This is a technical paper on a new extension to the Transformer model, which is a popular neural network architecture used in machine translation tasks.

Summary

The authors propose an extension to the self-attention mechanism, which can be used to incorporate relative position information for sequences. This extension improves performance for machine translation tasks.

Key contributions

1. **Relative position information**: The authors introduce a new way to consider relative positions between elements in a sequence. This is done by introducing two new attention mechanisms: aV and aK.
2. **Improved performance**: The extension improves performance on machine translation tasks, such as WMT14 English-to-German translation.
3. **Ablation study**: The authors perform an ablation study to evaluate the effectiveness of their proposed extension.

Methodology

1. **Transformer architecture**: The Transformer model is used as a baseline for comparisons.
2. **Self-attention mechanism**: The self-attention mechanism is extended to consider relative position information using aV and aK.
3. **Machine translation task**: The authors use the WMT14 English-to-German translation task to evaluate their proposed extension.

Experimental results

The experimental results show that the proposed extension improves performance on machine translation tasks, compared to the baseline Transformer model without relative position information.

Conclusion

The authors conclude that their proposed extension to self-attention can be used to incorporate relative position information for sequences, improving performance on machine translation tasks. They also highlight the need for further research on more general applications of this mechanism, such as considering arbitrary directed labeled graph inputs.

Future work

1. **Arbitrary directed graph inputs**: The authors plan to extend their proposed mechanism to consider arbitrary directed, labeled graph inputs.
2. **Nonlinear compatibility functions**: The authors are interested in exploring nonlinear compatibility functions to combine input representations and edge representations.

Overall, this paper presents a new extension to the Transformer model, which can be used to improve performance on machine translation tasks.