How does Attention Affect the Model?

Cheng Zhang[1],**Qiuchi Li**[2], **Lingyu Hua**[3], **Dawei Song** [3*]
[1]College of Intelligence and Computing, Tianjin University, China
[2]Department of Information Engineering, University of Padua, Italy
[3]School of Computer Science, Beijing Institute of Technology, China
cheng.zhang@tju.edu.cn
qiuchili@dei.unipd.it
hualingyu@hotmail.com
dwsong@bit.edu.cn

Abstract

The attention layer has become a prevalent component in improving the effectiveness of neural network models for NLP tasks. Figuring out why attention is effective and its interpretability has attracted a widespread deliberation. Current studies mostly investigate the effect of attention mechanism based on the attention distribution it generates with one single neural network structure. However they do not consider the changes in semantic capability of different components in the model due to the attention mechanism, which can vary across different network structures. In this paper, we propose a comprehensive analytical framework that exploits a convex hull representation of sequence semantics in an n-dimensional Semantic Euclidean Space and defines a series of indicators to capture the impact of attention on sequence semantics. Through a series of experiments on various NLP tasks and three representative recurrent units, we analyze why and how attention benefits the semantic capacity of different types of recurrent neural networks based on the indicators defined in the proposed framework.

1 Introduction and Motivation

The first appearance of the attention mechanism in natural language processing (NLP) can be traced back to its successful application in Neural Machine Translation (NMT). Bahdanau et al. (2014a) proposed an attention mechanism in an Encoder-Decoder model, which achieved great success, and showed that attention weight produced in this mechanism improved the interpretability of the model by providing a way of aligning the source and target languages through a simple quantitative analysis. Subsequently, the assumption that attention could improve the interpretability and transparency of a model was acquiesced by many later works, such as AEN (Song et al., 2019) (applied to targeted sentiment classification), ATAE-LSTM (Wang et al., 2016) (applied to aspect-level sentiment classification), and CMLA (Wang et al., 2017) (applied to semantic sentiment analysis).

More recently, this hypothetical premise has aroused controversies. For example, Serrano and Smith (2019) and Jain and Wallace (2019) used an erasure-based approach and advocated the attention weight does not necessarily correspond to importance. Wiegreffe and Pinter (2019) and Vashishth et al. (2019) considered attention to be interpretable, using a more model-driven approach and manual verification. These investigations focus on whether the attention distribution is unique and the correlation between attention weight and model prediction results, based on a similar neural network setting that consists of an embedding layer, a specific Recurrent Neural Network (RNN) and an attention component. Complex components such as the encoder-decoder structure were removed from the network as they may bias the analysis on the effect of attention weights. However, these works fail to explain two critical issues as follows:

(1) Neglecting the changes in the rest of the model before and after introducing attention, especially the word embedding layer and the RNNs' hidden layer. Figure 1 shows the transition before and after the introduction of attention. When a model does not introduce attention, the model generates an embedding sequence $\mathcal{E} = \{e_1, e_2, ..., e_n\}$ from the original one-hot word representation. Subsequently, the sequence $\mathcal{E}$ is processed by a specific RNN and converted into a hidden layer sequence $\mathcal{H} = \{h_1, h_2, ..., h_n\}$, which is used to produce the output. The introduction of attention will cause the model to change the gradient during the back-propagation in the training phase, which will lead to the embedding and hidden sequences to move away from the pre-

* Corresponding Author: Dawei Song

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 256–268

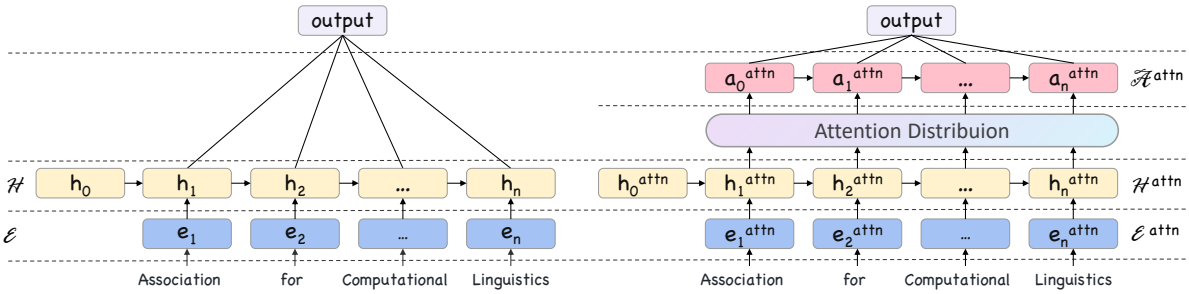


Figure 1: The transformation before and after the introduction of the attention in the RNNs.

vious values $\mathcal{E}$ and $\mathcal{H}$ when the model training step is done. The new embedding sequence is represented by $\mathcal{E}^{attn} = \{e_1^{attn}, e_2^{attn}, ..., e_n^{attn}\}$, and the hidden layer sequence is represented by $\mathcal{H}^{attn} = \{h_1^{attn}, h_2^{attn}, ..., h_n^{attn}\}$. After the attention layer, $\mathcal{H}^{attn}$ is adjusted by the attention distribution to produce a new sequence $\mathcal{A}^{attn} = \{a_1^{attn}, a_2^{attn}, ..., a_n^{attn}\}$. At this point, $\mathcal{A}^{attn}$ is used to generate the output of the model.

The existing works have focused on whether the attention distribution is unique or reasonable (if it is not unique). However, they ignore the extent of semantic changes in the sequences caused by the attention mechanism, including the changes in the embedding sequence ($\mathcal{E} \rightarrow \mathcal{E}^{attn}$), in the hidden layer sequence ($\mathcal{H} \rightarrow \mathcal{H}^{attn}$), and even in the emerging attention sequence ($\mathcal{A}^{attn}$). We posit that such ignorance would lead to an unfair and biased analysis of the attention.

(2) Lacking a systematic study of the attention effect on different types of RNNs. The attention layer is compatible with various types of RNNs, regardless of which recurrent unit out of Vanilla-RNN (Mikolov et al., 2010), LSTM (Hochreiter and Schmidhuber, 1997) or GRU (Cho et al., 2014), is used, or whether it has a uni-directional or bi-directional structure. Although the existing works have experimented on many datasets, they solely focus on a single type of RNN preceding the attention layer at a time. We argue that a comprehensive comparison of the changes before and after introducing attention to different types of RNNs mentioned above will better reveal the intrinsic interpretability of attention.

To address these two issues, we propose to explore the effect of attention from a new perspective by conducting a systematic investigation on the semantic changes across different sequences of a RNN model with or without attention, and comparing the differences in the changes across mainstream recurrent units. Based on the analysis results, we expect to better understand what happens before and after the introduction of attention into the model.

