# Impact of Modern Transformer Architectures on Federated Learning for Remote Sensing

Students: Sebastian Völkers, Kenneth Weitzel, Felix Zailskas

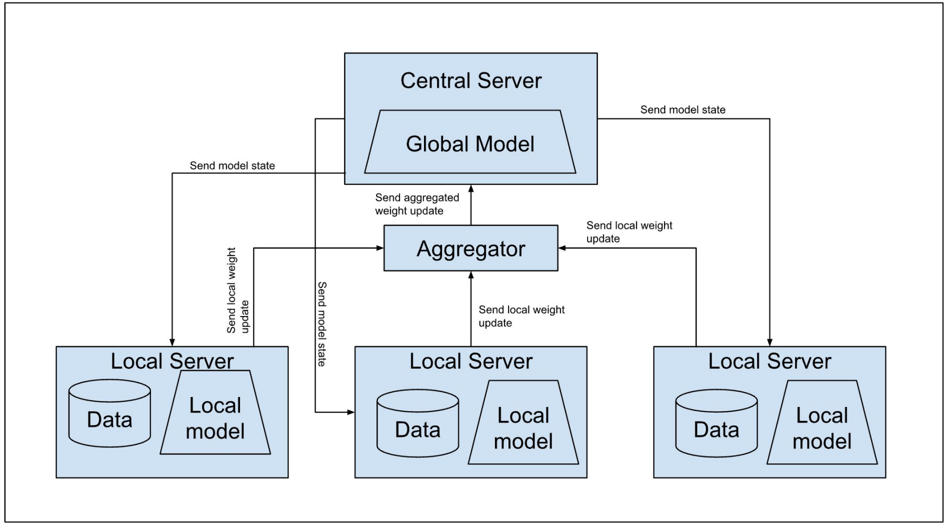
Supervisor: Baris Buyuktas

# Table of Contents

# The Federated Learning Setup

# Federated Learning (FL) Idea & Issues

- Often datasets are distributed across multiple clients and not available for everyone
    - Privacy, commercial interest, legal regulations [1]
- Train local models on available data of the client and aggregate results to train global model




- Client data may be non-IID
    - Distribution of labels may vary
    - Different amounts of data
    - Different distribution for the same class (concept drift)
- This can lead to diverging local weight updates
- Hinders global convergence
- Transformers can help tackle data heterogeneity in FL [2]



[1] Büyüktaş, B., Sümbül, G., & Demir, B. (2023). Federated learning across decentralized and unshared archives for remote sensing image classification. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2311.06141
[2] Qu, L., Zhou, Y., Liang, P. P., Xia, Y., Wang, F., Adeli, E., Li, F., & Rubin, D. L. (2022). Rethinking architecture design for tackling data heterogeneity in federated learning. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr52688.2022.00982

# Solution attempts for non-IID data in FL

## Local Training Focused Algorithms [1]

➢ Adjust the local training in the clients by improving the empirical risk minimization

1. Add auxiliary terms to the loss function
   - Auxiliaries try to reduce the deviation of local updates from each other
   - Local convergence may be hindered
2. Adjust the gradient
   - Increase local model's generalization capabilities
   - Adjust local gradient based on global update

## Model Aggregation Focused Algorithms [1]

➢ Adjust the way local updates are aggregated into the global update

1. Weighted model averaging
   - Normalize parameter updates
2. Personalized model averaging
   - Cluster clients and create global model for each cluster
3. Knowledge distillation
   - Use the outputs of local models (their learned knowledge) for the global update rather than the direct parameter updates



