연세대학교
YONSEI UNIVERSITY

Less is More: Attention Supervision with Counterfactuals for Text Classification

Seungtaek Choi, Haeju Park, Jinyoung Yeo, Seung-won Hwang
Department of Computer Science, Yonsei University, Seoul, Korea
{hist0613, phj0225, jinyeo, seungwonh}@yonsei.ac.kr

Agenda

What is Attention?
Supervised Attention
Less-is-More with Our Approach (SANA) – **Causality**!
Experiments

What is Attention?

Attention is for improving NLP models as a **neural weighting** function: $\sum_i \alpha_i \cdot x_i$.

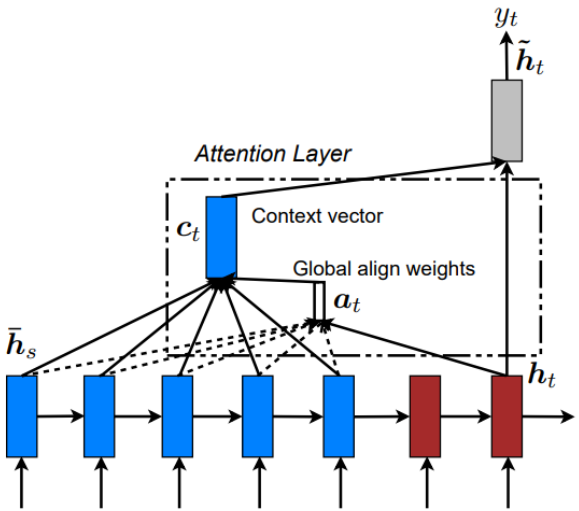


Figure 1: Attention mechanism [1]

What is Attention?

Attention is for enhancing **human understanding** of models.

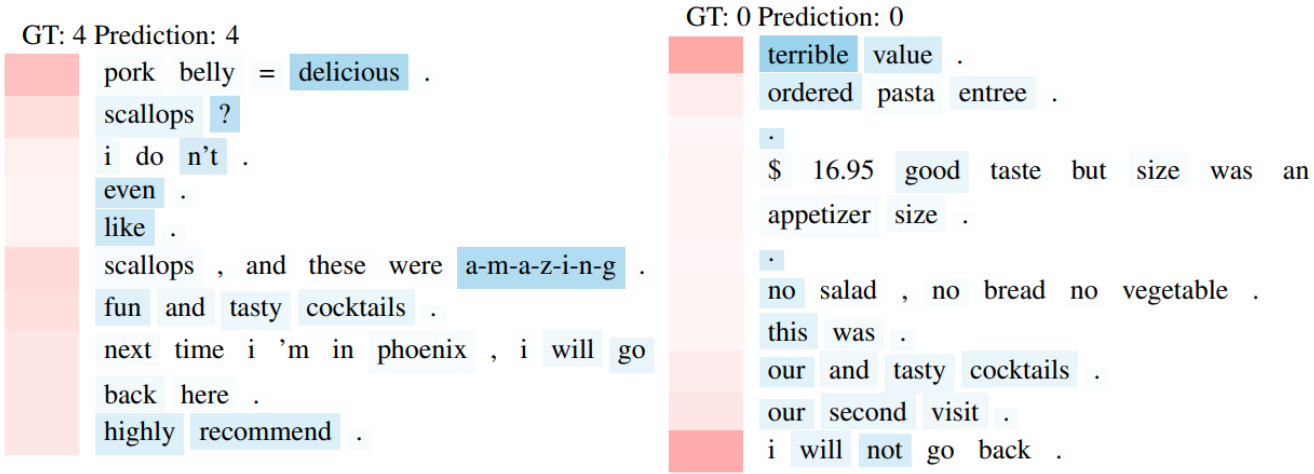


Figure 2: *Interpretable* document classification [2]

Let’s improve attention **with supervision!**

Supervised Attention

Attention can be supervised with **human attention**: $L_{att}(\hat{\alpha}, \alpha) = -\sum_{i=1}^{T} \alpha_i \log(\hat{\alpha}_i)$.

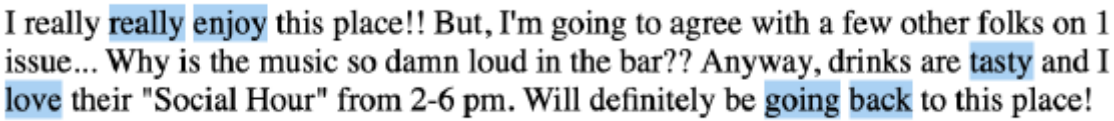
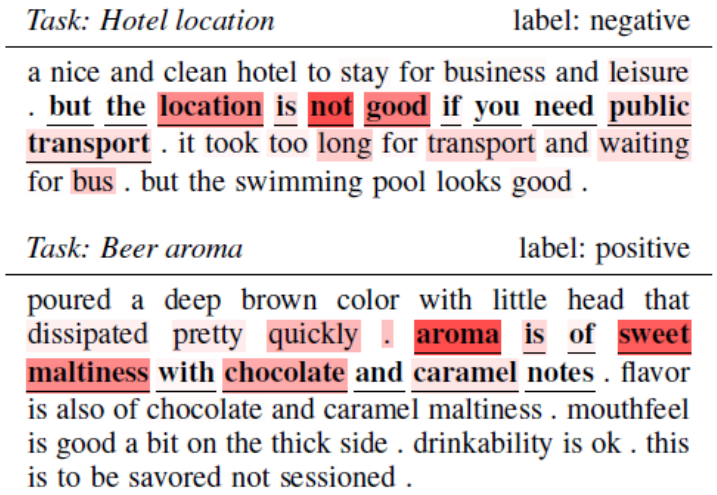


