Conclusion

Summary of Contributions and Research

In this thesis, we presented an efficient end-to-end self-supervised approach to vision-language learning, that is significantly cheaper to train and smaller in size compared to existing multimodal models. Our experiments show, especially on retrieval tasks, promising results