The proposed analysis requires a comprehensive framework with reasonable metrics to evaluate the quality of sequence semantics based on their vector representations. For this purpose, we adopt the concept of Convex Hull in n-dimensional Semantic Euclidean Space ($\mathbb{SR}^n$) (Zhang et al., 2020) to represent the semantics of a sequence. Since an attention mechanism always produces a point in the convex hull of its preceding hidden units, we can establish suitable metrics based on the convex hull formed by a sequence of vectors in $\mathbb{SR}^n$, to facilitate the analysis of attention effect. Section 2 will briefly introduce the Semantic Euclidean Space and the convex hull representation of the sequence semantics. In Section 3, we analyze the attention mechanism and establish the relationship between the attention weight and the sequence meaning (as convex hull in $\mathbb{SR}^n$). Section 4 formulates a set of indicators to analyze the semantic changes before and after attention. With the proposed framework, we conduct comparative experiments on various datasets concerning text classification and sentiment analysis tasks in Section 5. Based on the experimental results, we conduct in-depth analysis from the perspective of why and how the attention mechanism benefits the semantic capacity of different recurrent units of RNNs.

2 Background

Zhang et al. (2020) proposed an n-dimensional Semantic Euclidean Space ($\mathbb{SR}^n$), which defines the mapping relationship between points in an Euclidean Space ($\mathbb{R}^n$) and their semantics. As a se-

mantic extension of $\mathbb{R}^n$, $\mathbb{SR}^n$ is defined as:

$$\mathbb{SR}^n = \{\forall x = (x_1, ..., x_n) \in \mathbb{R}^n | x \rightarrow \text{semantics}\} \quad (1)$$

In $\mathbb{SR}^n$, the points are divided into *specific semantic points* and *abstract semantic points*. A specific semantic point has a word corresponding to it, which can also be regarded as a word embedding. An abstract semantic point does not have a specific word corresponding to it, such as a point generated by the hidden layer of RNNs.

Zhang et al. (2020) then proposed to use the convex hull and centroid of points in $\mathbb{SR}^n$ to measure the **meaning** and **central idea** of a sequence of words. It provides a theoretical basis for exploring the semantic changes that occur before and after introducing attention into a model.

2.1 Meaning of a Sequence

Definition. *The meaning of a sequence composed of semantic points is represented by the convex hull composed of these points.*

Given a sequence $\mathcal{X}$ composed of semantic points, its meaning, denote as **ME**$(\mathcal{X})$, is formulated as:

$$\text{ME}(\mathcal{X}) = \text{Conv}(\mathcal{X}) \quad (2)$$

$\text{Conv}(\mathcal{X})$ denotes the convex hull of a finite point set $\mathcal{X}$, as the set of all convex combinations of the points (Faux and Pratt, 1979). In a convex combination, each point x_i in $\mathcal{X}$ is assigned with a weight or coefficient α_i in such a way that the coefficients are all non-negative and sum to one. These weights are used to produce a weighted average of points. It is formulated as:

$$\text{Conv}(\mathcal{X}) = \left\{ \sum_{i=1}^{|\mathcal{X}|} \alpha_i x_i \middle| \alpha_i \geq 0 \wedge \sum_{i=1}^{|\mathcal{X}|} \alpha_i = 1 \right\} \quad (3)$$

The mapping between the definition of the convex hull and the meaning of a sequence is intuitive. A sentence (sequence) consists of words (semantic points). In addition to the semantics expressed by the individual words, a sentence should also include the implicit semantics (abstract semantic points) produced by all possible combinations of words.

2.2 Central Idea of a Sequence

Definition. *The central idea of a sequence composed of semantic points is represented by the centroid of the sequence's meaning.*

The central idea of a sequence $\mathcal{X}$ of semantic points is denoted as $\text{Centroid}(\mathcal{X})$, formulated as:

$$\text{CI}(\mathcal{X}) = \text{Centroid}(\text{ME}(\mathcal{X})) \quad (4)$$

The centroid a subset $\mathcal{X}$ of $\mathbb{R}^n$ is the mean position of all the points in all coordinate directions. It is computed as:

$$\text{Centroid}(\mathcal{X}) = \frac{\int x g(x) dx}{\int g(x) dx} \quad (5)$$

where the integrals are taken over the whole space $\mathbb{R}^n$, and g is the characteristic function, which is 1 if a point is inside $\mathcal{X}$ and 0 otherwise (Protter and Morrey, 1977).

However, the central idea of a sequence needs to be calculated as $\text{Centroid}(\text{Conv}(\mathcal{X}))$, instead of $\text{Centroid}(\mathcal{X})$ directly, to guarantee that the central idea of the sequence lies within the convex hull (meaning) of the sequence. In contrast, even though the geometric centroid of a convex object always lies within the area representing its meaning, a non-convex object might also have a centroid that is outside the area, which is undesirable. As introduced above, ME scopes the meaning of a sequence as an area in $\mathbb{SR}^n$, while the central idea of the sequence should be at the centre of the ME area.

The central idea of a sequence can be considered as a "summary" of the sentence's meaning. Operationally, it is the centroid of the convex hull representation of the sentence meaning, within a $\mathbb{SR}^n$. Take the phrase "The Association for Computational Linguistics" as an example, the central idea of this phrase can be summarized as a semantic point, which corresponds to the abbreviation "ACL". Considering another phrase, "good enough but not excellent", the central idea of this phrase also can be summarized as a semantic point, but for the time being, there is no word that corresponds to this semantic point. Perhaps with the development of natural language, people will soon create a word to describe this semantic point. This is actually the specific semantic point and abstract semantic point defined in $\mathbb{SR}^n$. More explanations about $\mathbb{SR}^n$ can be found in Zhang et al..

3 Attention With Convex Hull

Motivated by the ability of $\mathbb{SR}^n$ to measure the meaning and central idea of sequences, we will theoretically analyze the role of attention from the perspective of semantic change.

Take the well-known Scaled Dot-Product attention (Vaswani et al., 2017) as an example (this will be abbreviated as **dot-attn** later). When a sequence $\mathcal{X} = \{x_1, x_2, ..., x_n\}$ passes through a dot-attn layer, the specific calculation process is shown as follows:

$$\alpha_i = softmax(\frac{x_i X}{\sqrt{m}}) \tag{6}$$

$$y_i = \alpha_i X \tag{7}$$

$X \in \mathbb{R}^{n \times m}$ denotes the matrix of word vectors corresponding to the input sequence $\mathcal{X}$, where m is the dimensionality of word vectors. Essentially this process can be described as the following two steps:

1. Construct an attention distribution $\alpha = (\alpha_1, \alpha_2, ..., \alpha_n)$ through the input sequence $\mathcal{X}$ and the softmax function,

2. Use the attention weight and $\mathcal{X}$ to generate a new sequence $\mathcal{Y} = \{y_1, y_2, ..., y_n\}$, which is called an attention sequence.