Figure 4: Human attention [4]

Supervised Attention

Attention (supervised with human rationales) improves accuracy and better selects **important words**.



Task: Hotel location label: negative

a nice and clean hotel to stay for business and leisure . **but the location is not good if you need public transport** . it took too long for transport and waiting for bus . but the swimming pool looks good .

Task: Beer aroma label: positive

poured a deep brown color with little head that dissipated pretty quickly . **aroma is of sweet maltiness with chocolate and caramel notes** . flavor is also of chocolate and caramel maltiness . mouthfeel is good a bit on the thick side . drinkability is ok . this is to be savored not sessioned .

Figure 5: Attention similar with **human rationales** [5]

However, human attention is **too expensive**!

Human annotator is required to highlight important words
specific to a sample and its class label.

“this place is small and crowded but the service is quick” (negative)

“this place is small and crowded but the service is quick” (positive)

Sample-level (rationale)

Alternative: Task-Level Supervision

An alternative with lower overhead is **annotating vocabulary**, rather than each sample, which is often publicly available as resources (e.g., *SentiWordNet*) or tools (e.g., *AllenNLP NER*).

"this place is small and crowded but the service is quick" (negative)

"this place is small and crowded but the service is quick" (positive)

Sample-level (rationale)

	Sample-level	Task-level	Reduction ratio
SST2	208K	16K	-92.3%
IMDB	5M	124K	-97.5%
20NG	232K	22K	-90.5%

Table 1: Comparison of annotation space

"this place is small and crowded but the service is quick" (negative/positive)

Task-level

Alternative: Task-Level Supervision

An alternative with lower overhead is **annotating vocabulary**, rather than each sample, which is often publicly available as resources (e.g., *SentiWordNet*) or tools (e.g., *AllenNLP NER*).

"this place is small and crowded but the service is quick" (negative)

"this place is small and crowded but the service is quick" (positive)

Sample-level (rationale)

	Sample-level	Task-level	Reduction ratio
SST2	208K	16K	-92.3%
IMDB	5M	124K	-97.5%
20NG	232K	22K	-90.5%

Table 1: Comparison of annotation space

"this place is small and crowded but the service is quick" (negative/positive)

Task-level

"*is sample-level rationale more effective than task-level supervision*?".

Less-is-More with Our Approach (SANA)

We propose Sample-level AttentioN Adaptation (SANA), to augment less human supervision with **counterfactual (machine) supervisions**.

1. Counterfactuals ($\hat{\alpha}$ vs. $\bar{\alpha}$) as causal signals ($\hat{y}$ vs. $\bar{y}$)
2. Adaptation of task-level annotation $\alpha_t \leftarrow \gamma \cdot \alpha_t$

What is Counterfactual?

A method of examining the **causality** in machine learning model.

A famous example in loan application:

1. M : “Seungtaek cannot receive the loan!”
2. M’: ”**If Seungtaek had a higher salary**, his loan application would have been accepted.”

What is Counterfactual?

A method of examining the **causality** in machine learning model.

A famous example in loan application:

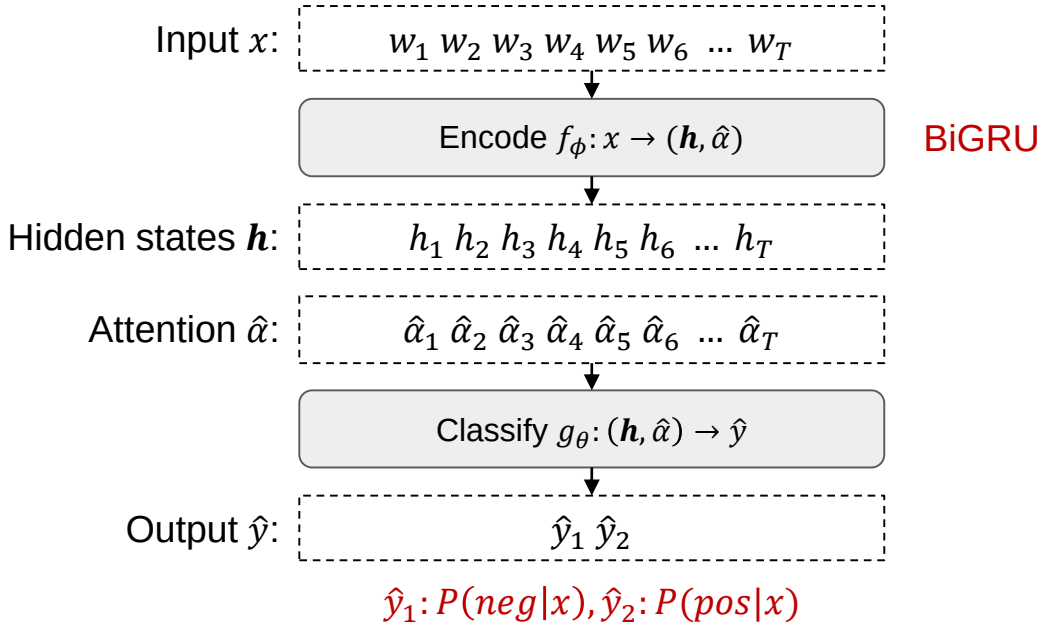
1. M : “Seungtaek cannot receive the loan!”
2. M’: ”**If Seungtaek had a higher salary**, his loan application would have been accepted.”

