NAACL2022

Contrastive Learning for Prompt-based Few-shot Language Learners

Yiren Jian, Chongyang Gao, Soroush Vosoughi

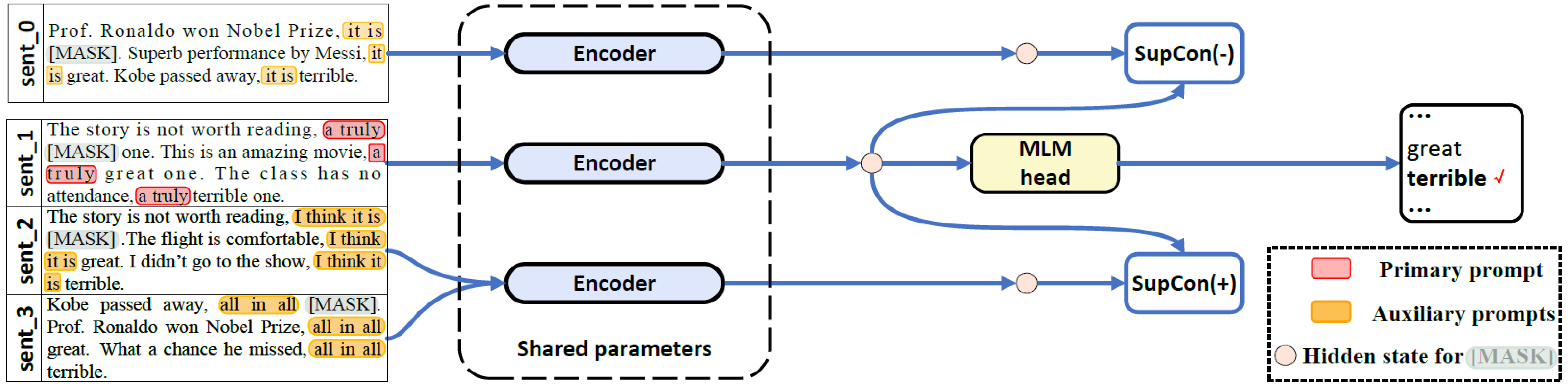
DARTMOUTH Department of Computer Science

