Comparing CNNs and ViTs for Computer Vision Tasks

This report examines the comparison between Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) in the field of computer vision. CNNs have been the de facto standard for visual recognition tasks for the past decade, and their hierarchical structure of convolutional, pooling, and fully connected layers allows them to learn complex representations of visual data. ViTs, originally designed for natural language processing tasks, have recently been adapted for image recognition tasks and have demonstrated the potential to surpass CNNs in terms of accuracy and efficiency. The report examines the workings of both CNNs and ViTs, their strengths and weaknesses, and provides a comprehensive comparison to determine their respective roles in the future of computer vision.

The report finds that ViTs have been shown to outperform traditional CNNs on various computer vision tasks, and a recent study found that ViTs outperformed ConvNets in 13 out of 15 tasks, including fine-grained classification, scene recognition, open-domain classification, and face recognition. The study also found that ViTs had a stronger ability to capture subtle differences and comprehend complex scenarios, making them more suitable for transfer learning. Additionally, the report finds that ViTs are more robust to adversarial perturbations compared to CNNs, due to their attention mechanism, but are not inherently invariant to geometric transformations, such as translations, and are sensitive to changes in input distributions.

The report references several research papers that compare the performance of VITs and CNNs, including Raghu et al. (2021), Bai et al. (2021), Park and Kim (2022), and Liu et al. (2022). It also references two blog posts from Google and Becoming Human AI, which provide an overview of VITs and their applications. Finally, it references Dosovitskiy et al. (2020), which provides an in-depth look at VITs and their potential for image recognition. Ultimately, the choice between CNNs and ViTs depends on the specific requirements of the application, including the desired classification accuracy, computational efficiency, and inference speed.