In our problem,

1. M: “*this place is **small** and crowded, but the service is quick*” = positive
2. M’: “*this place is **small** and crowded, but the service is quick*” = positive
3. “***small***” is not important!

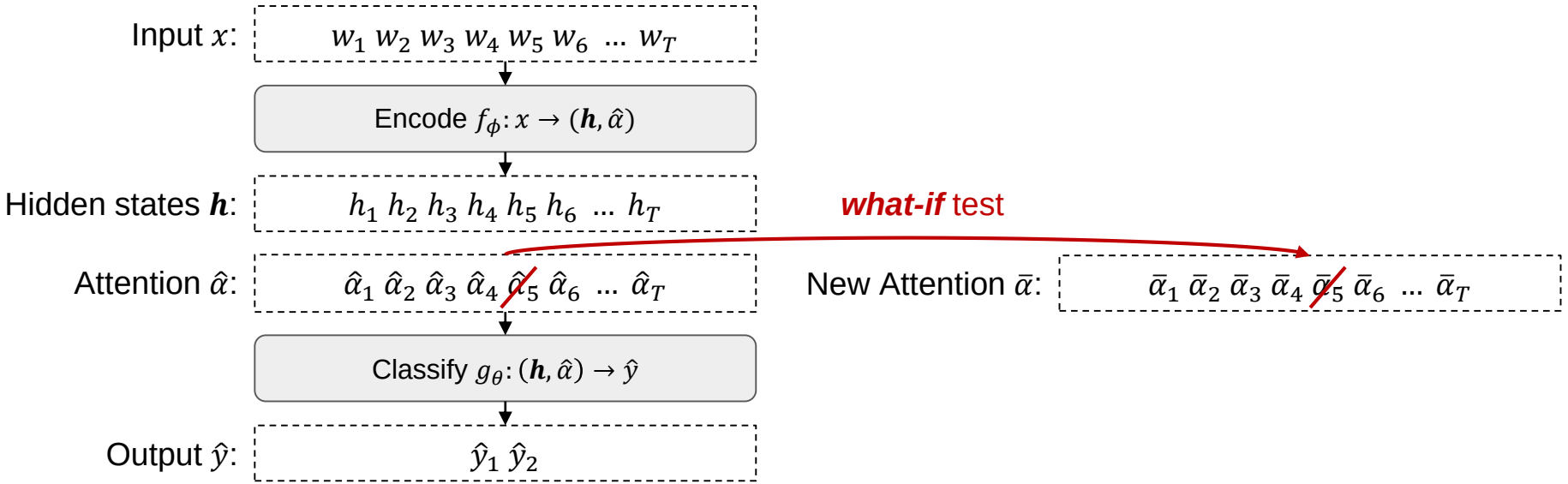
Less-is-More with Our Approach (SANA)

Typical process of text classification.



