AD-DROP: Attribution-Driven Dropout for Robust Language Model Fine-Tuning

Tao Yang[1], Jinghao Deng[1], Xiaojun Quan[1], Qifan Wang[2], Shaoliang Nie[2]

[1] School of Computer Science and Engineering, Sun Yat-sen University
[2] Meta AI

NeurIPS 2022

Outlines

Outlines

Introduction

- **Dropout**

- Fine-tuning PrLMs is apt to suffer from overfitting. (Large model v.s. Small data)
- Dropout that randomly dropping a proportion of units is a widely used regularizer to mitigate overfitting.
- While existing research has rarely examined its effect on the self-attention mechanism.

Introduction

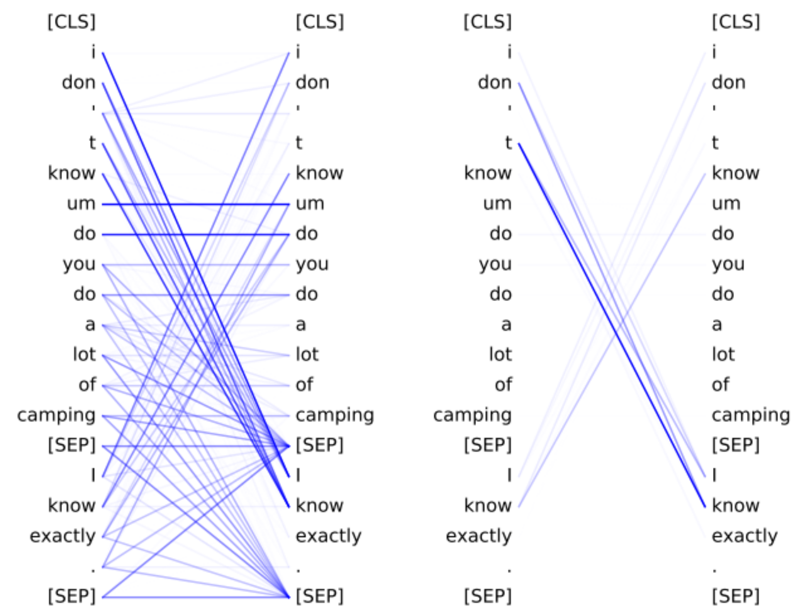
- **Attribution**

  - Attribution is an interpretability method that attributes predictions to the input features.

- **Self-attention Attribution**

  - Integrated Gradient

  - Provide a more accurate saliency measure than attention score.



(a) Attention Score (b) Attribution Score

Yaru Hao, et. al. Self-attention attribution: Interpreting information interactions inside transformer. AAAI, volume 35, pages 12963–12971, 2021.

