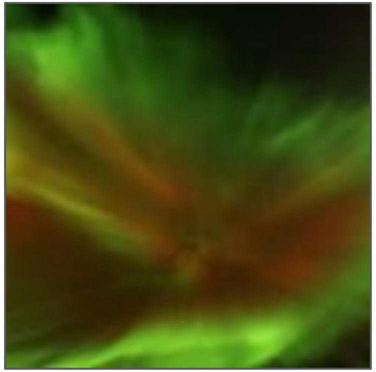
Evaluating Pretraining Methods for Deep Learning on Geophysical Imaging Datasets

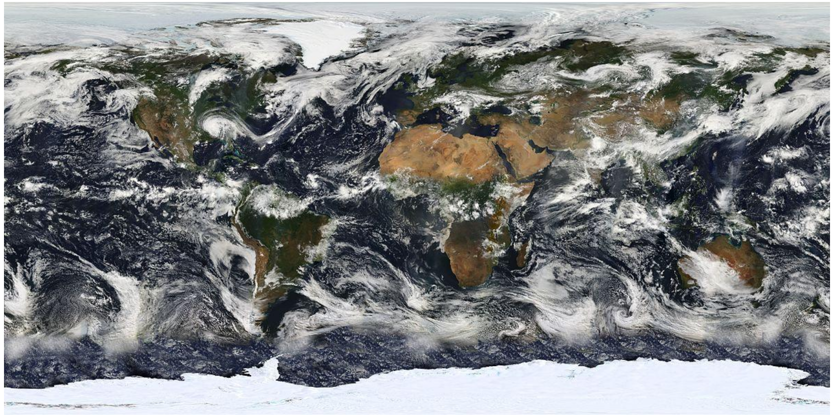
James Chen

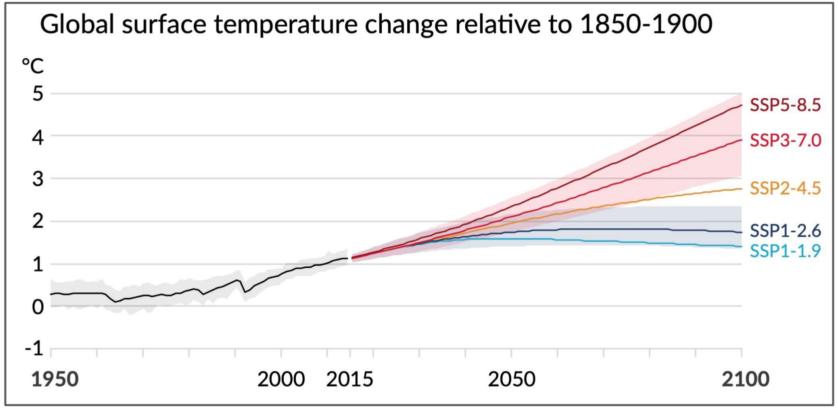


NeurIPS 2021 Tackling Climate Change with Machine Learning Workshop

Motivation: Clouds and Climate

- Clouds play a vital role in sensitivity of climate to changes in CO_2 concentration
- Some types of clouds trap heat; others reflect heat away
- IPCC 2021: “Clouds remain the largest contribution to overall uncertainty in climate feedbacks”
  - Leads to larger error bars in projections of future climate scenarios