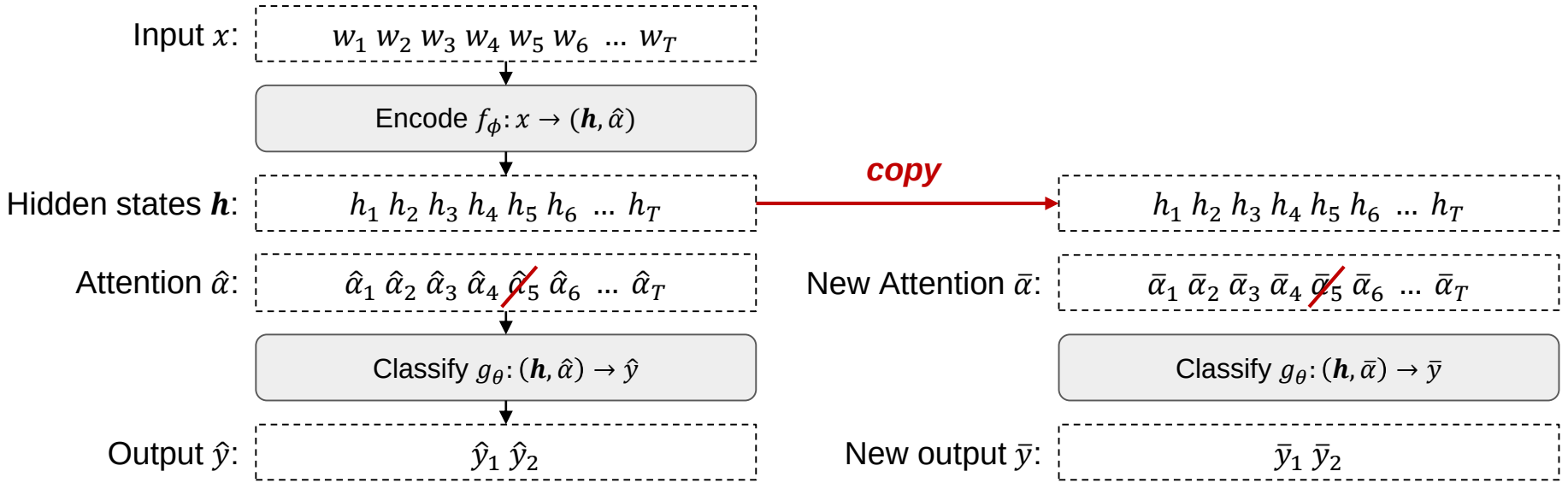
Less-is-More with Our Approach (SANA)

We obtain counterfactual attention by zeroing-out a word w_t.



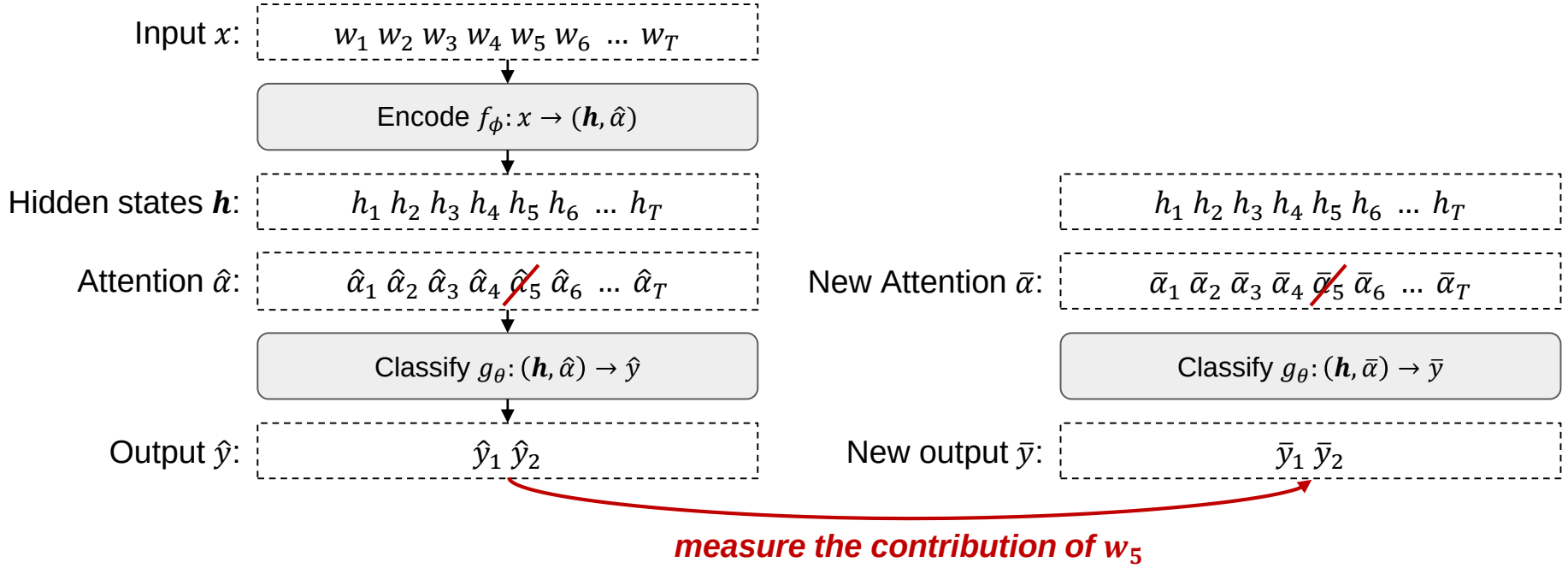
Less-is-More with Our Approach (SANA)

Then, we compute its corresponding prediction from modified attention by re-using $\boldsymbol{h}$.