Northwestern McCORMICK SCHOOL OF ENGINEERING Computer Science

Contributions

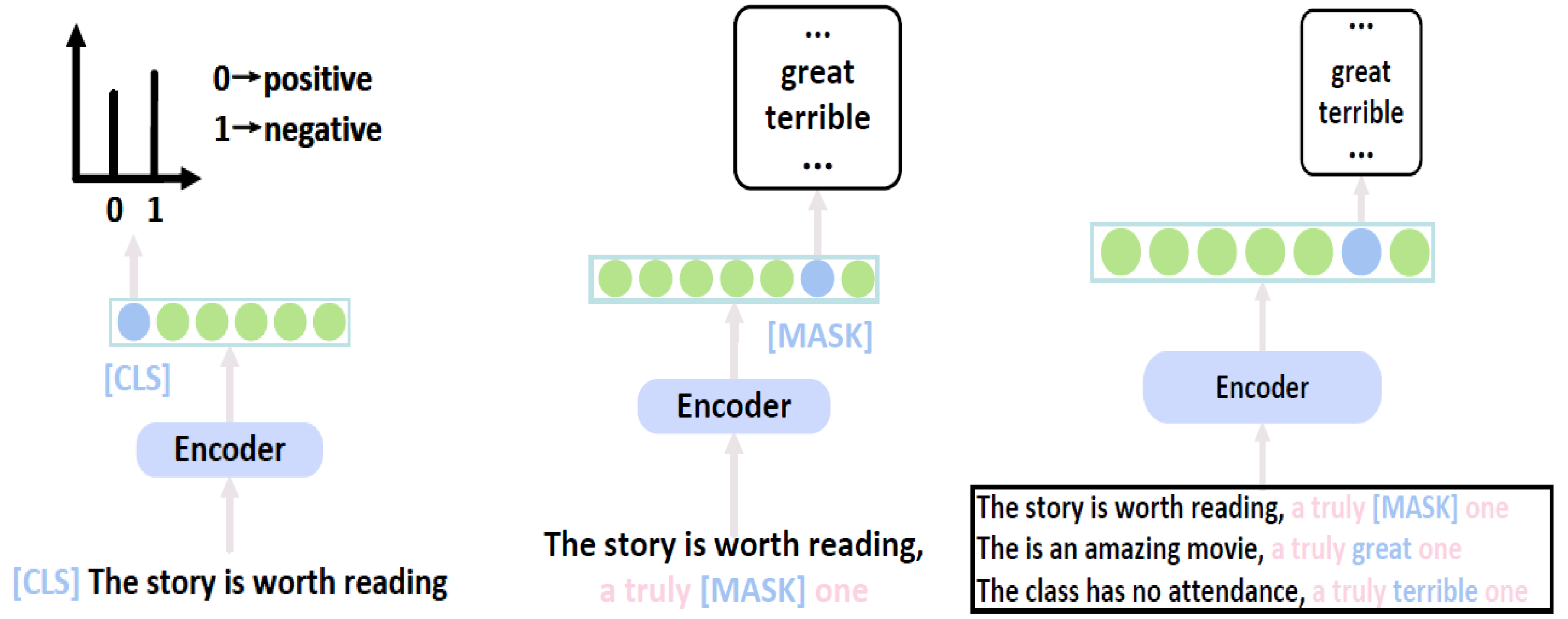
- A Supervised Contrastive Learning framework for prompt-based few-shot learners
- An effective data augmentation method using prompts for contrastive learning with prompt-based learners.

Overview



Overview of our proposed method.

Different kinds of fine tuning



Conventional fine tuning, classifying at hidden states of [CLS].

Prompt-based fine tuning, predicting the best suited missing word.

Prompt-based fine tuning with demonstrations to help model to know what is "great" / "terrible".

Random templates/demonstrations vs.

Task	LM-BFF	SR	RI	RS	RD	EDA	ours
SST-2	89.2	90.7	**90.8**	90.7	90.7	90.5	90.6
Subj	88.6	90.6	90.8	**91.0**	90.5	89.1	90.4
SST-5	47.9	47.7	49.2	48.2	47.9	46.7	**49.5**
CoLA	6.1	5.8	6.5	4.9	4.0	3.9	**10.2**
TREC	82.8	78.1	80.7	79.0	80.7	80.6	**83.3**
MNLI	61.0	61.8	62.4	61.0	58.1	58.9	**64.0**
-mm	62.5	63.6	64.8	62.7	60.3	60.9	**65.5**
SNLI	66.9	63.1	66.4	67.2	65.2	62.2	**69.9**
QNLI	60.7	65.3	65.3	**67.4**	64.8	62.5	66.4†
QQP	62.5	64.5	65.8	68.0	63.2	61.0	**68.8**
RTE	64.3	61.4	61.4	61.3	62.1	61.1	**65.1**
MRPC	75.5	77.6	77.7	**79.3**	78.7	79.1	78.2†
MR	83.3	85.5	85.5	85.5	85.3	85.6	**85.8**
MPQA	83.6	82.2	84.4	84.4	83.9	82.8	**84.6**
CR	88.9	88.9	88.2	88.3	88.5	87.1	**89.4**

Comparing our random templates/demonstrations as data augmentation to Synonym Replacement (SR), Random Insertion (RI), Random Swapping (RS), Random Deletion (RD) and EDA [4].

