Rebuttal Information

Because of page number restriction, we decided to add more supplement information for our paper.
This supplement information include four main parts.

- The Advantage Of FMHSA
- Ablation Experiment
- Difference Between IRU And RRU
- Some Abbreviation Of Some Terms And Writing Mistakes In Paper

The following pages will discuss about these topics. Thanks for your patience to view this rebuttal!

The Advantage Of FMHSA

In the original self-attention module, the input X is transformed into query Q, key K, and value V through a linear transformation. The dimensions of the input, key, and value are represented by d, dk, and dv, respectively, and the number of patches is represented by $n = H \times W$. The self-attention module is then applied to Q, K, and V using the formula $\mathrm{Attn}(Q, K, V) = \mathrm{Softmax}(QK/\sqrt{dk})V$, where Softmax denotes the softmax function. To reduce the computational overhead, we use a $k \times k$ depth-wise convolution with stride 1 to reduce the spatial size of K and V, and three sub-linear layers are applied before the attention operation, see Fig.1.

To further optimize the computation, we introduce the Fast Multi-head Self-attention (FMHSA) module, which involves h "heads." In other words, h Fast functions are applied to the input. Each head produces a sequence of size $n \times (d/h)$, and these h sequences are concatenated into a $n \times d$ sequence. The FMHSA module effectively enables parallel processing by allowing multiple heads to operate simultaneously on the input data, thereby reducing computation time.

The FMHSA module extends the original self-attention module by introducing multiple heads to enable parallel processing, and thus significantly reducing computation time. By optimizing the computation in this way, we can effectively apply self-attention to large datasets and achieve state-of-the-art results in various machine learning tasks.

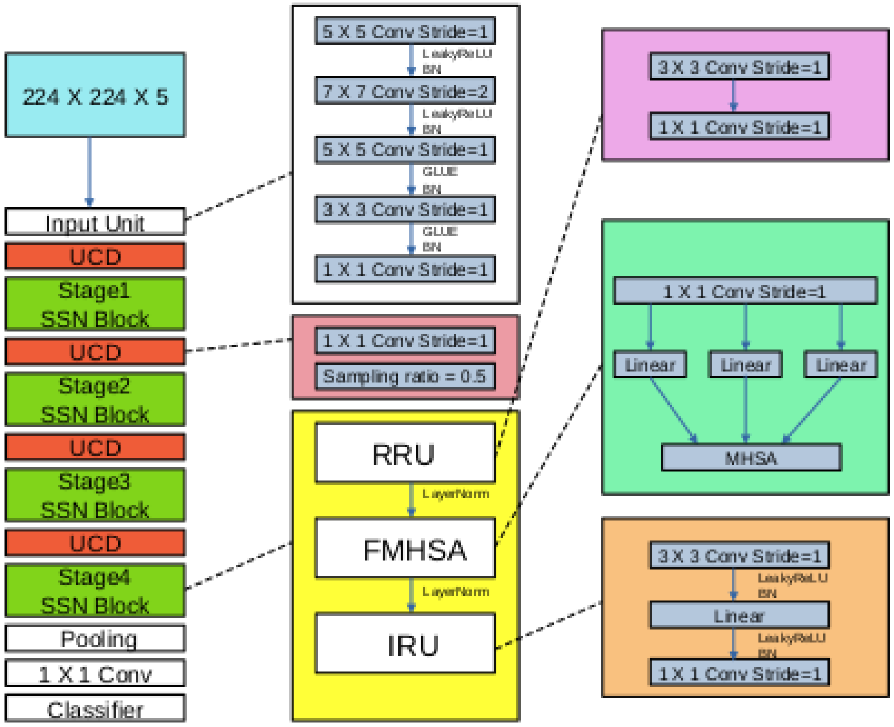

Figure 1: SSN structure

Ablation Experiment

The experiment focused on the SSN modula which consists of three main units (RRU, FMHSA, IRU). The RRU was replaced with two FC layer (maintaining the input and output dimensions and total number of layers). The results in Table.1 show a slight decrease in processing time (around 7% per iteration), but total collision times increased to 28.2.

The contribution of the FMHSA unit was evaluated by replacing it with three FC layers (again, maintaining the input and output dimensions and total number of layers) to convert the double sequential process (Bi-LSTM + FMHSA) to a single sequential process (Bi-LSTM). The results in Table.1 indicate a significant decrease in processing time (around 31% per iteration), but a significant increase in total collision times to 60.6.

Finally, the IRU unit was replaced with three FC layers (maintaining the input and output dimensions and total number of layers). The results in

Table.1 show a slight decrease in processing time (around 11% per iteration), but a slight increase in total collision times to 23.1. Overall, these experiments demonstrate the importance of each component in the BCSSN and how optimizing each component can improve the performance of the network.

Structure	**Front collision**	**Side collision**	**Rear collision**	**Total collision**	**Process time** (A5000)
SSN	2.6	13.3	3.5	19.4	1.30μs
Without RRU	5.2	9.6	13.4	28.2	1.21μs
Without FMHSA	9.6	17.7	33.3	60.6	0.90μs
Without IRU	3.3	6.1	13.7	23.1	1.16μs

Table 1: Ablation Result

Difference Between IRU And RRU

IRU has one linear layer so that we can reduce the input size rapidly. Besides, RRU didn't contain a normalization process since during that process we did not need to keep the data value in a fixed range. Besides,we also did an Ablation Experiment to validate the functions of these Units. We upload code on Github.

Some Writing Mistakes In Paper

“The proposed SSN block consists of a Reinforcement Region Unit (RRU), a Fast Multi-Head Self-Attention (FMHSA) module and an Information Refinement Unit (IRU), as shown in Fig. 4. We will describe these four components in the following.” Here should be three components not four.

CNN: Convolutional Neural Network

RNN: Recurrent Neural Network

STDN: Spatio-Temporal Deep Learning Network

VGG: Visual Geometry Group

NLP: Natural Language Processing

CV: Computer Vision