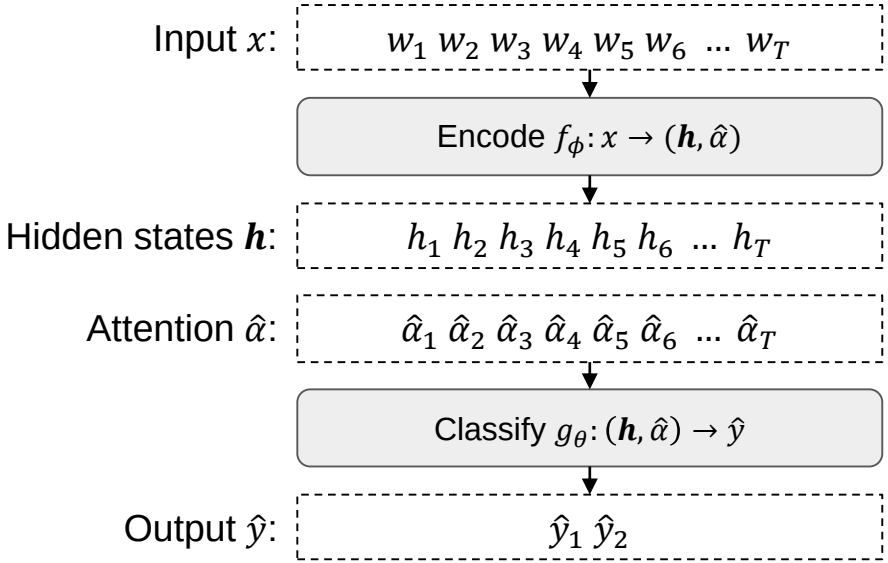
Less-is-More with Our Approach (SANA)

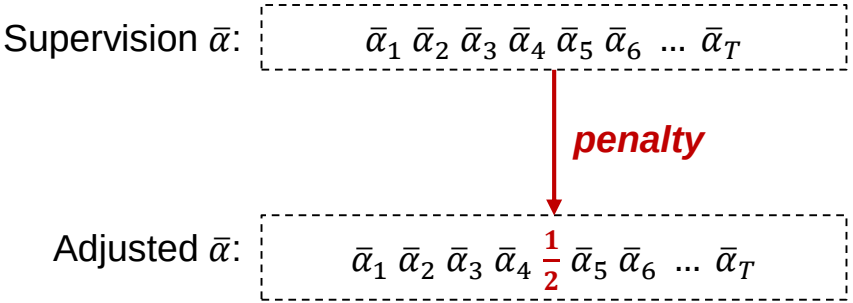
We measure how much the word w_t contributes to the original prediction via attention mechanism: $TVD(\hat{y}, \bar{y}_t) = \frac{1}{2}\sum_{c=1}^{C} |\hat{y}^c - \bar{y}_t^c|$[3].



Less-is-More with Our Approach (SANA)

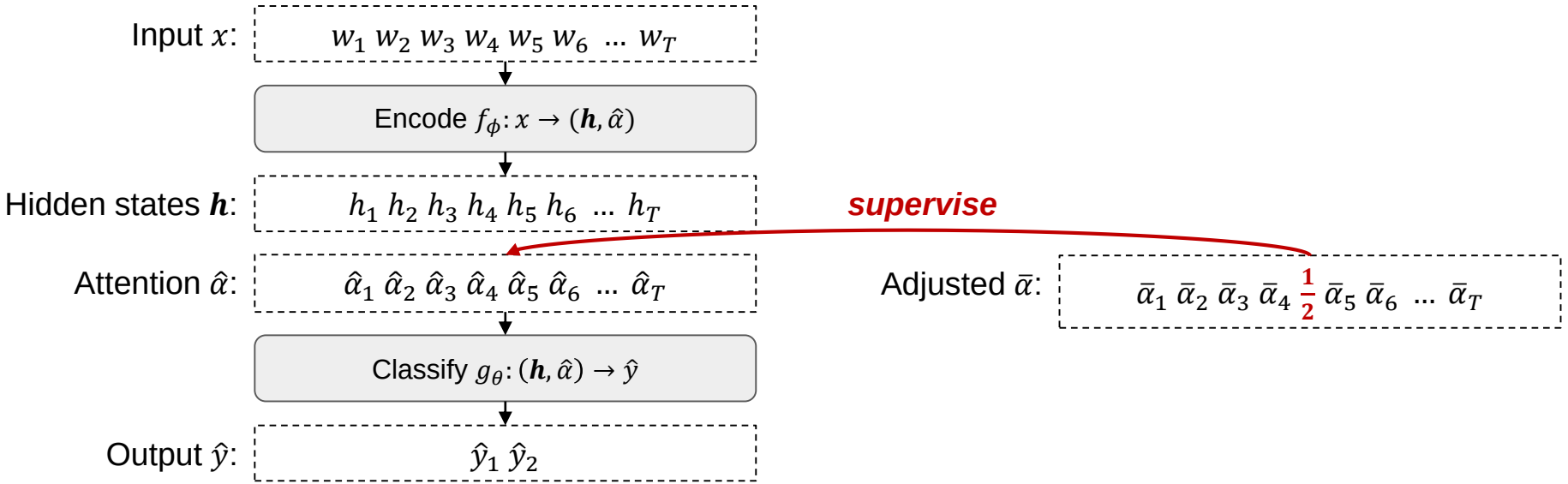
Then, we can give a penalty by decaying the supervision: $\alpha_t \leftarrow \gamma \cdot \alpha_t$,
where we set $\boldsymbol{\gamma = 0.5}$.