Algorithm

Algorithm 1 Our method

$Max\_Step = 1000$,
LM: Language model,
$Train\_Set$: Training set,
$Sample$: Randomly sampling function,
$Concatenate$: The function to concatenate two strings,
CE: Cross Entropy loss,
$SupCon$: Supervised Contrastive loss.
for i in Max_Step **do**
  $sent, y = Sample(Train\_Set)$
  $demo_1 = Sample(Train\_Set)$
  $demo_2 = Sample(Train\_Set)$
  $input_1 = concatenate(sent, demo_1)$
  $input_2 = concatenate(sent, demo_2)$
  ▷ Learning from MLM Loss
  $output_1 = LM(input_1)$
  $L_{MLM} = CE(output_1, y)$
  $L_{MLM}.backward()$
  $optimizer.step()$
  ▷ Learning from SupCon Loss
  $output_2 = LM(input_2)$
  $L_{SupCon} = SupCon(output_1, output_2)$
  $L_{SupCon}.backward()$
  $optimizer.step()$
end for

Ensemble vs. our single model

Task	LM-BFF +ours	LM-BFF ensemble
SST-5 (acc)	**49.5** (1.1)	48.0 (0.8)
CoLA (Matt.)	**10.2** (5.8)	7.5 (4.7)
MNLI (acc)	**63.3** (2.4)	62.2 (1.8)
MNLI-mm (acc)	**65.1** (2.4)	64.0 (1.8)
QNLI (acc)	**66.4** (3.5)	63.8 (2.7)
MR (acc)	**85.8** (0.6)	85.7 (0.7)

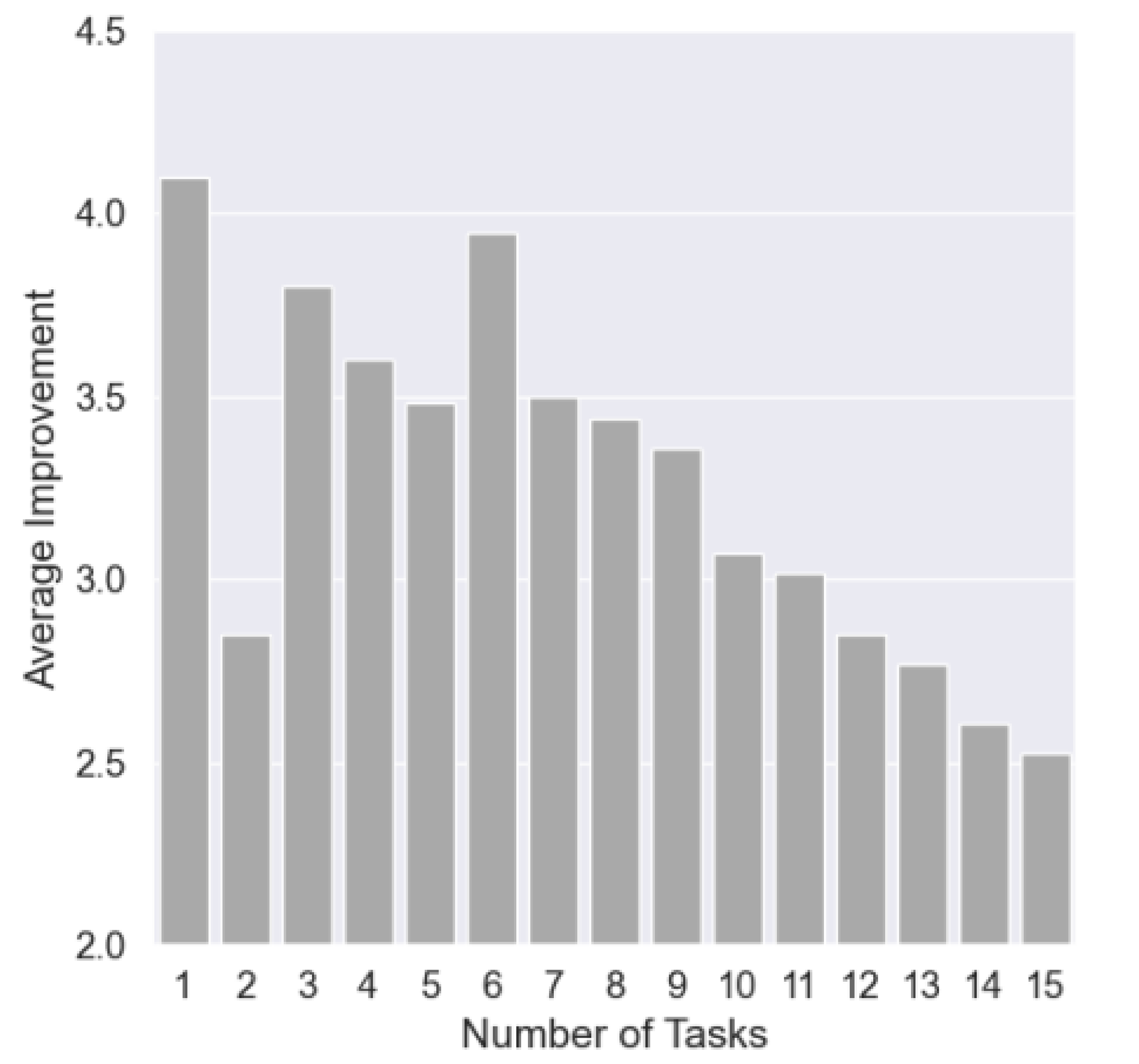
Comparing our single model trained with Sup-Con loss to an ensemble of 20 models