α_i is a probability distribution generated from the softmax function, which ensures that each component in it will be in the interval $(0, 1)$ and the components will add up to 1. Focusing on a specific vector y_i in $\mathcal{Y}$, it can be expressed as follows:

$$y_i = \sum_{j=1}^{n} \alpha_i^j x_j \Big| \sum_{j=1}^{n} \alpha_i^j = 1, \alpha_i^j \geq 0 \tag{8}$$

Comparing Formula 8 with Formula 3, we can find that under the action of dot-attn, the process of converting $\mathcal{X}$ to $\mathcal{Y}$ is indeed a process of continuously selecting new semantic points from the convex hull of $\mathcal{X}$ (i.e., the meaning of $\mathcal{X}$, ME($\mathcal{X}$)) to form a new semantic sequence $\mathcal{Y}$. Therefore, the new sequence $\mathcal{Y}$ is a semantic transformation of the original sequence $\mathcal{X}$ to some extent. An example of $\mathcal{X} \xrightarrow{\text{dot-attn}} \mathcal{Y}$ is shown in Figure 2.

Furthermore, although the convex hull of $\mathcal{X}$ can be used to express the meaning of a sentence, the model usually uses $\mathcal{X}$ to construct a vector $c = \frac{1}{n}\sum_{i=1}^{n} x_i$ as representation of a sentence, followed by a dense layer and activation function to generate prediction result. The form of sentence representation c is consistent with the definition of the central idea in $\mathbb{SR}^n$. Therefore, from the sentence representation's perspective, attention is to transform the central idea expressed by the original sequence (CI($\mathcal{X}$)) to a new semantic point CI($\mathcal{Y}$) (the central idea of $\mathcal{Y}$). The offset from the yellow diamond to the red triangle in Figure 2 represents the conversion from CI($\mathcal{X}$) to CI($\mathcal{Y}$).

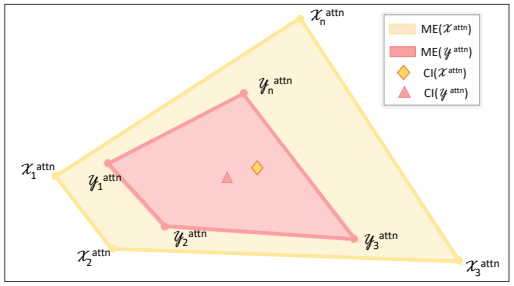


Figure 2: Use attention to convert sequence $\mathcal{X}$ to sequence $\mathcal{Y}$. The meaning of $\mathcal{X}$ is represented by the yellow shaded part, and the meaning of $\mathcal{Y}$ is represented by the red part.

In summary, the model has undergone the following changes after the introduction of attention:

1. Attention adjusts the meaning expressed by the original sequence by adjusting each semantic point in the sequence.

2. Attention changes the central idea (an instance representation) of the original sequence.

Through the above analysis, we have a deeper understanding of how attention transforms the original input from the perspective of the convex hull at the theoretical level. It is important to note that our proposed attention analysis framework above is applicable to other forms of attention, such as tanh attention (Zhou et al., 2016). No matter a popular dot-attn or a traditional tanh attention is used, they can be regarded as firstly adjusting the sequence $\mathcal{X}$ to sequence $\mathcal{Y}$, and then further averaging them to make predictions. The subtle difference between them lies in the dimensions of attention distributions. For an input sequence, the dot-attn generates a 2-dimensional distribution, while the tanh attention generates a 1-dimensional distribution.

However, only the above analysis framework is not enough. During the training process, due to the introduction of attention, the gradient change in the

back-propagation process will cause the original sequence $\mathcal{X}$ to be converted into a new sequence $\mathcal{X}^{attn}$, and this has also been ignored in previous studies. To this end, we will first construct relevant evaluation indicators, and then give our complete analysis framework based on the theoretical analysis of attention in this section.

4 Assessing the Effect of Attention

Following the typical settings in this area, we use an RNN model as the basis to systematically analyze the changes between different sequences in the process of introducing attention to the model. We specify multiple indicators to measure these changes and accordingly present our analysis framework.

4.1 Various Sequences Before and After the Introduction of Attention

As shown in Figure 3, the model takes the one-hot representation of the word sequence as the initial input. When attention is not used by the model, the model contains an embedding sequence $\mathcal{E} = \{e_1, e_2, ..., e_n\}$ inferred from a dense layer and a hidden sequence $\mathcal{H} = \{h_1, h_2, ..., h_n\}$ produced by an RNN. When an attention mechanism is introduced, the model weight changes due to the gradient changes during the training process. Hence, $\mathcal{E}$ and $\mathcal{H}$ will be converted to new sequences $\mathcal{E}^{attn} = \{e_1^{attn}, e_2^{attn}, ..., e_n^{attn}\}$ and $\mathcal{H}^{attn} = \{h_1^{attn}, h_2^{attn}, ..., h_n^{attn}\}$. The introduction of attention will be further transformed $\mathcal{H}^{attn}$ into an attention sequence $\mathcal{A}^{attn} = \{a_1^{attn}, a_2^{attn}, ..., a_n^{attn}\}$. For the above five sequences, there is the following progressive relationship:

$$\mathcal{E} \rightarrow \mathcal{H} \tag{9}$$

$$\mathcal{E}^{attn} \rightarrow \mathcal{H}^{attn} \rightarrow \mathcal{A}^{attn} \tag{10}$$

The differences between the above two links of sequences should be carefully examined to explore the impact of attention on the model. In this work, we first define a series of indicators to measure the semantic expression ability of an independent sequence and the semantic relationship between two sequences that belong to the same link. A framework is proposed to systematically compare the differences between the two links to assess the effect of attention.