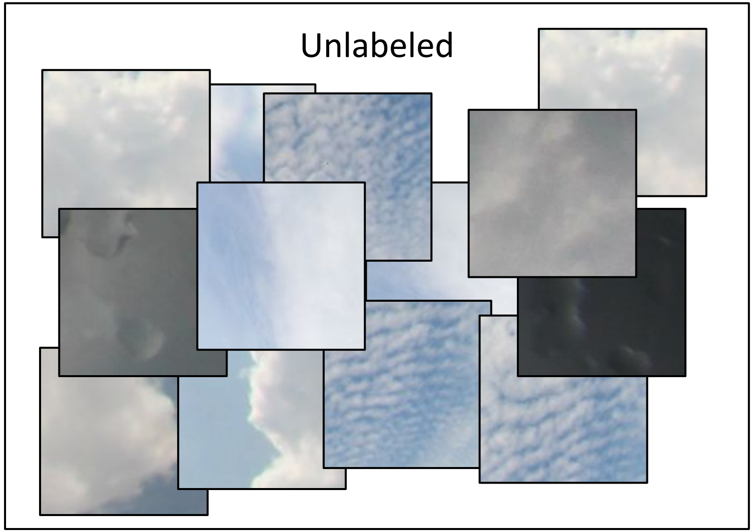


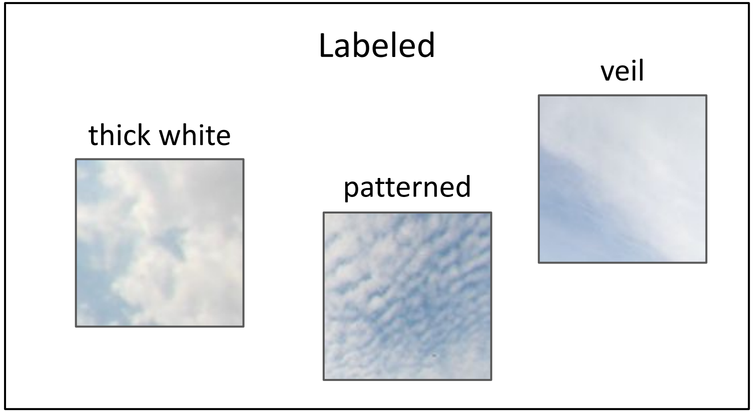


Temperature projections under different scenarios (IPCC 2021)

Limited Availability of Labeled Data

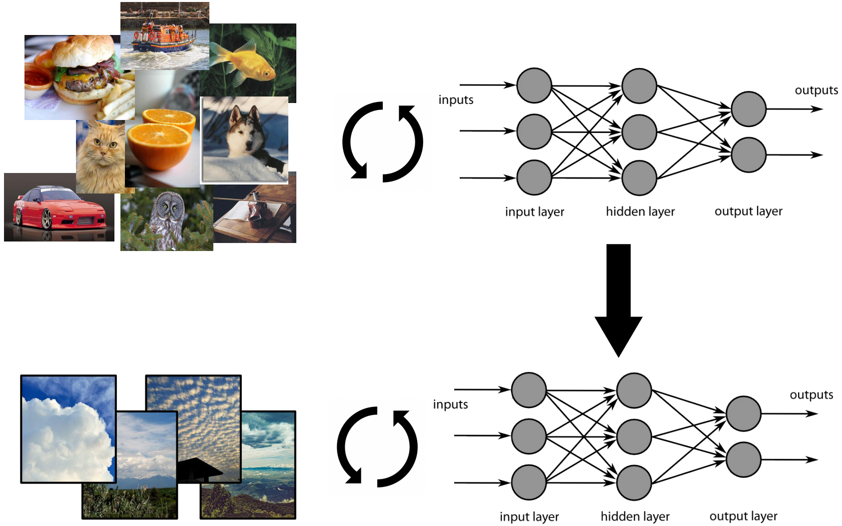
- Machine learning can automatically classify cloud types to improve climate modeling
- Need large amount of labeled data for automatic classification; however, a lot of human effort required to label cloud images
- Many raw images of clouds but very few labeled images (100s to few 1000s in past work)





Transfer Learning

- Pretrain neural network on auxiliary "source" dataset (e.g. internet images)
- Finetune on your own "target" dataset (e.g. cloud images)
- Transfers patterns learned in source dataset to warm-start training on target dataset



1) Pretrain on "source" dataset

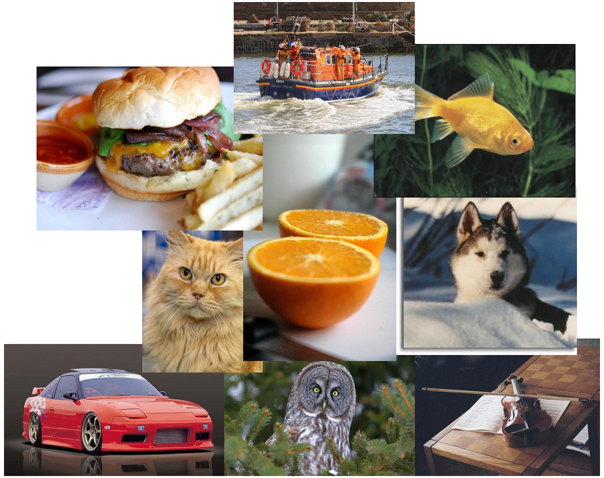
2) Copy pretrained weights

3) Continue training on "target" dataset

Smaller set of other
cloud images

What data should we pretrain on?

