a data augmentation method
  - Using a Denoising Auto-Encoder as generative model
  - Using a reconstruction function to project back on

Framework


Training Set
$\mathcal{D} = \{(x_1, y_1) \ldots (x_n, y_n)\}$
Corruption Function
Corrupted Examples
$x' \sim q(x'|x)$
Reconstruction Function
Reconstructed Examples
$\hat{x} \sim r(\hat{x}|x')$
Augmented Dataset
$\mathcal{D}^{aug} = \{(\hat{x}_1, \hat{y}_1) \ldots (\hat{x}_n, \hat{y}_n)\}$
Original Model
Optional
Self-Supervision
$\hat{y} \leftarrow f(\hat{x})$

Algorithm

Algorithm 1 SSMBA

Require: perturbation function q
reconstruction function r

Input: Dataset $\mathcal{D} = \{(x_1, y_1) \dots (x_n, y_n)\}$
number of augmented examples m

function SSMBA($\mathcal{D}$, m)
  train a model f on $\mathcal{D}$
  for $(x_i, y_i) \in \mathcal{D}$ **do**
    for $j \in 1 \dots m$ **do**
      sample perturbed $x'_{ij} \sim q(x'|x_i)$
      sample reconstructed $\hat{x}_{ij} \sim r(\hat{x}|x'_{ij})$
      generate $\hat{y}_{ij} \leftarrow f(\hat{x}_{ij})$ or preserve the original y_i
    end for
  end for
  let $\mathcal{D}^{aug} = \{(\hat{x}_{ij}, \hat{y}_{ij})\}_{i=1\dots n, j=1\dots m}$
  augment $\mathcal{D}' \leftarrow \mathcal{D} \cup \mathcal{D}^{aug}$
  return $\mathcal{D}'$
end function

Baseline

1. Easy Data Augmentation (EDA): randomly replaces words by synonyms, insert, swaps, deletes words
2. Conditional Bert Contextual Augmentation (CBERT): finetune a class-condition BERT model and use it to generate sentences
3. Unsupervised Data Augmentation (UDA): translate and back translate
4. Reward Augmented Maximum Likelihood (RAML): sample noisy target sentences based on Hamming distance(MT only)
5. Word Dropout: randomly set embedding of words to zeros
6. SwitchOut: apply RAML on both source and target sentences (MT only)

Result (Sentiment Analysis)

		AR-Full		AR-Clothing		Movies		Yelp		Average	
Model	Augmentation	ID	OOD	ID	OOD	ID	OOD	ID	OOD	ID	OOD
RNN	None	69.46	66.32	69.25	67.80	90.74	71.94	62.51	61.28	70.16	66.17
	EDA	67.32	64.47	66.87	65.21	88.43	68.3	58.39	57.19	67.56	63.55
	CBERT	69.94	66.77	69.56	68.10	**91.01**	72.11	63.17	61.75	70.17	66.57
	UDA	69.92	66.97	69.98	68.24	90.05	69.73	63.40	62.13	70.64	66.53
	SSMBA	**70.38***†	**67.41***†	**70.19**	**68.60***†	89.61	**73.20**	**63.85**	**62.83***†	**70.96**	**67.31**
CNN	None	70.67	67.64	70.14	68.52	92.92	72.11	65.13	64.46	71.68	67.63
	EDA	68.52	66.03	67.76	66.17	91.22	74.20	60.99	59.88	69.13	65.65
	CBERT	70.62	67.70	70.13	68.23	92.92	71.56	65.09	64.19	71.65	67.49
	UDA	70.80	68.06	70.29	68.70	92.63	72.55	65.22	64.32	71.77	67.89
	SSMBA	**71.10***	**68.18***	**70.74**	**69.04***	**92.93**	**74.67**	**65.59**	**64.81***†	**72.11**	**68.33**

Table 2: Average in-domain (ID) and out-of-domain (OOD) accuracy (%) for models trained on sentiment analysis datasets. Average performance across datasets is weighted by number of domains contained in each dataset. Accuracies marked with a $*$ and $\dagger$ are statistically significantly higher than unaugmented models and the next best model respectively, both with $p < 0.01$.

