This work introduces an end-to-end learnable, sequence length agnostic and parallelizable method for sequence to sequence learning tasks. Several novel techniques such as self-attention, layer-norm seems to be deciding factor over the encouraging results. Adding to that, the model already achieves state-of-the-art results on the WMT datasets, which is a major advantage considering broad area of application in this field.

Advantages:

+ Parallelized training will make it faster to train this architecture.
+ Due to sequence length agnostic nature, it is naturally adaptive to a large range of language translation and other NLP tasks.
+ A direct replacement of RNN and convolution architectures.

Disadvantages:

- I feel that many minor level details are hidden underneath over a tightly fitted details in this paper which makes it hard to follow in first go. Authors may provide a detailed explanation of the paper through some other platform.
- Many hyperparameters are provided without any intuition behind them. Though, I understand it could be difficult to analytically examine the quantitative values of hyperparameters here, the work falls short on interpretability part and thus hard to extend on a solid ground.

Overall, considering the novelty of architecture and the magnitude of applicability of this work in various field, it is a strong accept from my side.