Less-is-More with Our Approach (SANA)

Finally, the network is re-trained with adjusted supervision.



Experiments: Dataset

Three text classification datasets, where two is sentiment analysis task and the other one is news categorization task, which are widely used and statistically diverse.

SST2

1) sentiment analysis (2 classes)
2) sentence (max_seq_len 30)
3) 11K samples

IMDB

1) sentiment analysis (2 classes)
2) document (max_seq_len 180)
3) 50K samples

20NG

1) news categorization (2 classes)
2) document (max_seq_len 300)
3) 1.1K samples

Experiments: Research Questions

We present the empirical findings for the following four research questions:

1. Does SANA improve model accuracy?
2. Does SANA improve model robustness?
3. Is SANA effective for data-scarce cases?
4. Does SANA improve attention explainability?

RQ1: Does SANA improve model accuracy?

1. SANA with task-level annotation outperforms all baselines in all the datasets.
2. The largest improvement is found in 20NG, which has the smallest training data.
3. SANA is effective even in model distillation setting.

	Accuracy		
	SST2	IMDB	20NG
BERT	91.67	94.10	93.25
unsupervised			
BiGRU	83.96	88.07	86.04
model distillation			
BiGRU	83.53	86.93	85.12
+ SANA	84.35	88.03	88.23
task-level annotation			
BiGRU	85.12	89.30	87.19
+ SANA	**85.72**	**90.10**	**89.13**

RQ1: Does SANA improve model accuracy?

1. SANA with task-level annotation outperforms all baselines in all the datasets.
2. The largest improvement is found in 20NG, which has the smallest training data.
3. SANA is effective even in model distillation setting.

	Accuracy		
	SST2	IMDB	20NG
BERT	91.67	94.10	93.25
unsupervised			
BiGRU	83.96	88.07	86.04
model distillation			
BiGRU	83.53	86.93	85.12
+ SANA	84.35	88.03	88.23
task-level annotation			
BiGRU	85.12	89.30	87.19
+ SANA	**85.72**	**90.10**	**89.13**

RQ1: Does SANA improve model accuracy?

1. SANA with task-level annotation outperforms all baselines in all the datasets.
2. The largest improvement is found in 20NG, which has the smallest training data.
3. SANA is effective even in model distillation setting.

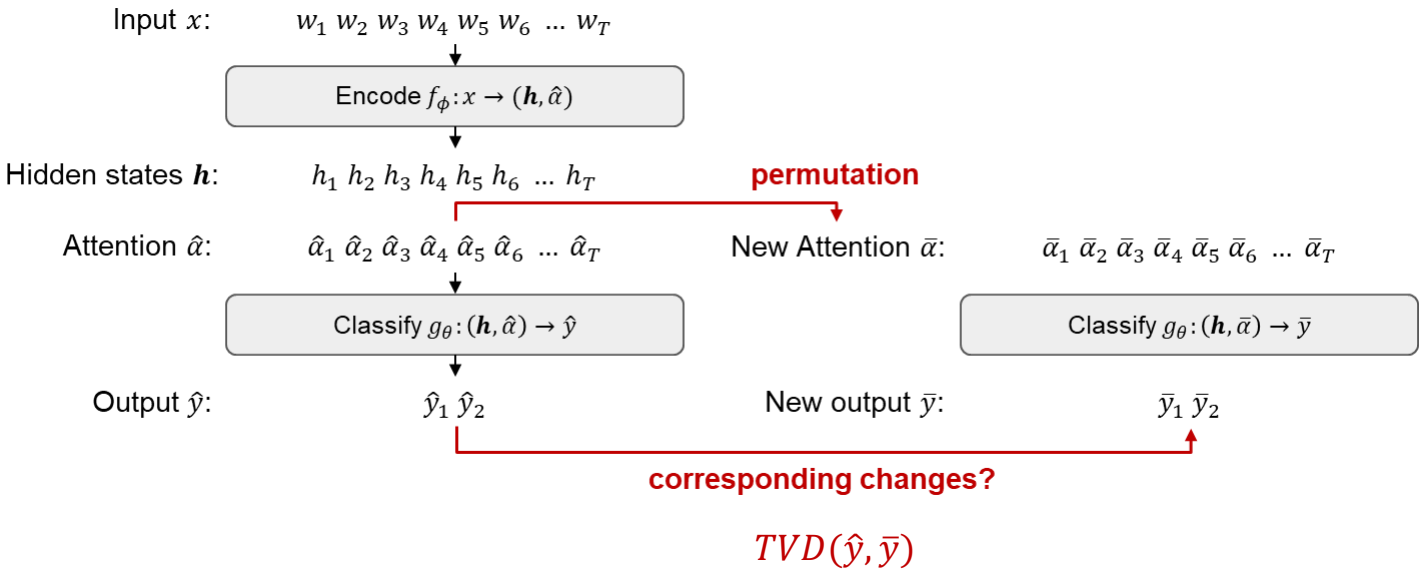
	Accuracy		
	SST2	IMDB	20NG
BERT	91.67	94.10	93.25
unsupervised			
BiGRU	83.96	88.07	86.04
model distillation			
BiGRU	83.53	86.93	85.12
+ SANA	84.35	88.03	88.23
task-level annotation			
BiGRU	85.12	89.30	87.19
+ SANA	**85.72**	**90.10**	**89.13**

RQ4: Does SANA improve attention explainability?