4.2 Degree of Semantic Unsaturation

In $\mathbb{SR}^n$, the meaning of a sequence $\mathcal{X}$ of length $|\mathcal{X}|$ is calculated by ME. We define the degree of semantic unsaturation of a sequence as follow:

$$\text{DSU}(\mathcal{X}) = \frac{\text{ME}(\mathcal{X})}{|\mathcal{X}|} \tag{11}$$

$\text{DSU}(\mathcal{X})$ reflects the degree of semantic unsaturation regarding $\mathcal{X}$. Normally, the smaller the semantic space contained in the meaning of a sequence, the more precise the semantics expressed. Specifically, for sequences $\mathcal{X}$ and $\mathcal{Y}$ have same sequence length, if the meaning expressed by $\mathcal{X}$ is more precise than the meaning expressed by $\mathcal{Y}$, i.e. $\text{ME}(\mathcal{X}) < \text{ME}(\mathcal{Y})$, then $\text{DSU}(\mathcal{X})$ is less than $\text{DSU}(\mathcal{Y})$, which means that the degree of unsaturation of $\mathcal{X}$ is lower than $\mathcal{Y}$. For this reason, the smaller $\text{DSU}(\mathcal{X})$, the better.

4.3 Semantic Coverage

For two sequences $\mathcal{X}, \mathcal{Y}$, the sequence $\mathcal{Y}$ is a semantic transformation of the previous sequence $\mathcal{X}$ (this transformation may be synonymous transformation or even semantic extraction), we use semantic coverage (SC) (Zhang et al., 2020) to indicate the overlap between two sequences:

$$\text{SC}(\mathcal{X}, \mathcal{Y}) = \text{ME}(\mathcal{X}) \cap \text{ME}(\mathcal{Y})$$

Since $\mathcal{X}$ is the original sequence and $\mathcal{Y}$ is the converted sequence, then three indicators Semantic Coverage Precision (SCP), Semantic Coverage Recall (SCR), and Semantic Coverage F-Measure (SCF) can be naturally defined to observe the changes between the two sequences:

$$\text{SCP}(\mathcal{X}, \mathcal{Y}) = \frac{\text{SC}(\mathcal{X}, \mathcal{Y})}{\text{ME}(\mathcal{Y})} \tag{12}$$

$$\text{SCR}(\mathcal{X}, \mathcal{Y}) = \frac{\text{SC}(\mathcal{X}, \mathcal{Y})}{\text{ME}(\mathcal{X})} \tag{13}$$

$$\text{SCF}(\mathcal{X}, \mathcal{Y}) = \frac{2 \times \text{SCP} \times \text{SCR}}{\text{SCP} + \text{SCR}} \tag{14}$$

4.4 Central Idea Offset

In addition to the difference in meaning between two sequences, it is crucial to check the deviation of the central idea between the two sequences $\mathcal{X}, \mathcal{Y}$. The offset distance between the central idea of $\mathcal{Y}$ and that of the original sequence $\mathcal{X}$ is called Central Idea Offset (CIO), formulated as follows:

$$\text{CIO}(\mathcal{X}, \mathcal{Y}) = \|\text{CI}(\mathcal{X}), \text{CI}(\mathcal{Y})\| \tag{15}$$

4.5 Analysis Framework

Base on the definition of the above five indicators, we propose a framework to analyze the impact of

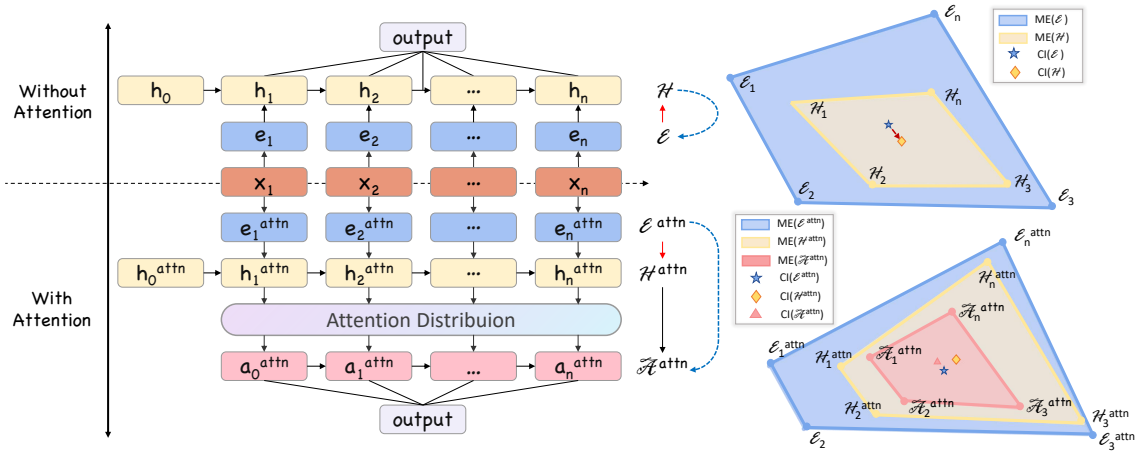


Figure 3: Splice the two models before and after the attention is introduced in a symmetrical manner. The word vector sequence and hidden layer sequence linked by the red arrow in the figure represent the observation from the corresponding perspective. The sequence linked at both ends of the blue dashed arrow represents observation from a shift perspective.

attention on a certain model. Recalling the two links in Eq. 9, since the number of sequences contained in each link is different, we propose two perspectives for comparison: the corresponding perspective and the shift perspective.

4.5.1 The Corresponding Perspective

The introduction of attention to a model has caused changes in its embedding sequence ($\mathcal{E} \rightarrow \mathcal{E}^{attn}$) and its hidden layer sequence ($\mathcal{H} \rightarrow \mathcal{H}^{attn}$). This observation on the changes in the corresponding layers of the model is called the corresponding perspective. Using Δ to represent the difference, $\Delta(\rho(\mathcal{E}), \rho(\mathcal{E}^{attn}))$ reflects the influence of the introduction of attention on the embedding layer from the corresponding perspective, and similarly for $\mathcal{H}$ and $\mathcal{H}^{attn}$.

In addition to the changes on a single sequence, the difference between the links between adjacent sequences ($\mathcal{E} \rightarrow \mathcal{H}$, $\mathcal{E}^{attn} \rightarrow \mathcal{H}^{attn}$) can also be used to observe the impact of attention. For example, $\Delta(\text{SCP}(\mathcal{E}, \mathcal{H}), \text{SCP}(\mathcal{E}^{attn}, \mathcal{H}^{attn}))$ is employed to compare the changes of semantic coverage precision between embedding layer and hidden layer before and after introduction of attention. This difference can also be computed for SCP, SCF and CIO.

4.5.2 The Shift Perspective

According to the analysis in Section 3, before attention is added, the generated attention sequence $\mathcal{H}$ is actually a conversion of the embedding layer. In the presence of attention, the embedding sequence $\mathcal{E}^{attn}$ is converted to the sequence $\mathcal{H}^{attn}$, which is further transformed to $\mathcal{A}^{attn}$ by the attention mechanism. Therefore, the difference between $\text{CIO}(\mathcal{E}, \mathcal{H})$ and $\text{CIO}(\mathcal{E}^{attn}, \mathcal{A}^{attn})$ can be used to reflect the influence of attention on the overall semantic shift along the link.