With demonstrations vs. Without demonstrations

Task	LM-BFF	LM-BFF + ours	PET	PET + ours
SST-2 (acc)	89.2 (1.3)	**90.6** (0.1)	88.4 (1.0)	**89.9** (0.6)
Subj (acc)	88.6 (3.3)	**90.4** (1.1)	89.2 (1.5)	**90.6** (1.6)
SST-5 (acc)	47.9 (0.8)	**49.5** (1.1)	46.0 (0.9)	**48.8** (1.2)
CoLA (Matt.)	6.1 (5.3)	**10.2** (5.8)	3.5 (3.4)	**5.9** (3.3)
TREC (acc)	82.8 (3.1)	**83.3** (1.5)	77.8 (9.1)	**82.3** (4.6)
MNLI (acc)	61.0 (2.1)	**64.0** (2.0)	58.2 (1.1)	**58.9** (3.1)
MNLI-mm (acc)	62.5 (2.1)	**65.5** (2.7)	59.8 (1.2)	**61.0** (3.3)
SNLI (acc)	66.9 (2.4)	**69.9** (2.4)	63.1 (2.5)	**65.7** (3.9)
QNLI (acc)	60.7 (1.7)	**66.4** (3.5)	61.5 (3.3)	**63.5** (3.7)
QQP (acc)	62.5 (2.6)	**68.8** (3.8)	61.9 (3.5)	**65.7** (4.3)
RTE (acc)	64.3 (2.7)	**65.1** (3.5)	60.9 (4.7)	**65.1** (3.5)
MRPC (F1)	75.5 (5.2)	**78.2** (3.1)	70.6 (6.0)	**75.7** (6.1)
MR (acc)	83.3 (1.4)	**85.8** (0.6)	85.0 (0.6)	**85.2** (0.9)
MPQA (acc)	83.6 (1.8)	**84.6** (1.5)	81.3 (2.6)	**81.8** (2.4)
CR (acc)	88.9 (1.0)	**89.4** (1.0)	89.3 (1.0)	**90.5** (0.5)

Few-shot results of baseline methods and ours.

Average improvements



The average improvements achieved by our method on the top K hardest tasks

[1] Khosla, Prannay, et al. "Supervised contrastive learning." Neurips 33 (2020): 18661-18673.
[2] Gao, Tianyu, Adam Fisch, and Danqi Chen. "Making Pre-trained Language Models Better Few-shot Learners." ACL/ *IJCNLP*. 2021.
[3] Schick, Timo, and Hinrich Schütze. "Exploiting cloze questions for few shot text classification and natural language inference." arXiv preprint arXiv:2001.07676 (2020).
[4] Wei, Jason, and Kai Zou. "EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks." *EMNLP-IJCNLP*. 2019.