We measure whether attention correlates with class prediction, which we call **causal** explanation.

RQ4: Does SANA improve attention explainability?

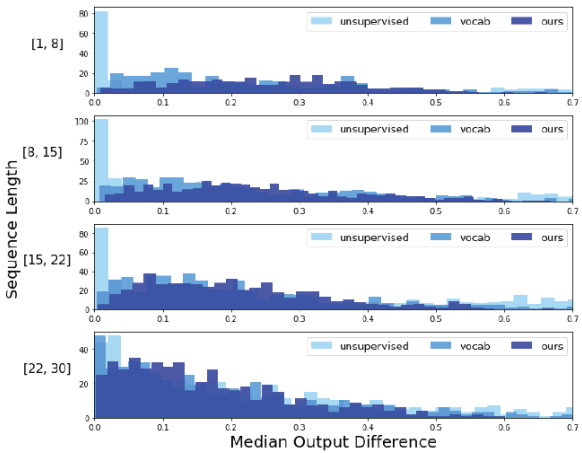
For causal explanation, [3] assumes that, if attention explains the machine decision, alternative attention weight ought to yield **corresponding changes** in prediction.



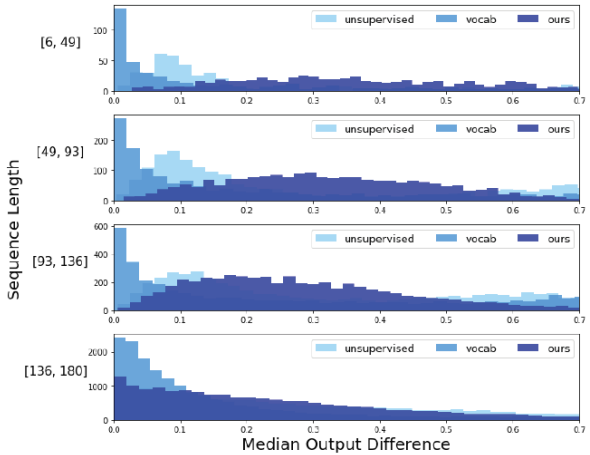
RQ4: Does SANA improve attention explainability?

x-axis: TVD values, i.e., the difference of model predictions

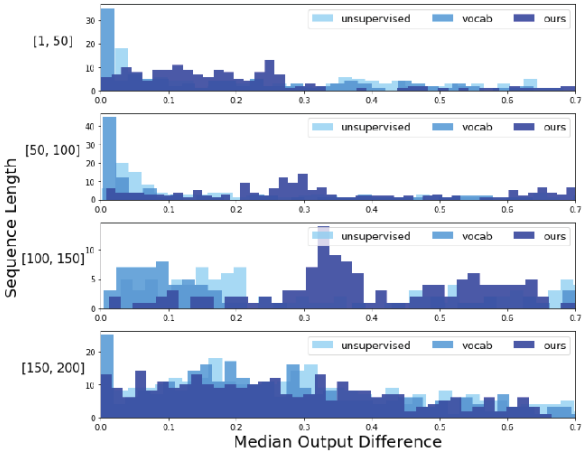
y-axis: the frequency of what-if simulations on their returning TVD value.



(a) SST2



(b) IMDB

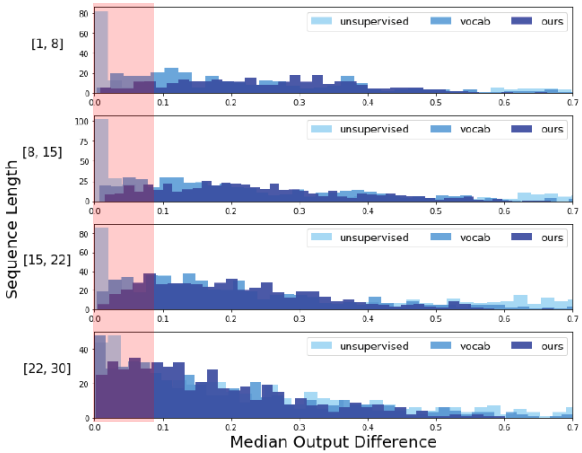


(c) 20NG

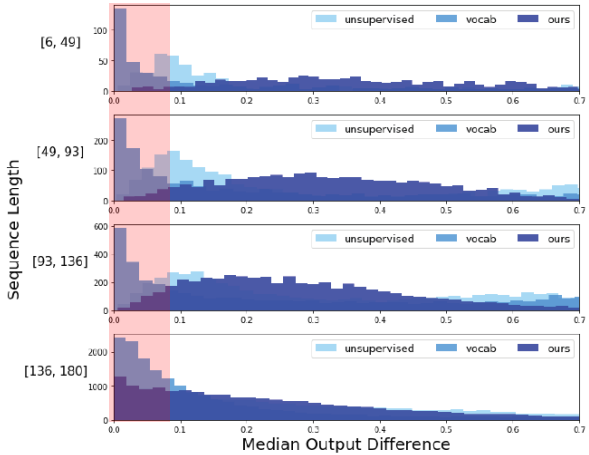
"If TVD is lower, the (original) learned attention has a weak mapping with the model prediction, and vice versa."