In the mean time, an input sentence is finally converted into $\mathcal{H}$ or $\mathcal{A}^{attn}$ to express the meaning of the sentence, so we can alternatively express this change by $\Delta(\text{DSU}(\mathcal{H}), \text{DSU}(\mathcal{A}^{attn}))$.

Dataset	Train/Valid/Test Size	Vocab Size	Label Size
AG_News	96000/24000/7600	95812	4
SST	8544/1101/2210	16583	4

Table 1: Dataset statistics.

5 Exploring the Attention

We first introduce the dataset and models used in the experiments and then explore the impact of attention using the analysis framework above.

5.1 Experimental Setup

In order to make our analysis concise, our experiments focused on both text classification task (Stanford Sentiment Treebank (SST) (Socher et al., 2013)) and sentiment analysis task (AGNews [1]). In the future, we will extend our work to more data

[1] http://groups.di.unipi.it/~gulli/AG_corpus_of_news_articles.html

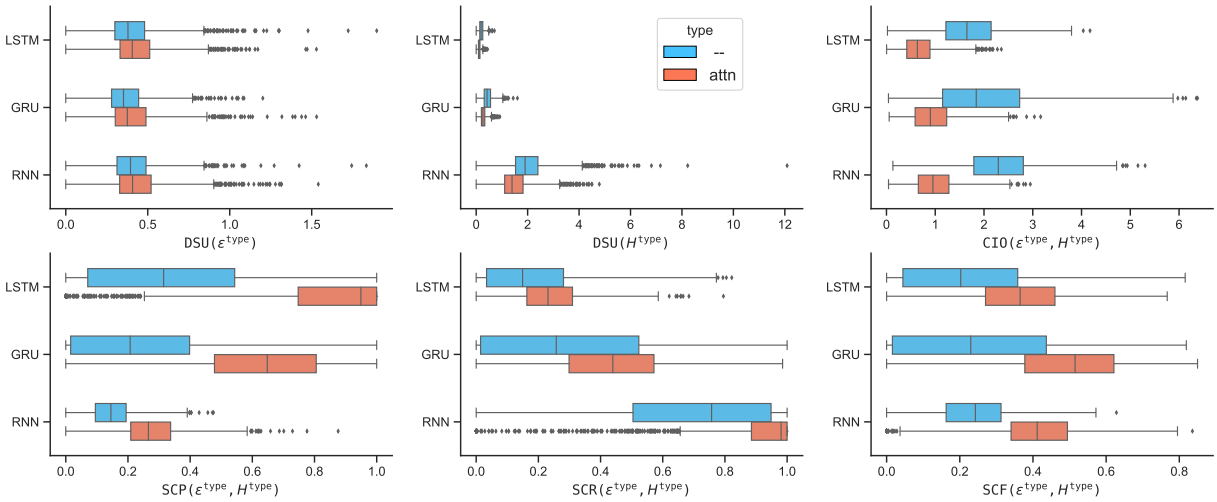


Figure 4: From the corresponding perspective observe the impact of the introduction of attention on the model. In each bin, top (blue) is the model without attention, bottom (red) is the model with attention.

	AG_News		SST	
	-	dot	-	dot
LSTM	0.969	0.976	0.918	0.946
Bi-LSTM	0.974	0.977	0.941	0.951
GRU	0.966	0.973	0.938	0.946
Bi-GRU	0.968	0.970	0.945	0.947
RNN	0.953	0.967	0.848	0.948
Bi-RNN	0.971	0.972	0.892	0.948

Table 2: The accuracy of different types of recurrent units before and after the introduction of attention.

sets, especially the machine translation dataset in the encoder-decoder model.

Since the AGNews dataset does not have a predefined validation set, the training set is split into a training set and validation set at a ratio of 8:2. The statistics of datasets are shown in Table 1. For each dataset, the base model we used for training is shown in Figure 1. It has an embedding layer for convert one-hot representation to distribution representation, a specific RNN-layer (recurrent units can be Vanilla-RNN, LSTM or GRU. The overall structure can be uni-directional or bi-directional, resulting in 6 different combinations.), without or with a dot-attn layer, followed by an additive layer and softmax prediction. The accuracy results of these models on the validation set are shown in Table 2, we calculate and analyze our indicators on the test set (The distribution of sentence length in the test set is shown in Appendix A). It should be pointed out that for the problem that the convex hull of high-dimensional vectors cannot be calculated temporarily, we use t-SNE to reduce the collected vectors to 2-dimensional at first (like the work of Zhang et al.), and further continue calculate the convex hull of the sequence, and use the area to represent the semantic size covered by the convex hull (The reproducibility is shown in Appendix B).

Both in the datasets, regardless of the uni-directional or bi-directional RNNs structure is used, the experimental results' trends are similar. Therefore, we only show the results generate by the use of bi-directional RNNs to compare different recurrent units before and after the introduction of attention on the dataset SST. The more experimental results can found in Appendix C, such as uni-directional RNNs in SST dataset, uni/bi-directional RNNs on AGNews.

5.2 Analysis from the Corresponding Perspective

As shown in Figure 4, by observing the difference in semantic density from the corresponding perspective. We can find that the embedding sequences ($\mathcal{E}, \mathcal{E}^{\text{attn}}$) learned by the model is basically the same for different recurrent types with or without attention. However, from the hidden layer sequence, $\mathcal{H}, \mathcal{H}^{attn}$, we can observe the change of this difference, no matter what type of recurrent unit, $\mathcal{H}^{\text{attn}}$ are less than $\mathcal{H}$, this shows that the hidden layer sequence of RNNs using the attention escapes or abstracts the original text with a smaller semantic range, and the semantics expressed are more accurate. After using the attention, the hidden layer sequence improves the accuracy of semantic

expression (decrease the degree of semantic unsaturation) also brings about the improvement of the model effect, as shown in Table 2. The attention mechanism's introduction also led to the shortening of the central idea offset between embedding sequences and hidden sequences.

From the semantic coverage perspective, the introduction of attention makes the semantic conversion between the embedding layer and the hidden layer generally improve in all of the three indicators, semantic coverage recall, semantic coverage accuracy, and semantic coverage F-Measure. This improvement shows that the semantic closeness between embedding and hidden layer and the degree of unsaturation of the hidden layer greatly influence the model results. The introduction of attention improve the accuracy of hidden layer sequence expression semantics and makes the semantic conversion between the embedding sequence and the hidden layer sequence more natural. If we compare different types of recurrent units, it is not difficult to find that RNNs is significantly worse than LSTM and GRU on most indicators, which shows that the prediction accuracy of RNNs is lower than LSTM and GRU is truthfully reflected in our indicators.