Lots of images
from different task



Target dataset

Smaller set of other cloud images

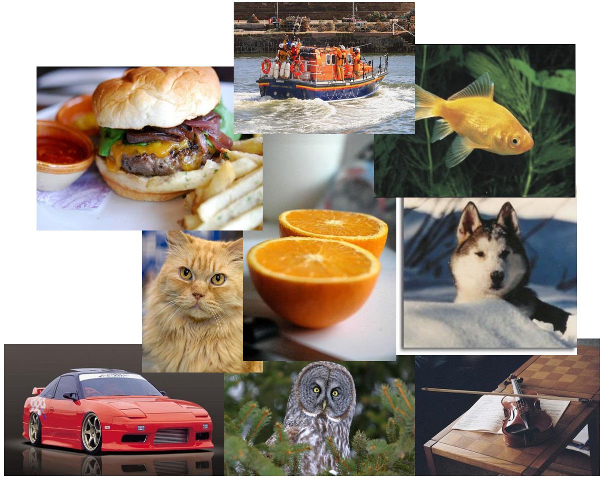
- Clausen et al., 2018
- Marmanis et al., 2016
- Zhong et al., 2020

What data should we pretrain on?

No clear consensus

Target dataset

Lots of images from different task



- Ham et al., 2019
- Rasp et al., 2021
- Zhang et al., 2018

Datasets

Cloud Classification:

- CCSN - 2,543 cloud images, 11 classes
- SWIMCAT - 784 cloud images, 5 classes

Aurora Classification:

- Kiruna - 3,846 aurora images, 7 classes
- YR1 - 1,200 aurora images, 4 classes
- YR2 - 8,001 aurora images, 4 classes

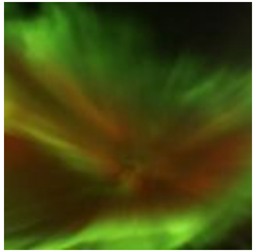
General Purpose Classification:

- ImageNet - 1,350,000 images, 1,000 classes

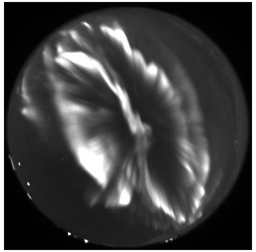
CCSN

SWIMCAT

Kiruna



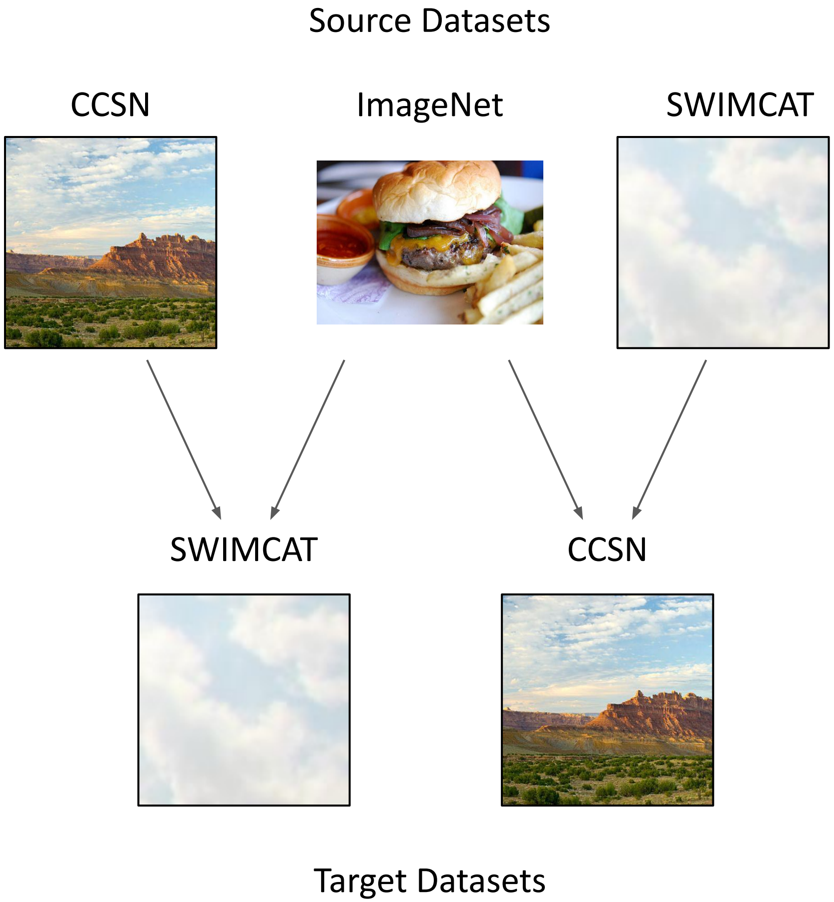
YR1/2



ImageNet

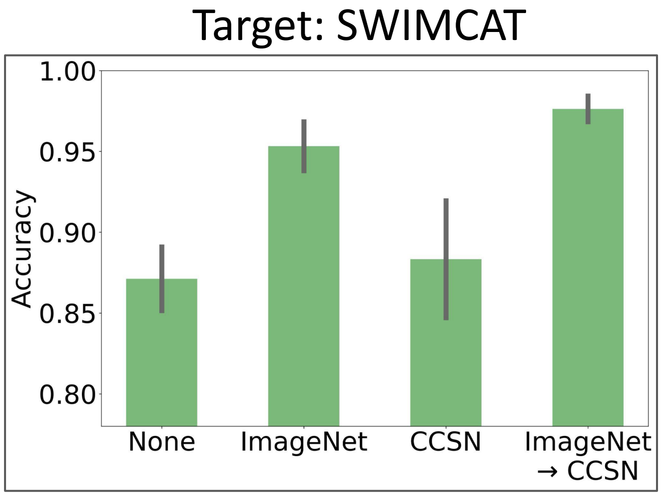
Experimental Setup

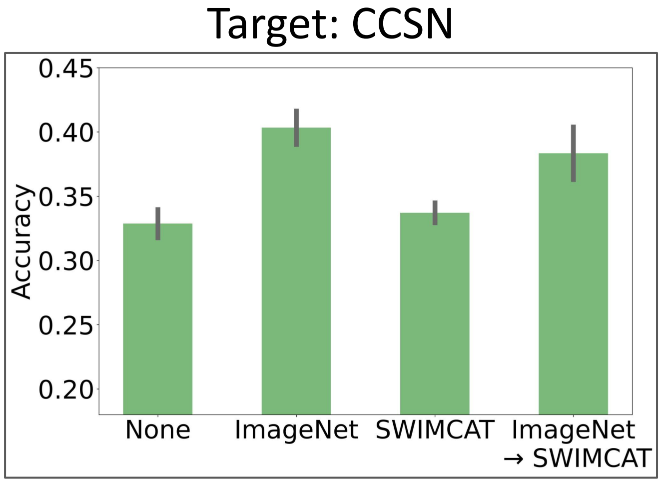
- Use each cloud/aurora dataset as the target dataset
- For each target dataset, try pretraining on all other datasets (i.e. with SWIMCAT as target, try pretraining on CCSN and ImageNet)
- Also try pretraining on multiple source datasets in sequence (i.e. ImageNet → CCSN → SWIMCAT)
- Model architecture: ResNet18 (Convolutional Neural Network)



Cloud Classification Results

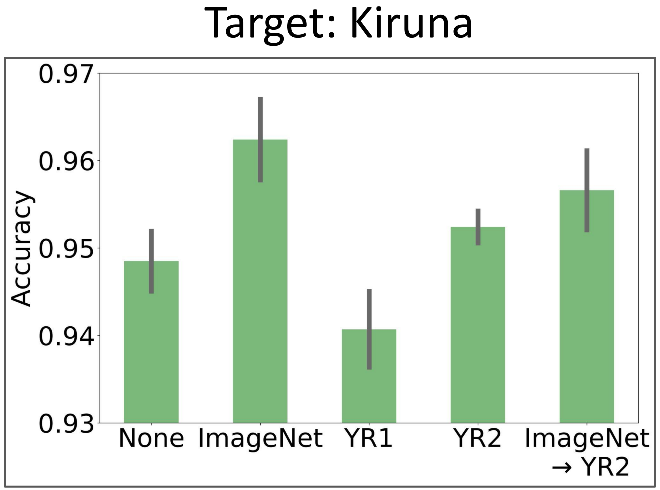
- Transfer learning can significantly improve accuracy, *depending on the source dataset*
- ImageNet was the best single source dataset: improves accuracy over 7% in both cases
- With SWIMCAT as target, multiple transfer learning steps improved accuracy by an additional 2%