Introduction

- **Prior Attribution Experiments**

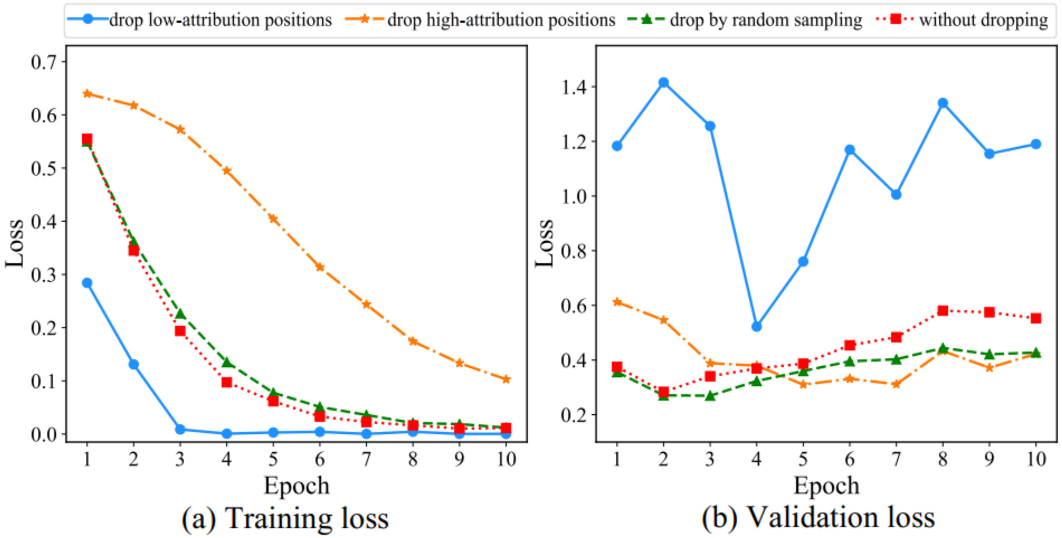


Figure 2: Results of training and validation losses when fine-tuning RoBERTa with different dropping strategies on MRPC. The dropping rate is set to 0.3 if it applies.

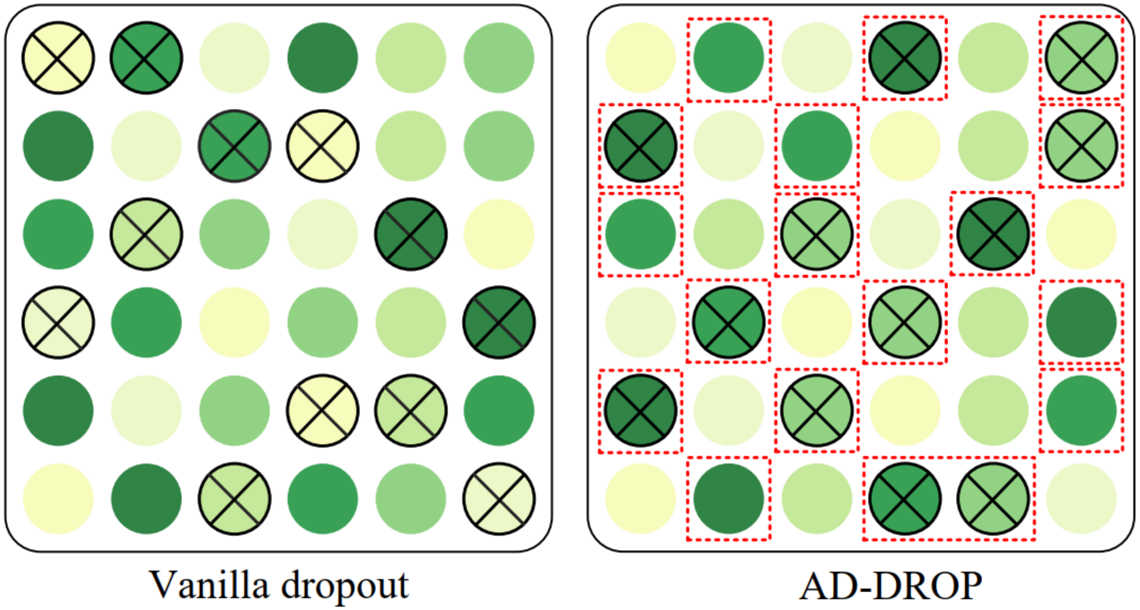
- Dropping low attribution positions makes the model fit the training data rapidly, whereas it performs poorly on the development set. (Accelerate Overfitting)

- Dropping high attribution positions reduces the fitting speed significantly.

Introduction

- **AD-DROP**

  - Attention positions are not equally important in preventing overfitting.



  - Darker attention positions indicate higher attribution scores.
  - Red-dotted boxes refer to candidate discard regions with high attribution scores.
  - AD-DROP focuses on dropping positions in candidate discard regions.