5.3 Analysis from the Shift Perspective

The result of the shift perspective is shown in Figure 5, the picture on the top reflects $\Delta(\text{CIO}(\mathcal{E}, \mathcal{H}), \text{CIO}(\mathcal{E}^{attn}, \mathcal{A}^{attn}))$, the bottom picture reflects $\Delta(\text{DSU}(\mathcal{H}), \text{DSU}(\mathcal{A}^{attn}))$. After the introduction of attention, the sequence $\mathcal{A}^{attn}$ used for original semantic expression has a smaller degree of semantic unsaturation than $\mathcal{H}$ used for original semantic expression before the introduction of attention. At the same time, the distance between the central idea of the embedding sequence ($\text{CI}(\mathcal{E})$, $\text{CI}(\mathcal{E}^{attn})$) and the final vector used as an instance representation of the embedding sequence ($\text{CI}(\mathcal{H}), \text{CI}(\mathcal{A}^{attn})$) is also shortened. The improvement of these indicators is also reflected in the accuracy of the model.

It is worth mentioning that if we observe the changes from the perspective of different recurrent types, it is not difficult to find that the number of gate structures in the recurrent type is positively correlated with the degree of semantic unsaturation of embedding sequences and attention sequences. (There are three gate structures in LSTM, 2 in GRU and 0 in Vanilla-RNN.)

5.4 Analysis from the Holistic Perspective

Table 1 and Figure A in Appendix show that in terms of dataset size, vocabulary size and sentence length, the AGNews dataset is larger than SST. By observing the experimental results, it can be found that the changes in the many indicators of RNNs after adding attention on SST are the most obvious. On the other hand, the experimental results for LSTM and GRU without using attention are significantly better than RNN in term of semantic expression ability. However, this superiority is largely compromised by the introduction of attention, which can well recognize the central idea and effectively condense the semantics of a sequence. Therefore, for LSTM and GRU, the changes in performance caused by adding attention are relatively less than that for RNN. This also explains why the accuracy of any RNN variant can be greatly improved after the introduction of attention.

Through Figure 4 and Figure 7 in Appendix, we can see that the introduction of attention on the SST dataset has led to substantial improvements for all indicators, and the improvements on the AGNews dataset are significantly lower. Furthermore, for CIO indicators, RNN results are similar to LSTM and GRU after the introduction of attention. The CIO indicator measures the offset between central ideas and is directly related to the vector that the model finally uses to make predictions (see Formula 8 and Formula 15). Therefore, the high performance of RNN with dot-attn on SST validation set is explainable, especially a single-directional RNN model with dot-attn on the SST validation set reached 0.948.

All of these results verify that the analysis framework in our paper can objectively reflect the attention mechanism's effectiveness on the semantic expression ability.

6 Related Work

Guidotti et al. (2018) divided the problem of black-box models in detail, and the interpretability problem of attention discussed in this paper belongs to the model explanation problem.

RNNs can be said to be the basic ancestor model that introduced the attention mechanism in NLP tasks. Karpathy et al. (2015) established a mapping between the neurons of hidden layers and the content represented to explore the RNNs, Du et al. (2019) proposed a quantitative analysis framework to pave the way for effective quality analysis of

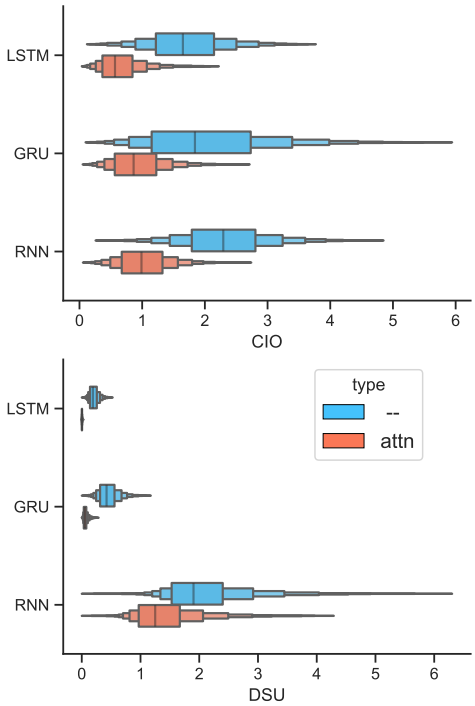


Figure 5: From the shift perspective observe the impact of the introduction of attention on the model. In each bin, top (blue) is the model without attention, bottom (red) is the model with attention.

RNNs, Zhang et al. (2020) assessing the memory ability of RNNs.

Bahdanau et al. (2014b) explained the attention from the perspective of translation alignment, Lee et al. (2017) presented an interactive interface for visualizing and intervening behavior of attention. Recently, a large amount of quantitative analysis work on the interpretability of attention has emerged, such as Vashishth et al. (2019); Jain and Wallace (2019); Serrano and Smith (2019); Wiegreffe and Pinter (2019); Jain et al. (2019), which explained the attention only focused on the attention distribution itself, and used an *erase* method.

7 Conclusions

In this paper, we have proposed a novel framework, based on a convex hull representation of sequence semantics over a Semantic Euclidean Space, to analyze the effect of attention on the semantic capacity of a RNN model and how the effect differs on different network structures. Extensive experiments on two NLP tasks provide in-depth insights on how and why attention impacts the model. From the corresponding perspective, the introduction of attention directly leads to (1) a reduction of semantic unsaturation in the hidden layer of RNNs, that is, an increase in accuracy of the original semantic expression, (2) narrowing the central idea distance between the hidden layer sequence and the embedding layer sequence, (3) an improved performance of semantic coverage between embedding layer sequence and hidden layer sequence. These are critical impacts of attention on the model and improve the capabilities of different types of RNNs. From the shift perspective, the attention layer sequence further reduces the degree of semantic unsaturation, and gets a closer proximity to the embedding layer sequence in the central idea. This is a critical factor in improving the model's accuracy. Our method illustrates how attention affects the model from the perspective of semantic transformation and makes up the limitations of the previous studies in which they only uses a single model to analyze attention.

We believe that the method proposed in this paper will help carry out more in-depth analysis of the role of attention and provide a brand-new perspective for semantic visualization in NLP tasks.

Acknowledgments

This work is funded in part by the National Key Research and Development Program of China (grant No. 2018YFC0831704) and Natural Science Foundation of China (grant No. U1636203, U1736103).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014a. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014b. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Xiaoning Du, Xiaofei Xie, Yi Li, Lei Ma, Yang Liu, and Jianjun Zhao. 2019. A quantitative analysis framework for recurrent neural network. In *2019 34th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pages 1062–1065.