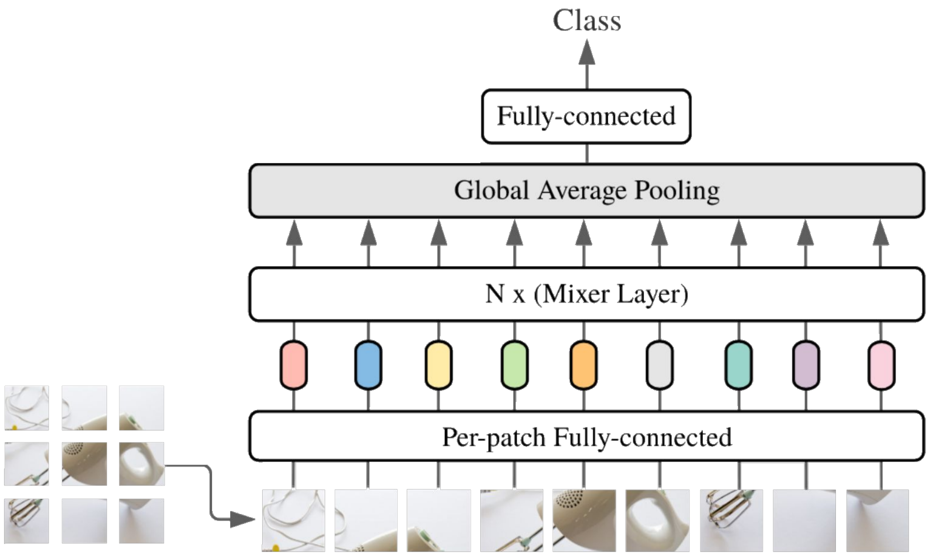
[1] Büyüktaş, B., Sümbül, G., & Demir, B. (2023). Federated learning across decentralized and unshared archives for remote sensing image classification. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2311.06141

# The MLP-Mixer Architecture [3]



[3] Tolstikhin, I., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J., Lučić, M., & Dosovitskiy, A. (2021). MLP-Mixer: an all-MLP architecture for vision. arXiv (Cornell University). https://arxiv.org/pdf/2105.01601.pdf

# MLP-Mixer Architecture



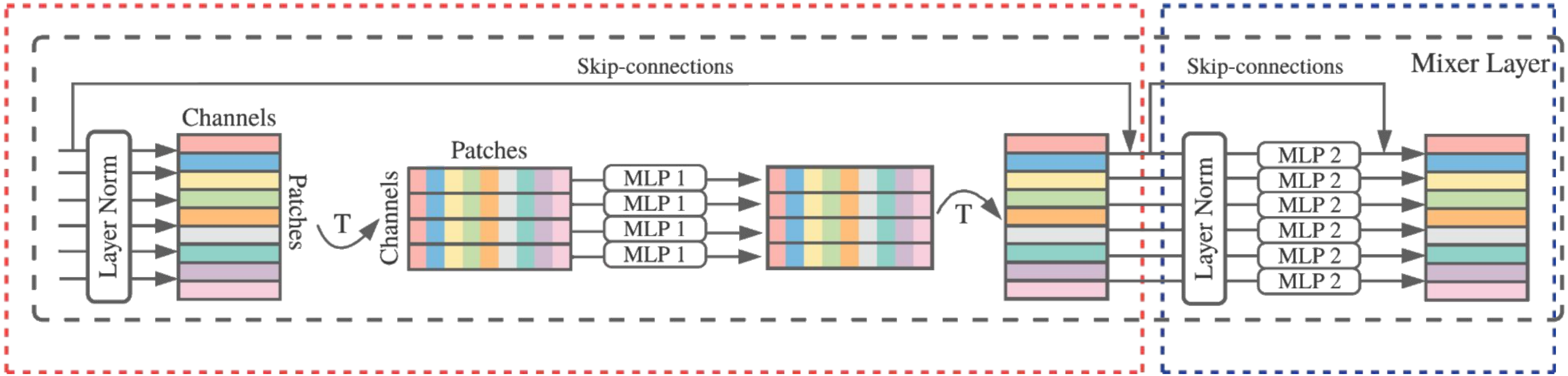
Model architecture of MLP-Mixer [3]

- Goal to create computer vision model without convolution and self-attention
- Model that can keep up with state of the art CNNs and Vision Transformers
- Ideas from recent VisionTransformer Paper
- Only using multilayer perceptrons that are repeatedly applied across spatial locations and feature channels



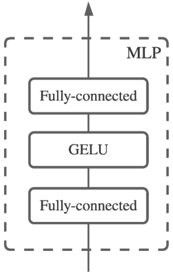
[3] Tolstikhin, I., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J., Lučić, M., & Dosovitskiy, A. (2021). MLP-Mixer: an all-MLP architecture for vision. arXiv (Cornell University). https://arxiv.org/pdf/2105.01601.pdf