Outlines

Methodology

- **AD-DROP**

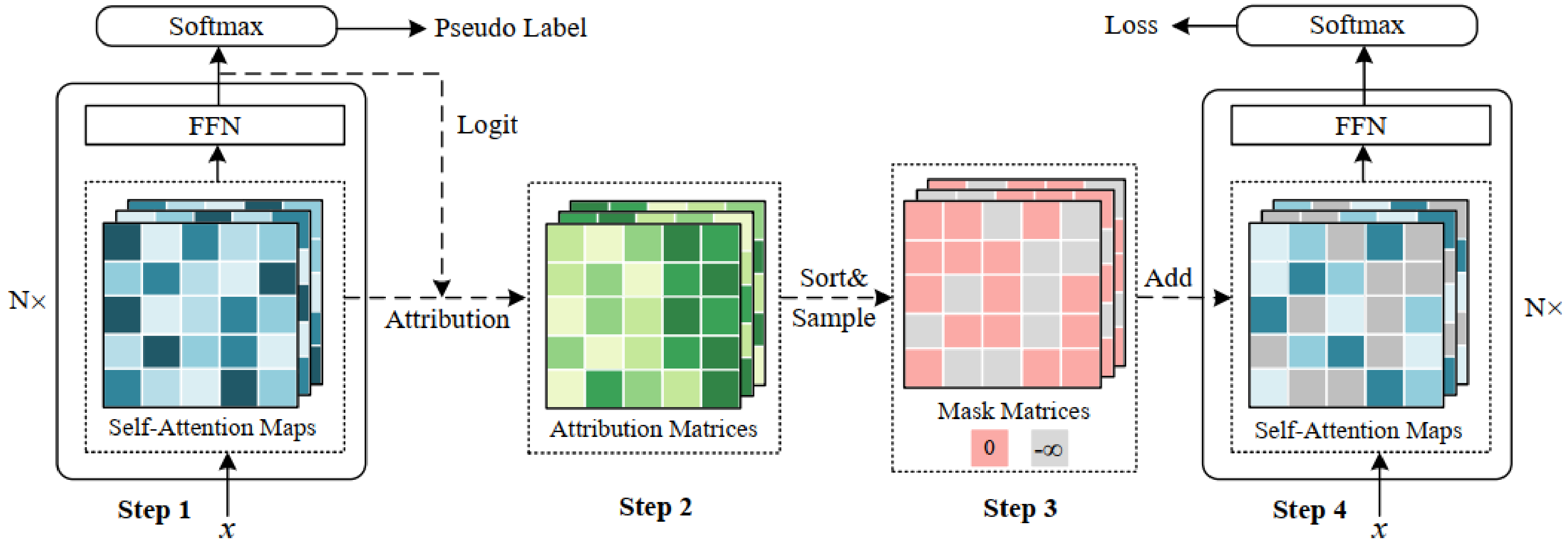


Figure 3: Illustration of AD-DROP in four steps. (1) Conduct the first forward computation to obtain pseudo label $\tilde{c}$. (2) Generate attribution matrices $\mathbf{B}$ via computing the gradient of logit output $F_{\tilde{c}}(\mathbf{A})$ with respect to each attention head. (3) Sort $\mathbf{B}$ and strategically drop some positions to produce mask matrices $\mathbf{M}$. (4) Feed $\mathbf{M}$ into the next forward computation to compute the final loss.

Methodology

- **Cross-tuning**

Algorithm 1 Cross-tuning

Input: shuffled training samples $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, PrLM F with parameters $\mathbf{W}$
Output: updated parameters $\widetilde{\mathbf{W}}$

Initialize F with $\mathbf{W}$, $epoch = 1$
while not converged **do**
    Calculate the prediction $P_F(y_i|x_i)$ and loss via forward computation.
    if $epoch\%2 == 1$ **then**
        Backpropagate the loss to update model parameters $\mathbf{W}$.
    else
        Perform AD-DROP by Eq. (4)-(7) to obtain mask matrices $\mathbf{M} = [\mathbf{M}_1, \mathbf{M}_2, \cdots, \mathbf{M}_H]$.
        Calculate the new prediction $P_F(y_i|x_i)$ and new loss by feeding $\mathbf{M}$ into Eq. (1).
        Backpropagate the new loss to update model parameters $\mathbf{W}$.
    $epoch = epoch + 1$
return $\widetilde{\mathbf{W}} = \mathbf{W}$

---

original fine-tuning

AD-DROP

Outlines

1. Introduction
2. Methodology
3. Experiment
4. Analysis
5. Conclusion

Experiment

- **Overall results**

Table 1: Overall results of fine-tuned models on the GLUE benchmark. The symbol † denotes results directly taken from the original papers. The best average results are shown in bold.