Ivor D Faux and Michael J Pratt. 1979. *Computational geometry for design and manufacture*. Horwood Chichester.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5).

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Sarthak Jain, Ramin Mohammadi, and Byron C. Wallace. 2019. An analysis of attention over clinical notes for predictive tasks. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 15–21, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Andrej Karpathy, Justin Johnson, and Li Fei-Fei. 2015. Visualizing and understanding recurrent networks. *arXiv preprint arXiv:1506.02078*.

Jaesong Lee, Joong-Hwi Shin, and Jun-Seok Kim. 2017. Interactive visualization and manipulation of attention-based neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 121–126, Copenhagen, Denmark. Association for Computational Linguistics.

Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Eleventh Annual Conference of the International Speech Communication Association*.

Murray H Protter and Charles Bradfield Morrey. 1977. *College calculus with analytic geometry*. Addison-Wesley.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Youwei Song, Jiahai Wang, Tao Jiang, Zhiyue Liu, and Yanghui Rao. 2019. Attentional encoder network for targeted sentiment classification. *arXiv preprint arXiv:1902.09314*.

Shikhar Vashishth, Shyam Upadhyay, Gaurav Singh Tomar, and Manaal Faruqui. 2019. Attention interpretability across nlp tasks. *arXiv preprint arXiv:1909.11218*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*, pages 6000–6010.

Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. 2017. Coupled multi-layer attentions for co-extraction of aspect and opinion terms. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based LSTM for aspect-level sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 606–615, Austin, Texas. Association for Computational Linguistics.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

Cheng Zhang, Qiuchi Li, Lingyu Hua, and Dawei Song. 2020. Assessing the memory ability of recurrent neural networks. In *Proceedings of the 24th European Conference on Artificial Intelligence*, pages 1658–1665. IOS Press.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 207–212, Berlin, Germany. Association for Computational Linguistics.

A Datasets Detail

The statistics of sentence length in the test dataset of SST and AGNews are shown in Figure 6. The distribution of sentence length in the figure shown the sentence length in the AGNews' test set is mainly concentrated in (0, 80), in the SST's test dataset is mainly concentrated in (0, 40).

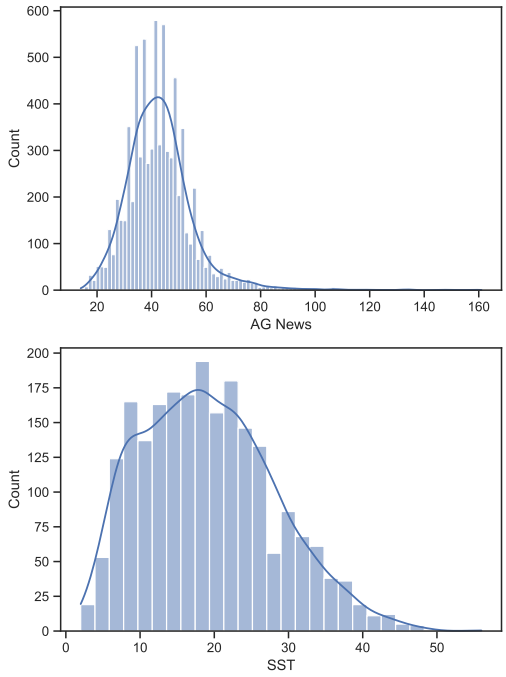


Figure 6: Sentence length statistics. The abscissa indicates the length of the sentence, and the ordinate indicates the count number.

B Reproducibility

Our experiment uses public dataset SST and AGNews. At the same time, in order to reproduce the experimental results more conveniently, we store the scores of each set of sequences in the dataset on defined indicators in a pickle binary file [2], which is convenient for you load it in and use Pandas [3] to view it. We uploaded all the pickle files saved under different models and different datasets to the code part and provided our drawing part of the code to view the experimental results disclosed in our paper, and the model code and training code will be released after some sorting.

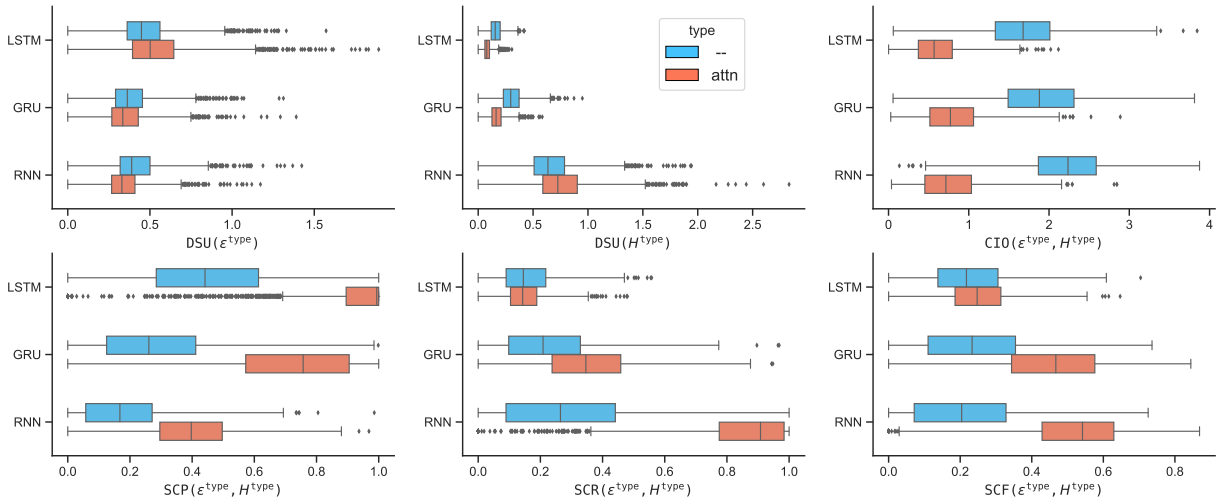
[2] https://docs.python.org/3/library/pickle.html