# MLP-Mixer Architecture




Mixer Layer [3]

- Two parts consists of Token-Mixer and Channel-Mixer
- Token-Mixer: cross-location mixing
- Channel-Mixer: per-location mixing
- 2 MLPs with the following architecture used that each use same parameters across inputs

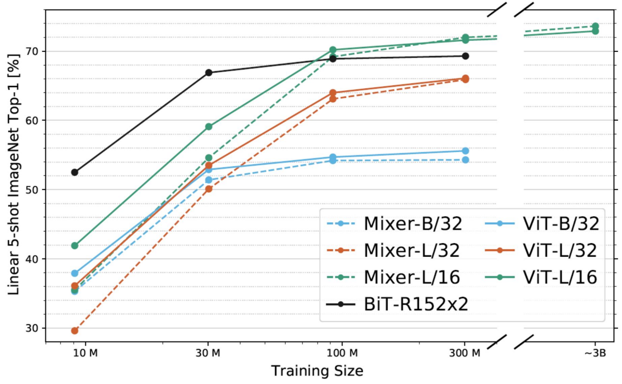





[3] Tolstikhin, I., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J., Lučić, M., & Dosovitskiy, A. (2021). MLP-Mixer: an all-MLP architecture for vision. arXiv (Cornell University). https://arxiv.org/pdf/2105.01601.pdf

# MLP-Mixer Paper Results

| | ImNet top-1 | ReaL top-1 | Avg 5 top-1 | VTAB-1k 19 tasks | Throughput img/sec/core | TPUv3 core-days |
|---|---|---|---|---|---|---|
| Pre-trained on ImageNet-21k (public) | | | | | | |
| ● HaloNet [51] | 85.8 | — | — | — | 120 | 0.10k |
| ● Mixer-L/16 | 84.15 | 87.86 | 93.91 | 74.95 | 105 | 0.41k |
| ● ViT-L/16 [14] | 85.30 | 88.62 | 94.39 | 72.72 | 32 | 0.18k |
| ● BiT-R152x4 [22] | 85.39 | — | 94.04 | 70.64 | 26 | 0.94k |
| Pre-trained on JFT-300M (proprietary) | | | | | | |
| ● NFNet-F4+ [7] | 89.2 | — | — | — | 46 | 1.86k |
| ● Mixer-H/14 | 87.94 | 90.18 | 95.71 | 75.33 | 40 | 1.01k |
| ● BiT-R152x4 [22] | 87.54 | 90.54 | 95.33 | 76.29 | 26 | 9.90k |
| ● ViT-H/14 [14] | 88.55 | 90.72 | 95.97 | 77.63 | 15 | 2.30k |




MLP-Mixer Results [3]

- ➢ MLP-Mixer results on the same level with other models
- ➢ Higher throughput with shorter computation time
- ➢ MLP-Mixer scalable and results improve with bigger datasets, even surpasses CNN and Vision Transformer results
- ➢ With smaller datasets results are a lot worse
- ➢ Generally good computation-accuracy trade-off

