Discussion Questions

1. What are the advantages of ViTs over ResNets?

2. What are the inductive biases of CNNs and why are they useful?

3. Are there any interesting/surprising empirical results that might not be obvious?

4. Is the comparison with ResNets fair (due to different # of params)?

5. How reasonable is this line of research given the amount of compute required (e.g., ~2.5K TPU days)? How can academics reproduce this work?

6. How does changing the patch size impact the accuracy?

7. Why do hybrid CNN-ViT models perform worse than ViTs when trained with large datasets?

8. Does the ViT paper have any significant contributions?

9. How does the trade-off between interpretability and performance impact the adoption of Transformer models in computer vision?

10. Will ViTs take over CNNs as the de facto standard model for computer vision tasks?

11. How does ViT's dependence on large-scale datasets for pre-training impact its applicability in real-world scenarios, particularly where large datasets are not available?

Discussion Questions

1. What are the advantages of ViTs over ResNets?

Discussion Questions

2. What are the inductive biases of CNNs and why are they useful?

Discussion Questions

3. Are there any interesting/surprising empirical results that might not be obvious?

Discussion Questions

4. Is the comparison with ResNets fair (due to different # of params)?

Discussion Questions

5. How reasonable is this line of research given the amount of compute required (e.g., ~2.5K TPU days)? How can academics reproduce this work?

Discussion Questions

6. How does changing the patch size impact the accuracy?

Discussion Questions

7. Why do hybrid CNN-ViT models perform worse than ViTs when trained with large datasets?

Discussion Questions

8. Does the ViT paper have any significant contributions?

Discussion Questions

9. How does the trade-off between interpretability and performance impact the adoption of Transformer models in computer vision?

Discussion Questions

10. Will ViTs take over CNNs as the de facto standard model for computer vision tasks?

Discussion Questions

11. How does ViT's dependence on large-scale datasets for pre-training impact its applicability in real-world scenarios, particularly where large datasets are not available?