Methods	SST-2	MNLI	QNLI	QQP	CoLA	STS-B	MRPC	RTE	Average	
Development										
$\text{BERT}_{\text{base}}$	92.3	84.6	91.5	91.3	60.3	89.9	85.1	70.8	83.23	
+*SCAL*† [17]	92.8	84.1	90.9	91.4	61.7	-	-	69.7	-	
+*SuperT*† [48]	93.4	84.5	91.3	91.3	58.8	89.8	87.5	72.5	83.64	
+*R-Drop*† [18]	93.0	85.5	92.0	91.4	62.6	89.6	87.3	71.1	84.06	
+AD-DROP	93.9	85.1	92.3	91.8	64.6	90.4	88.5	75.1	**85.21**	+1.98
$\text{RoBERTa}_{\text{base}}$	95.3	87.6	92.9	91.9	64.8	90.9	90.7	79.4	86.69	
+*R-Drop* [18]	95.2	87.8	93.2	91.7	64.7	91.2	90.5	80.5	86.85	
+*HiddenCut*† [15]	95.8	88.2	93.7	92.0	66.2	91.3	92.0	83.4	87.83	
+AD-DROP	95.8	88.0	93.5	92.0	66.8	91.4	92.2	84.1	**87.98**	+1.29
Test										
$\text{BERT}_{\text{base}}$	93.6	84.7	90.4	89.3	52.8	85.6	81.4	68.4	80.78	
+AD-DROP	94.3	85.2	91.6	89.4	53.3	86.6	84.1	68.7	**81.65**	+0.87
$\text{RoBERTa}_{\text{base}}$	94.8	87.5	92.8	89.6	58.3	88.7	86.3	75.1	84.14	
+AD-DROP	95.9	87.6	93.4	89.5	58.5	89.3	87.9	76.0	**84.76**	+0.62

Analysis

- **Ablation study**

Table 2: Results of ablation studies, where *r/w* means "replace with" and *w/o* means "without".

Methods	CoLA	STS-B	MRPC	RTE
$BERT_{base}$	60.3	89.9	85.1	70.8
+AD-DROP (GA)	**64.6**	90.4	**88.5**	**75.1**
r/w IGA	63.8	**90.7**	**88.5**	74.4
r/w AA	63.6	90.0	88.0	74.7
r/w RD	62.1	90.2	87.8	74.7
r/w gold labels	63.2	-	88.0	74.4
w/o cross-tuning	62.1	90.4	87.3	71.5
$RoBERTa_{base}$	64.8	90.9	90.7	79.4
+AD-DROP (GA)	66.8	91.4	**92.2**	**84.1**
r/w IGA	**68.1**	**91.6**	91.4	82.7
r/w AA	66.3	91.5	91.2	82.3
r/w RD	66.5	91.5	**92.2**	82.0
r/w gold labels	66.4	-	91.2	82.0
w/o cross-tuning	67.3	91.3	90.4	80.5

- Gradient-based attribution methods are better than others.
- IGA outperforms GA in some cases.
- AD-DROP improves the original models with any of the masking strategies.
- AD-DROP with pseudo labels for attribution is preferable.
- Removing cross-tuning causes noticeable performance degradation.

Outlines

Analysis

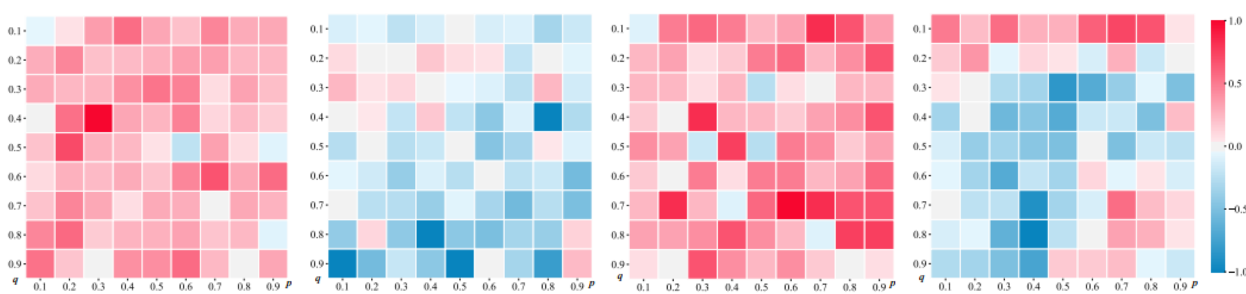
- ## Repeated Experiments

Table 3: Results of repeated experiments. Each score is the average of five runs with a standard deviation.