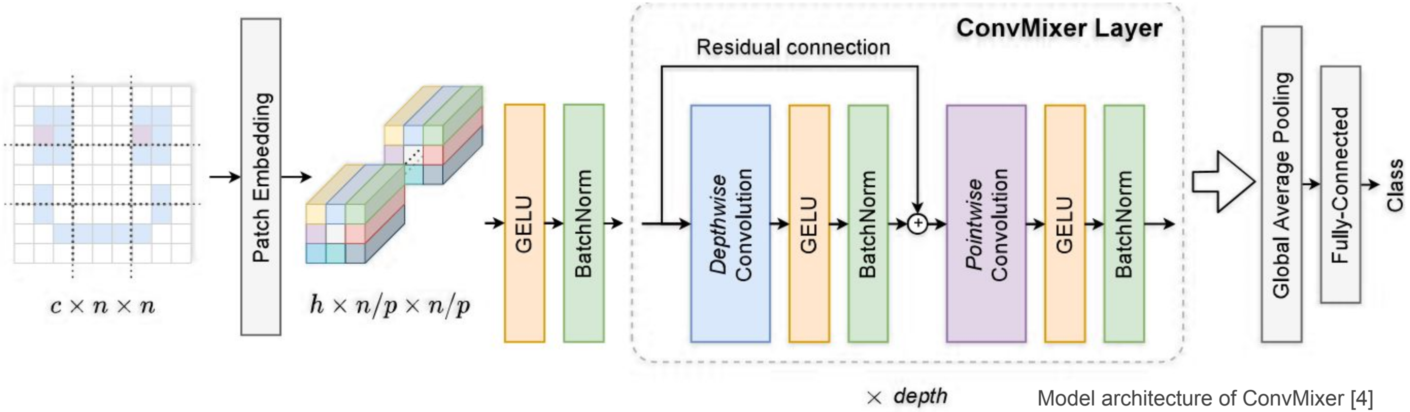


[3] Tolstikhin, I., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J., Lučić, M., & Dosovitskiy, A. (2021). MLP-Mixer: an all-MLP architecture for vision. arXiv (Cornell University). https://arxiv.org/pdf/2105.01601.pdf

# The ConvMixer Architecture [4]

[4] Trockman, A., & Kolter, J. Z. (2022). Patches are all you need? arXiv (Cornell University). https://doi.org/10.48550/arxiv.2201.09792

# ConvMixer Architecture




Model architecture of ConvMixer [4]

- Apply linear embedding to Patches instead of pixels, as in Vision Transformers [5]

- Built upon MLP-Mixer, with separate spatial and channel-wise mixing, while replacing MLPs by convolutional layers [3]



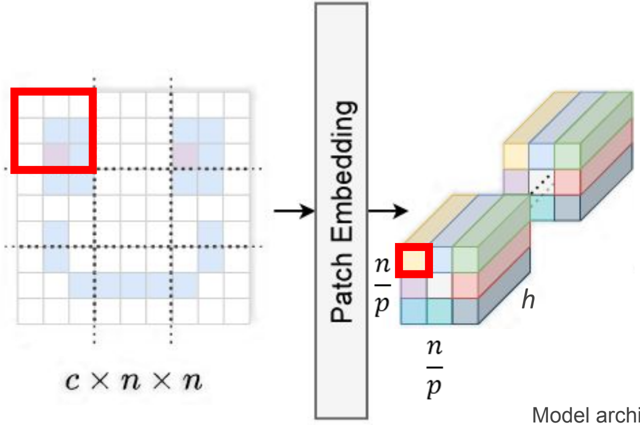
[3] Tolstikhin, I., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J., Lučić, M., & Dosovitskiy, A. (2021). MLP-Mixer: an all-MLP architecture for vision. arXiv (Cornell University). https://arxiv.org/pdf/2105.01601.pdf
[4] Trockman, A., & Kolter, J. Z. (2022). Patches are all you need? arXiv (Cornell University). https://doi.org/10.48550/arxiv.2201.09792
[5] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv (Cornell University). https://openreview.net/pdf?id=YicbFdNTTy

# Patch Embedding

**Input Shape:**
- Channel size $c$
- Spatial extent $n$

**Embedding via Convolution:**
- input channels $c$
- output channels $h$
- kernel size $p$
- stride $p$

**Embedding shape:**
- patch size $p$
- embedding dimension $h$




Model architecture of ConvMixer [4]



[4] Trockman, A., & Kolter, J. Z. (2022). Patches are all you need? arXiv (Cornell University). https://doi.org/10.48550/arxiv.2201.09792

