Discussion Questions

1. Why are vision transformers harder to scale compared to language transformers? Why is the proposed model only 22B parameters (rather than >500B)?

2. What are some aspects of performance that are not talked about as much as they should be?

3. Would the performance of ViT improve with even more parameters?

4. Could larger and larger transformers eventually solve any ML/CV task given the right dataset and training scheme?

5. What do you think about the environmental cost of training such large models? Should we keep scaling or instead try to make the models more resource-efficient?

6. How does the scaling of parameters in ViT impact their ability to generalize across diverse tasks?

7. What strategies or technologies can be used in the future to make large-scale image models more accessible for researchers and everyday users? Should the authors be required to open source such models?

8. What are some potential ethical issues of the application of the ViT-22B model? Does Internet provide good data for learning fair and unbiased computer vision models? How should we deal with this problem?

9. The paper has no technical contributions but presents lots of engineering and experimental details. Is this useful? Do we overemphasize the importance of technical contributions?

10. Were you impressed or underwhelmed by the performance of ViT-22B? Are the comparisons with other models fair?

Discussion Questions

1. Why are vision transformers harder to scale compared to language transformers? Why is the proposed model only 22B parameters (rather than >500B)?

Discussion Questions

2. What are some other aspects of performance that are not talked about as much as they should be?

Discussion Questions

3. Would the performance of ViT improve with even more parameters?

Discussion Questions

4. Could larger and larger transformers eventually solve any ML/CV task given the right dataset and training scheme?

Discussion Questions

5. What do you think about the environmental cost of training such large models? Should we keep scaling or instead try to make the models more resource-efficient?

Discussion Questions

6. How does the scaling of parameters in ViT impact their ability to generalize across diverse tasks?

Discussion Questions

7. What strategies or technologies can be used in the future to make large-scale image models more accessible for researchers and everyday users? Should the authors be required to open source such models?

Discussion Questions

8. What are some potential ethical issues of the application of the ViT-22B model? Does Internet provide good data for learning fair and unbiased computer vision models? How should we deal with this problem?

Discussion Questions

9. The paper has no technical contributions but presents lots of engineering and experimental details. Is this useful? Do we overemphasize the importance of technical contributions?

Discussion Questions

10. Were you impressed or underwhelmed by the performance of ViT-22B? Are the comparisons with other models fair?