RQ4: Does SANA improve attention explainability?

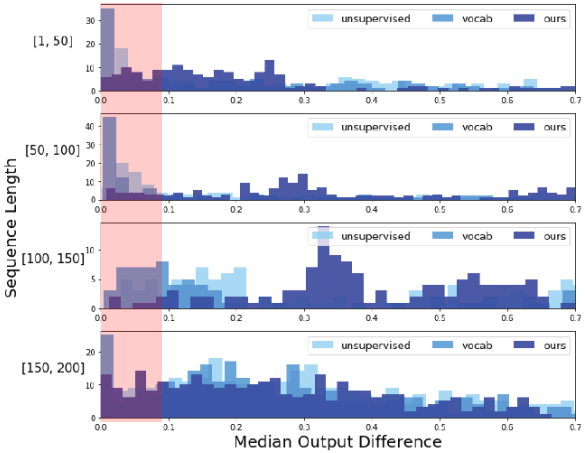
1. SANA has the lowest frequency on TVD=0 in all cases (right-skewed).
2. SANA even works well in long texts.



(a) SST2



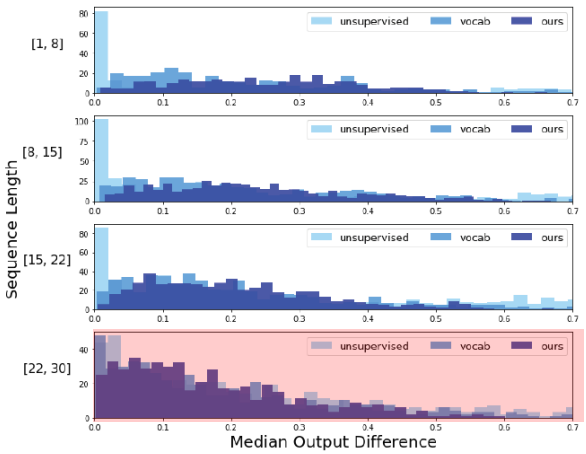
(b) IMDB



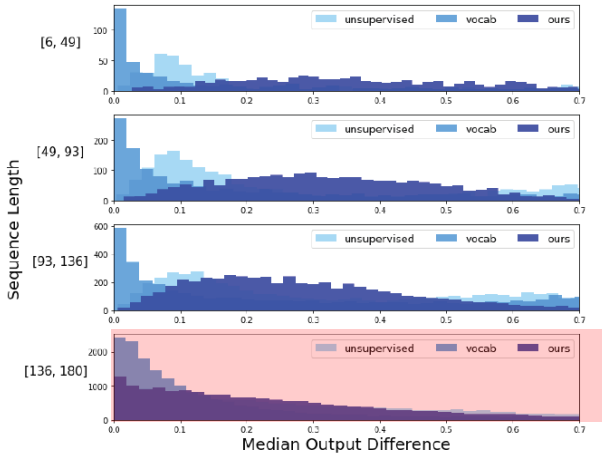
(c) 20NG

RQ4: Does SANA improve attention explainability?

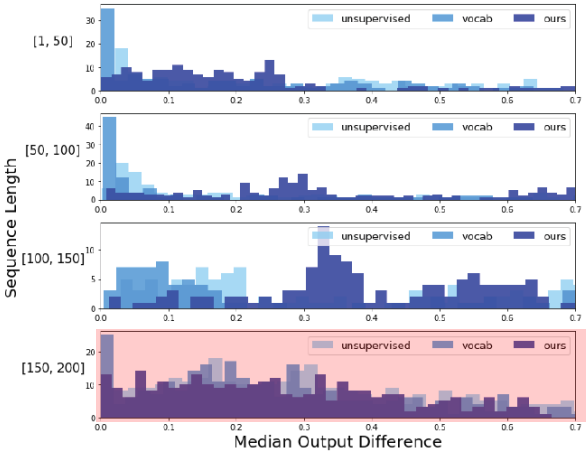
1. SANA has the lowest frequency on TVD=0 in all cases (right-skewed).
2. SANA even works well in long texts.



(a) SST2



(b) IMDB



(c) 20NG

Conclusion

We propose a counterfactual signal for self-supervision

1. to augment task-level human annotation
2. into sample-level machine attention supervision
3. to increase both the accuracy and explainability of the model.

Thanks!

Any question?