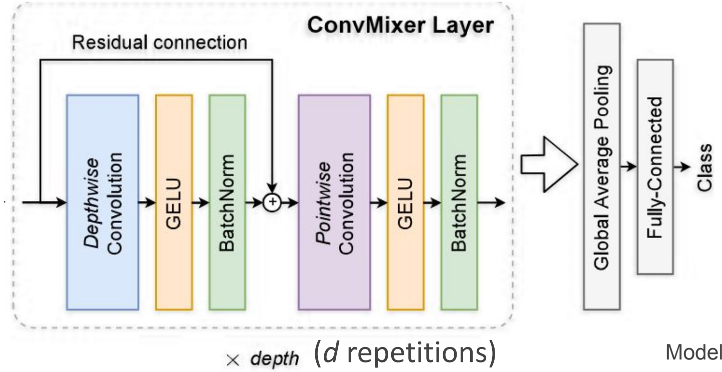
# ConvMixer Layer

## Depthwise convolution

- Spatial mixing
- Grouped convolution
- #groups = hidden dimension h
- Add layer input via Residual
- Large kernel size to mix distant spatial locations

## Pointwise convolution

- Channel mixing
- 1x1 Convolution



(*d* repetitions)

Model architecture of ConvMixer [4]



[4] Trockman, A., & Kolter, J. Z. (2022). Patches are all you need? arXiv (Cornell University). https://doi.org/10.48550/arxiv.2201.09792

# Results ImageNet1k

ConvMixer-*h*/*d*

| Current "Most Interesting" **ConvMixer** Configurations *vs.* Other Simple Models | | | | | | | |
|---|---|---|---|---|---|---|---|
| Network | Patch Size | Kernel Size | # Params ($\times 10^6$) | Throughput (img/sec) | Act. Fn. | # Epochs | ImNet top-1 (%) |
| ConvMixer-1536/20 | 7 | 9 | 51.6 | 134 | G | 150 | 81.37 |
| ConvMixer-768/32 | 7 | 7 | 21.1 | 206 | R | 300 | 80.16 |
| ResNet-152 | – | 3 | 60.2 | 828 | R | 150 | 79.64 |
| DeiT-B | 16 | – | 86 | 792 | G | 300 | 81.8 |
| ResMLP-B24/8 | 8 | – | 129 | 181 | G | 400 | 81.0 |

Table 1: Models trained and evaluated on $224 \times 224$ ImageNet-1k only. See more in Appendix A.

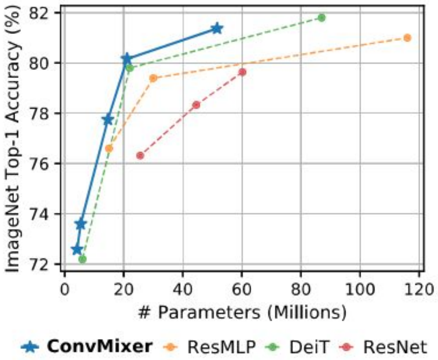



Figure 1: Accuracy *vs.* parameters, trained and evaluated on ImageNet-1k.

Results ConvMixer [4]



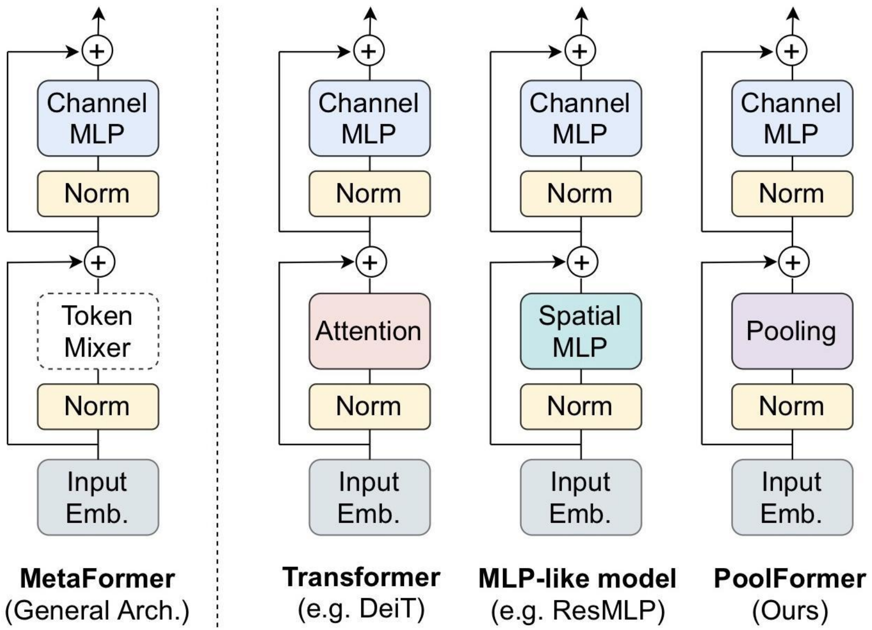
[4] Trockman, A., & Kolter, J. Z. (2022). Patches are all you need? arXiv (Cornell University). https://doi.org/10.48550/arxiv.2201.09792