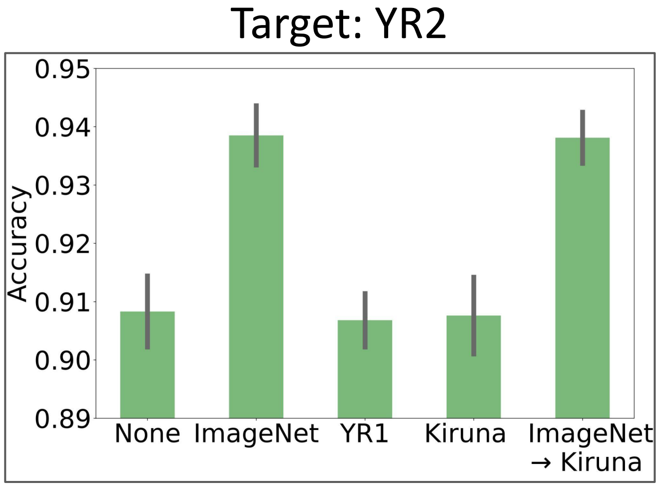




Aurora Classification Results

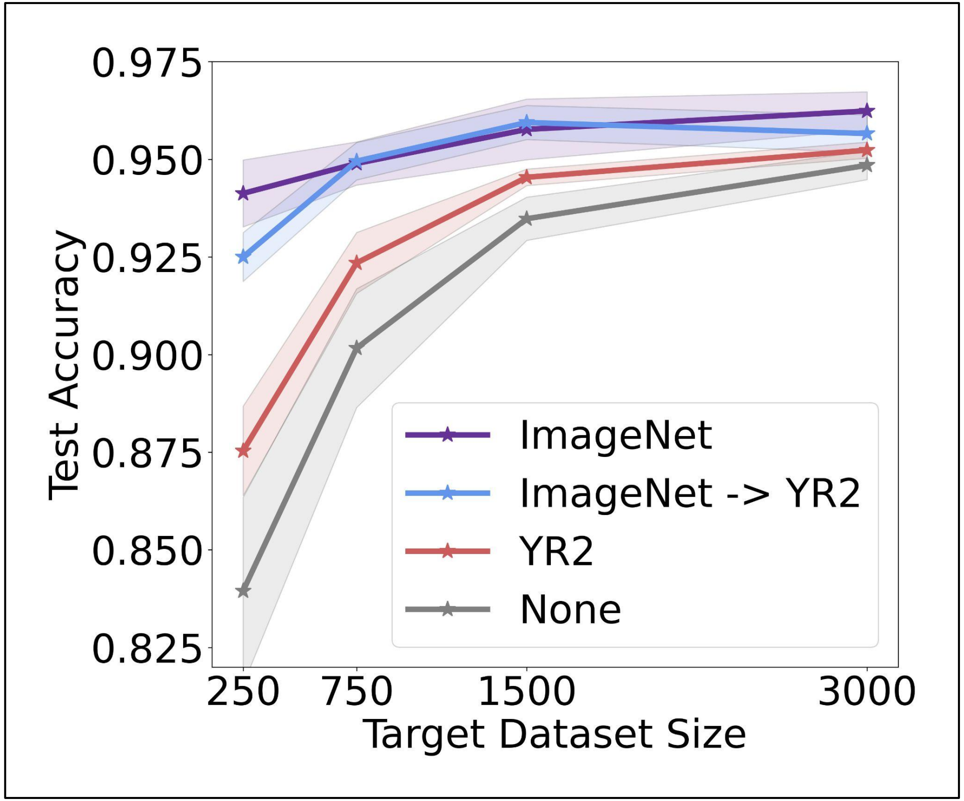
- ImageNet is best source dataset, still giving up to 3% increase in accuracy
- Pretraining on YR2 is much more effective than pretraining on YR1: the images are similar but YR2 is 8x larger
- Multiple transfer steps do no better than pretraining on ImageNet





Varying Target Dataset Size

- Artificially varied target dataset size by subsampling Kiruna dataset
- Choice of source dataset more important with less labeled target data, with differences of 10% accuracy for target dataset size 250
- Across sizes, ImageNet and ImageNet $\longrightarrow$ YR2 are best



Conclusion

- Size of source dataset matters most
- Benefits of using a large source dataset are greater with smaller target datasets
- Multiple transfer learning steps generally do not yield additional benefit, but were helpful in one instance
- Identifies best practices for using transfer learning for automated climate analysis

Smaller set of other cloud images

Lots of images from different task