Methods	CoLA	STS-B	MRPC	RTE
BERT$_{\text{base}}$	61.8$_{\pm 1.9}$	89.4$_{\pm 0.5}$	85.2$_{\pm 1.3}$	71.2$_{\pm 1.2}$
+AD-DROP	**63.4**$_{\pm 0.4}$	**90.1**$_{\pm 0.5}$	**87.4**$_{\pm 0.9}$	**73.9**$_{\pm 1.1}$
RoBERTa$_{\text{base}}$	64.3$_{\pm 0.9}$	91.0$_{\pm 0.2}$	89.8$_{\pm 0.8}$	79.1$_{\pm 1.7}$
+AD-DROP	**66.4**$_{\pm 0.9}$	**91.2**$_{\pm 0.1}$	**91.3**$_{\pm 0.7}$	**82.5**$_{\pm 0.9}$

- AD-DROP achieves better performance with lower deviations.

- ## Hyperparameter Sensitivity



(a) BERT on CoLA (b) RoBERTa on CoLA (c) BERT on RTE (d) RoBERTa on RTE

Figure 6: Results of sensitivity study on CoLA and RTE. Rows correspond to p and columns refer to q. Blue blocks indicate the results of AD-DROP below the baseline (FT), and red blocks mean the results of AD-DROP above the baseline. Darker colors mean greater gaps with the baseline.

- RoBERTa with AD-DROP is more sensitive than BERT.

Analysis

- ## Few-shot Scenario

Table 5: Testing AD-DROP in few-shot settings. RoBERTa with AD-DROP achieves higher performance and lower deviations than that with the original fine-tuning approach.

Methods	SST-2 16-shot	SST-2 64-shot	SST-2 256-shot	CoLA 16-shot	CoLA 64-shot	CoLA 256-shot
$\text{RoBERTa}_{\text{base}}$	$74.50_{\pm 3.03}$	$89.06_{\pm 0.83}$	$91.44_{\pm 0.17}$	$23.18_{\pm 6.38}$	$39.70_{\pm 4.68}$	$51.11_{\pm 1.64}$
+AD-DROP	**$80.16_{\pm 1.51}$**	**$91.61_{\pm 0.52}$**	**$92.61_{\pm 0.13}$**	**$26.70_{\pm 4.96}$**	**$46.41_{\pm 1.98}$**	**$52.47_{\pm 1.16}$**

- ## Computational Efficiency

Table 7: Results of performance and computational cost of AD-DROP with different masking strategies (GA, IGA, AA, and RD) relative to the original fine-tuning. The symbol ‡ means AD-DROP is only applied in the first layer. BERT is chosen as the base model.

Methods	CoLA Mcc	CoLA Time	STS-B‡ Pcc	STS-B‡ Time	MRPC Acc	MRPC Time	RTE Acc	RTE Time
RD	+1.8	×1.42	+0.3	×1.38	+2.7	×1.31	+3.9	×1.42
AA	+3.3	×1.42	+0.1	×1.48	+2.9	×1.94	+3.9	×1.58
GA	+4.3	×3.58	+0.5	×1.95	+3.4	×4.13	+4.3	×4.50
IGA	+3.5	×99.61	+0.8	×15.00	+3.4	×110.12	+3.6	×125.67

- AD-DROP with GA is more competitive than others.

Outlines

1/ Introduction

2/ Methodology

3/ Experiment

4/ Results

5/ Conclusion

Conclusion

- We proposed AD-DROP to mitigate overfitting when finetuning PrLMs on downstream tasks. AD-DROP focuses on discarding high attribution attention positions to prevent the model from relying heavily on these positions to make predictions.

- We proposed a cross-tuning strategy that performs the original finetuning and our AD-DROP alternately to stabilize the finetuning process.

- Extensive experiments and analysis demonstrate the effectiveness of AD-DROP.

Thanks !