# The MetaFormer Architecture [6]



[6] Yu, W., Luo, M., Zhou, P., Si, C., Zhou, Y., Wang, X., Feng, J., & Yan, S. (2022). MetaFormer is Actually What You Need for Vision. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr52688.2022.01055

# MetaFormer Architechture




MetaFormer block compared to other Transformer blocks [6]

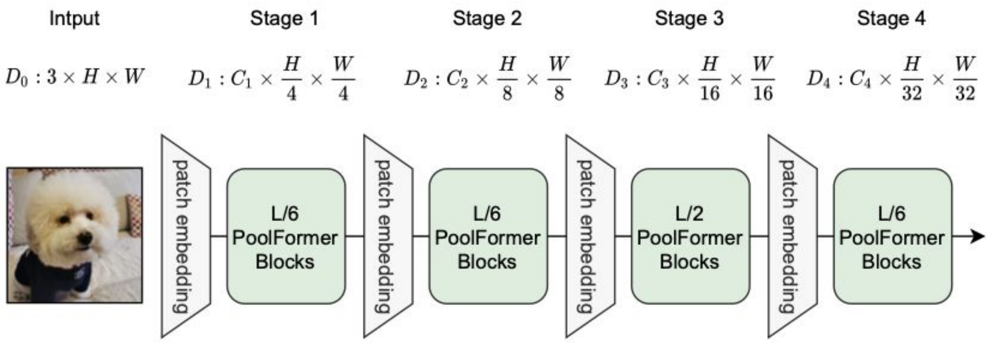
- Success of Transformer architecture attributed to the Attention token mixer
- Other research was focused on making the token mixer more complex or exchange it for another logic
- The MetaFormer architecture is the general shape of the Transformer with variable token mixer
- To test the effectiveness of the MetaFormer architecture PoolFormer was introduced using simple (untrainable) average pooling operation as a token mixer



[6] Yu, W., Luo, M., Zhou, P., Si, C., Zhou, Y., Wang, X., Feng, J., & Yan, S. (2022). MetaFormer is Actually What You Need for Vision. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr52688.2022.01055

# MetaFormer Architechture

- Consists of four stages
- Initial stage reduces input by a factor of 4 consecutive stages by a factor of 2
- Architecture proposed in two embedding sizes S=[64, 128, 320, 512] and M=[96, 192, 384, 768]
- Define L as total amount of PoolFormer blocks split among stages as L/6, L/6, L/2, L/6
- PoolFormer-<Size><L> determines which we are using e.g. PoolFormer-S12 uses small embedding sizes and 12 PoolFormer blocks in total

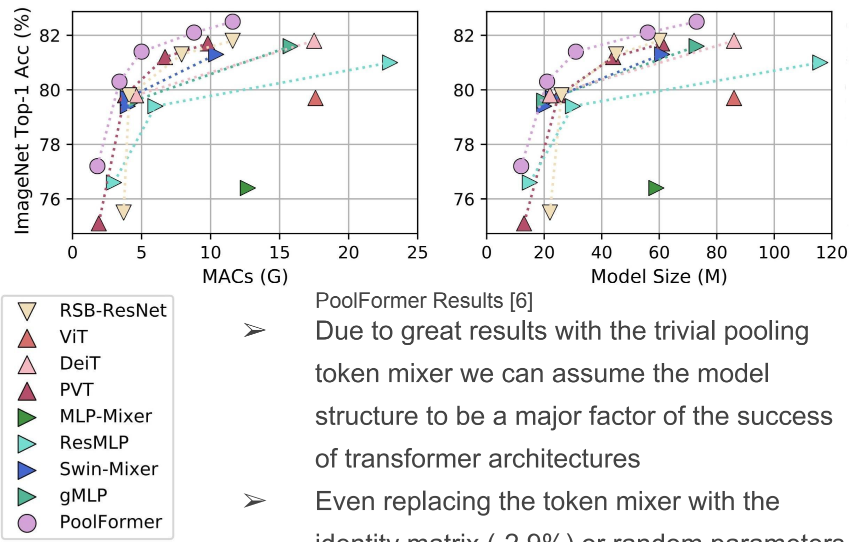



PoolFormer architecture [6]