MNLI and MT

Augmentation	MNLI ID	MNLI OOD	ANLI ID	ANLI OOD
None	84.29	80.61	42.54	**43.80**
EDA	83.44	80.34	45.59	42.77
CBERT	84.24	80.34	46.68	43.53
UDA	84.24	80.99	45.85	42.89
SSMBA	**85.71**	**82.44***†	**48.46***†	**43.80**

Table 3: Average in-domain and out-of-domain accuracy (%) for RoBERTa models trained on NLI tasks. Accuracies marked with a $*$ and $\dagger$ are statistically significantly higher than unaugmented models and the next best model respectively, both with $p < 0.01$.

Augmentation	OPUS ID	OPUS OOD	de→rm ID	de→rm OOD
None	**56.99**	10.24	51.53	12.23
Word Dropout	56.26	10.15	50.23	12.23
RAML	56.76	10.10	51.52	12.49
SwitchOut	55.50	9.27	51.34	13.59
SSMBA	54.88	**10.65**	**51.97**	**14.67***†

Table 5: Average in-domain and out-of-domain BLEU for models trained on OPUS (de→en) and de→rm data. Scores marked with a $*$ and $\dagger$ are statistically significantly higher than baseline transformers and the next best model, both with $p < 0.01$.

Discussion: Label generation

Label preservation: keep the **original label**

Generate **soft-label** using a poor classifier

Generate **hard-label** using a poor classifier

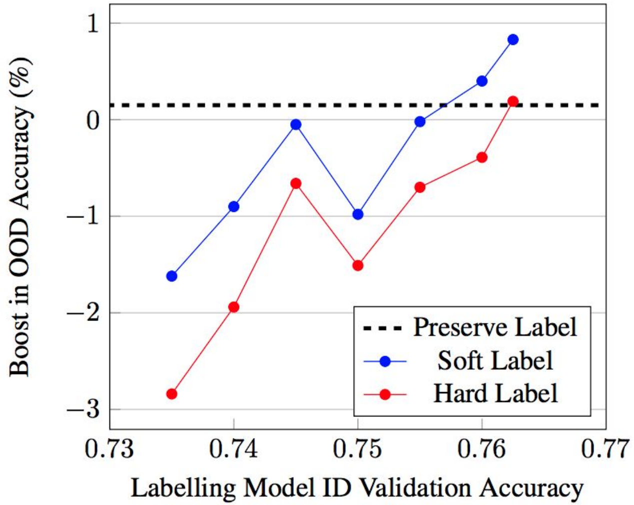


Figure 8: Boost in OOD accuracy (%) of models trained with augmented data labelled with different supervision models and label generation methods.

Discussion: Amount of Augmentation

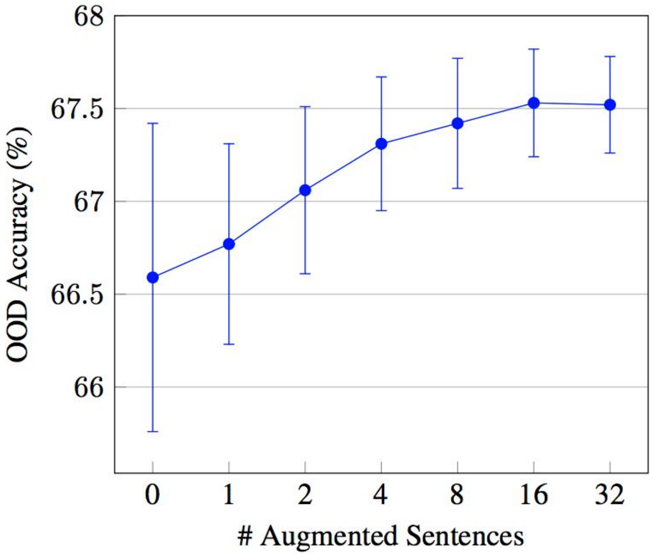


Figure 7: OOD accuracy (%) of models trained with different amounts of SSMBA augmentation. 0 augmentation corresponds to a baseline model. Error bars show standard deviation in OOD accuracy across models.