[3] https://pandas.pydata.org/

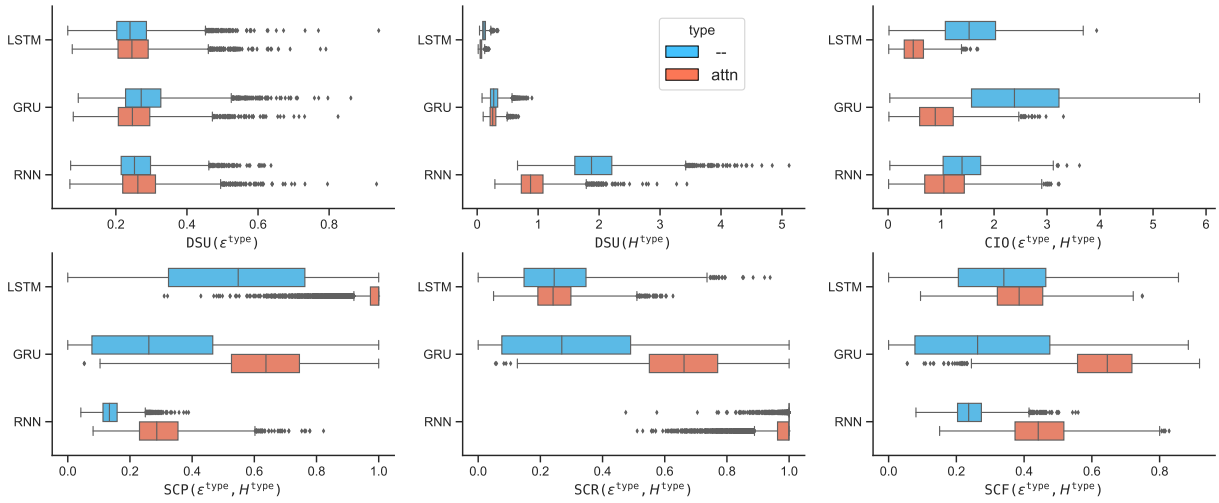
C Experiment

Figure 7 from the corresponding perspective observe the impact of the introduction of attention. Each set of pictures shows the experimental results under different experimental settings. Contains the dataset used in the experiment (SST or AGNews) and the directionality of RNNs (uni-directional or bi-directional), different types of recurrent units are compared on the ordinate of each picture.

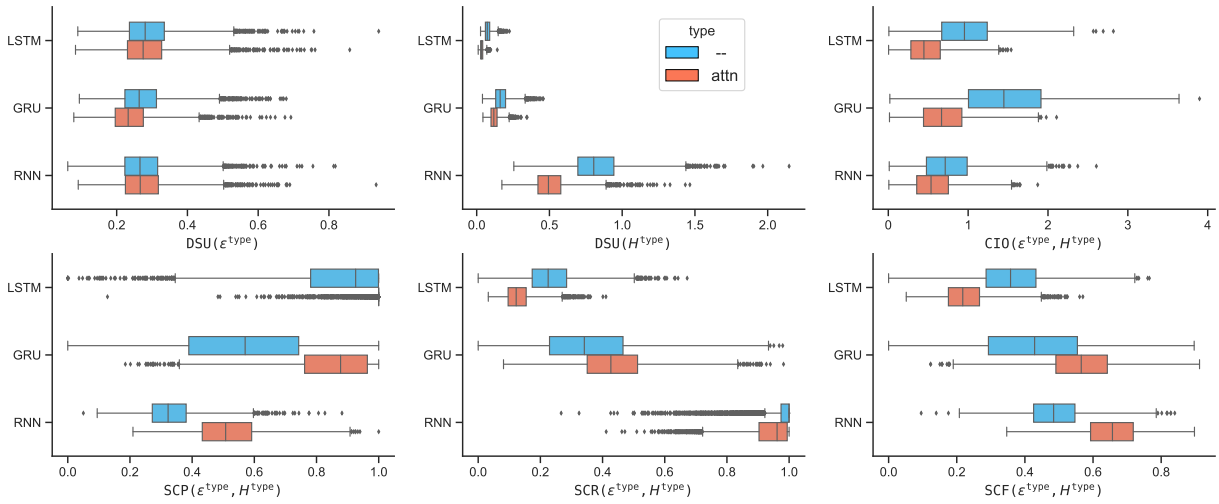
Figure 8 from the shift perspective observe the impact of the introduction of attention. Each set of pictures shows the experimental results under different experimental settings. Contains the dataset used in the experiment (SST or AGNews) and the directionality of RNNs (uni-directional or bi-directional), different types of recurrent units are compared on the ordinate of each picture.



(a) SST, uni-directional

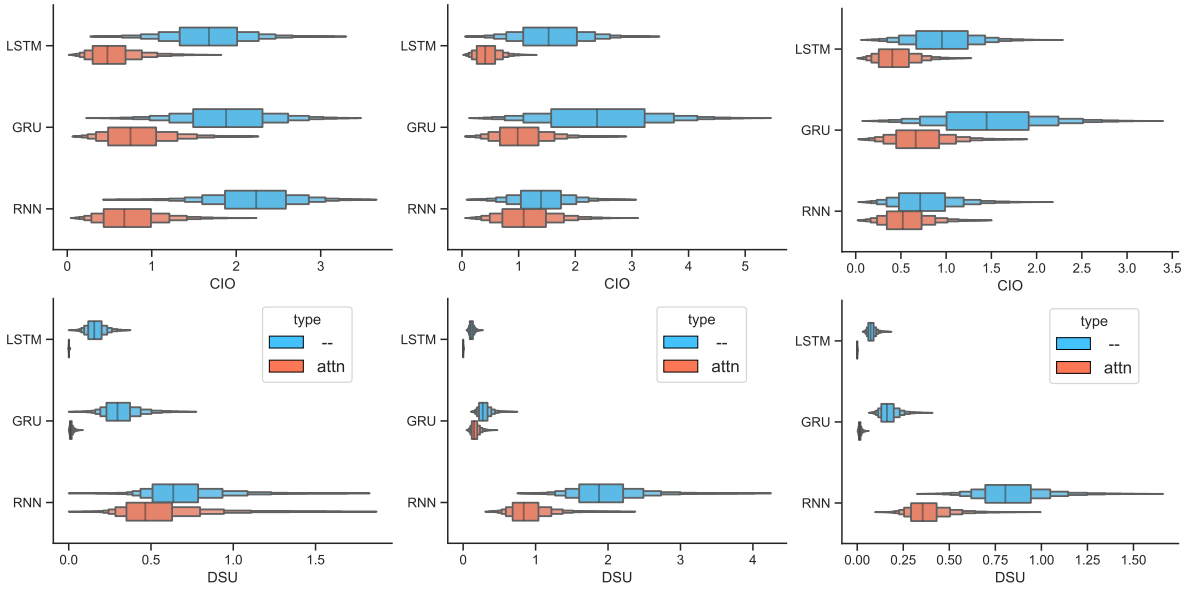


(b) AGNews, bi-directional



(c) AGNews, uni-directional

Figure 7: From the corresponding perspective observe the impact of the introduction of attention.



(a) SST, uni-directional (b) AGNews, bi-directional (c) AGNews, uni-directional

Figure 8: From the shift perspective observe the impact of the introduction of attention.