[6] Yu, W., Luo, M., Zhou, P., Si, C., Zhou, Y., Wang, X., Feng, J., & Yan, S. (2022). MetaFormer is Actually What You Need for Vision. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr52688.2022.01055

# MetaFormer/PoolFormer Paper Results




PoolFormer Results [6]

- Due to great results with the trivial pooling token mixer we can assume the model structure to be a major factor of the success of transformer architectures
- Even replacing the token mixer with the identity matrix (-2.9%) or random parameters (-1.4%) showed good results
- Key to improve transformers further might be in the general structure not in the token mixer

| Model | Accuracy | Parameter | MACs |
|---|---|---|---|
| PoolFormer-S24 | **80.3%** | **21M** | **3.4G** |
| RSB-ResNet-34 | 75.5% | 22M | 3.7G |
| DeiT-S | 79.8% | 22M | 4.6G |
| ResMLP-S24 | 79.4% | 30M | 6.0G |

Results for image classification for comparable model complexities. [6]



[6] Yu, W., Luo, M., Zhou, P., Si, C., Zhou, Y., Wang, X., Feng, J., & Yan, S. (2022). MetaFormer is Actually What You Need for Vision. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr52688.2022.01055

# Our Contribution

# Learning Setups

- We will use 3 data distributions to test our models on, all sampled from BigEarthNet [7]
  - Randomly distributed (low non-IID)
  - Split by country (moderate non-IID)
  - Split by country and season (high non-IID)
- We will test different FL aggregation algorithms selected from the ones tested in [1] on the presented Transformers and a ResNet
  - For now we tested FedAvg
  - Focus on model aggregation strategies
- First results retrieved using only the Serbia subset
- Data was randomly distributed to 3 clients (33%, 33%, 33%)
- FedAvg used for aggregation



[1] Büyüktaş, B., Sümbül, G., & Demir, B. (2023). Federated learning across decentralized and unshared archives for remote sensing image classification. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2311.06141
[7] Sümbül, G., Charfuelàn, M., Demir, B., & Markl, V. (2019). BigEarthNet: A Large-Scale Benchmark Archive for Remote Sensing Image Understanding. IEEE International Geoscience and Remote Sensing Symposium, 5901–5904. https://doi.org/10.1109/igarss.2019.8900532

# Results on BigEarthNet

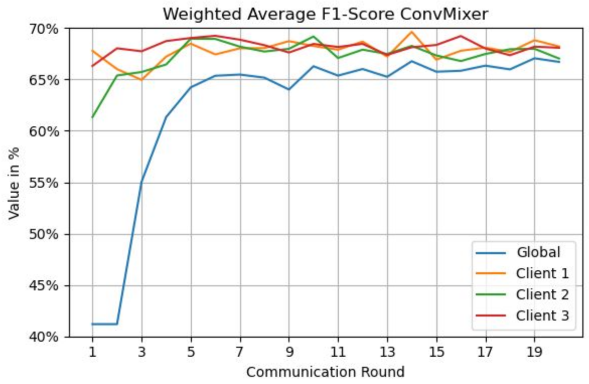
- Scores are measured as weighted average

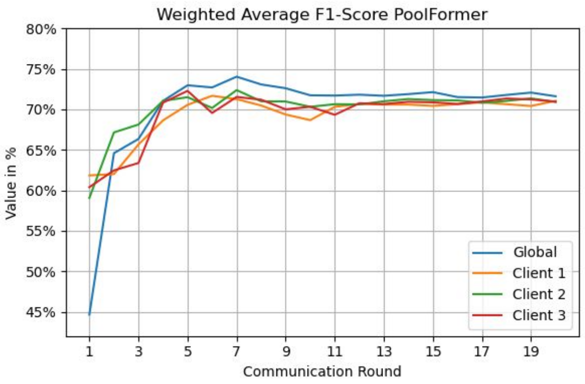
| Model Type | # Parameters | Com. Rounds | Epochs | Training Time (s) | F1-Score | Precision | Recall |
|---|---|---|---|---|---|---|---|
| ResNet18 | **11,239,571** | 20 | 10 | **6268** | **0.74** | **0.78** | **0.74** |
| MLP-Mixer | 59,145,823 | 20 | 10 | 62747[1] | 0.73 | 0.76 | 0.70 |
| ConvMixer -1024/20 | 24,782,867 | 20 | 10 | 7222 | 0.67 | 0.77 | 0.61 |
| PoolFormer -S12 | 11,433,875 | 20 | 10 | 6323 | 0.72 | 0.75 | 0.69 |

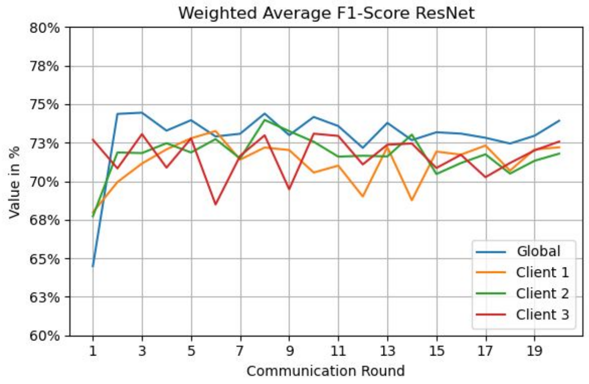


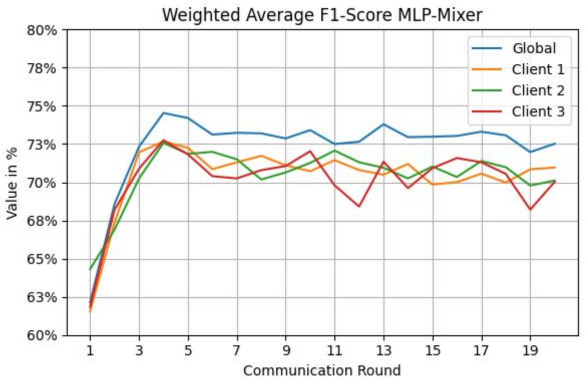
1 CPU Time

# Results on BigEarthNet Global Model vs Client Models



Thank you for your attention!

# Sources

[1] Büyüktaş, B., Sümbül, G., & Demir, B. (2023). Federated learning across decentralized and unshared archives for remote sensing image classification. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2311.06141

[2] Qu, L., Zhou, Y., Liang, P. P., Xia, Y., Wang, F., Adeli, E., Li, F., & Rubin, D. L. (2022). Rethinking architecture design for tackling data heterogeneity in federated learning. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr52688.2022.00982

[3] Tolstikhin, I., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J., Lučić, M., & Dosovitskiy, A. (2021). MLP-Mixer: an all-MLP architecture for vision. arXiv (Cornell University). https://arxiv.org/pdf/2105.01601.pdf

[4] Trockman, A., & Kolter, J. Z. (2022). Patches are all you need? arXiv (Cornell University). https://doi.org/10.48550/arxiv.2201.09792

[5] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv (Cornell University). https://openreview.net/pdf?id=YicbFdNTTy

[6] Yu, W., Luo, M., Zhou, P., Si, C., Zhou, Y., Wang, X., Feng, J., & Yan, S. (2022). MetaFormer is Actually What You Need for Vision. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr52688.2022.01055

[7] Sümbül, G., Charfuelàn, M., Demir, B., & Markl, V. (2019). BigEarthNet: A Large-Scale Benchmark Archive for Remote Sensing Image Understanding. IEEE International Geoscience and Remote Sensing Symposium, 5901–5904. https://doi.org/10.1109/igarss.2019.8900532