Deep Learning Sessions Portugal

Patches are all you need?

Reading group

November 8th 2021

~~Convolutions Attention MLPs~~ Patches Are All You Need? 🤷‍♂️

Anonymous authors
Paper under double-blind review

Abstract

Although convolutional networks have been the dominant architecture for vision tasks for many years, recent experiments have shown that Transformer-based models, most notably the Vision Transformer (ViT), may exceed their performance in some settings. However, due to the quadratic runtime of the self-attention layers in Transformers, ViTs require the use of patch embeddings, which group together small regions of the image into single input features, in order to be applied to larger image sizes. This raises a question: Is the performance of ViTs due to the inherently-more-powerful Transformer architecture, or is it at least partly due to using patches as the input representation? In this paper, we present some evidence for the latter: specifically, we propose the ConvMixer, an extremely simple model that is similar in spirit to the ViT and the even-more-basic MLP-Mixer in that it operates directly on patches as input, separates the mixing of spatial and channel dimensions, and maintains equal size and resolution throughout the network. In contrast, however, the ConvMixer uses only standard convolutions to achieve the mixing steps. Despite its simplicity, we show that the ConvMixer outperforms the ViT, MLP-Mixer, and some of their variants for similar parameter counts and data set sizes, in addition to outperforming classical vision models such as the ResNet. Our code is available at `https://github.com/tmp-iclr/convmixer`.

TL;DR

Deep Learning Sessions Portugal

- 4-page paper @ ICLR 2022;

- Propose **ConvMixer** Architecture
  = convolutions + patch embedding;

- Empirical validation using CIFAR-10 and ImageNet-1k

- Topics:

  Vision Transformers (ViT), Patch Embeddings

Under review as a conference paper at ICLR 2022

~~Convolutions~~ ~~Attention~~ ~~MLPs~~ Patches Are All You Need? 🤷

Anonymous authors
Paper under double-blind review

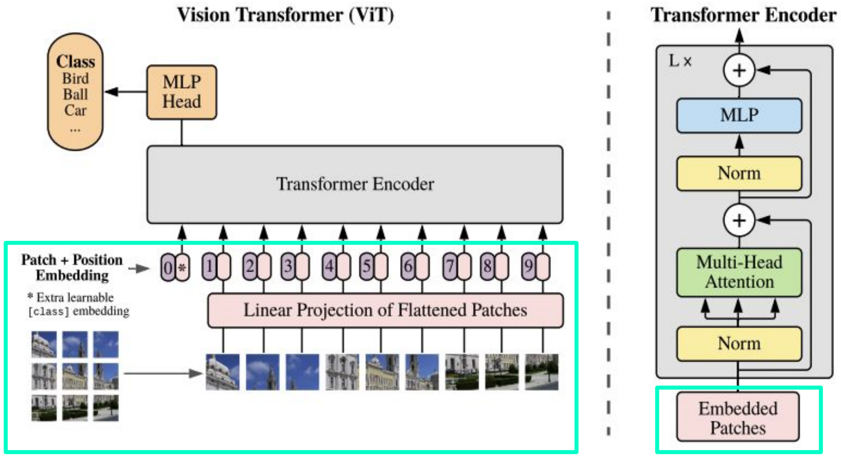
Abstract

Although convolutional networks have been the dominant architecture for vision tasks for many years, recent experiments have shown that Transformer-based models, most notably the Vision Transformer (ViT), may exceed their performance in some settings. However, due to the quadratic runtime of the self-attention layers in Transformers, ViTs require the use of patch embeddings, which group together small regions of the image into single input features, in order to be applied to larger image sizes. This raises a question: Is the performance of ViTs due to the inherently-more-powerful Transformer architecture, or is it at least partly due to using patches as the input representation? In this paper, we present some evidence for the latter: specifically, we propose the ConvMixer, an extremely simple model that is similar in spirit to the ViT and the even-more-basic MLP-Mixer in that it operates directly on patches as input, separates the mixing of spatial and channel dimensions, and maintains equal size and resolution throughout the network. In contrast, however, the ConvMixer uses only standard convolutions to achieve the mixing steps. Despite its simplicity, we show that the ConvMixer outperforms the ViT, MLP-Mixer, and some of their variants for similar parameter counts and data set sizes, in addition to outperforming classical vision models such as the ResNet. Our code is available at `https://github.com/tmp-iclr/convmixer`.

Research Question

Deep Learning Sessions Portugal

Is the performance of ViTs due to the use of patches?

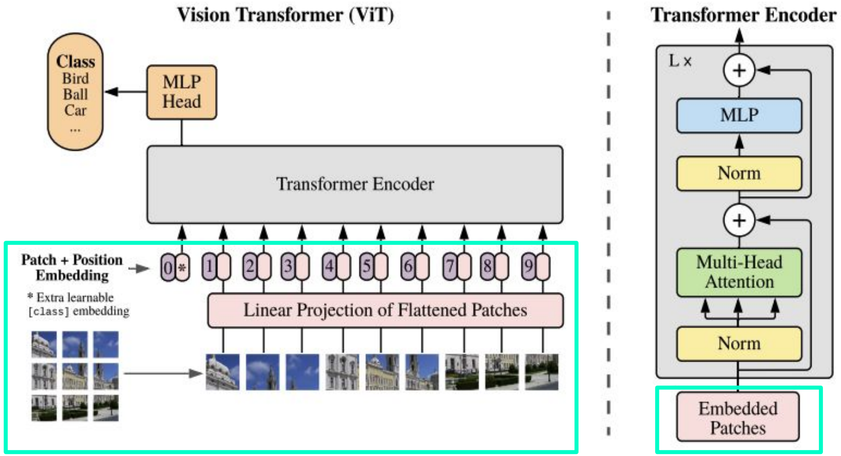


Source: *An image is worth 16x16 words: Transformers for recognition at scale* (ICLR 2021) - Paper, ViT Explained

Research Question

Deep Learning Sessions Portugal

Is the performance of ViTs due to the use of patches?

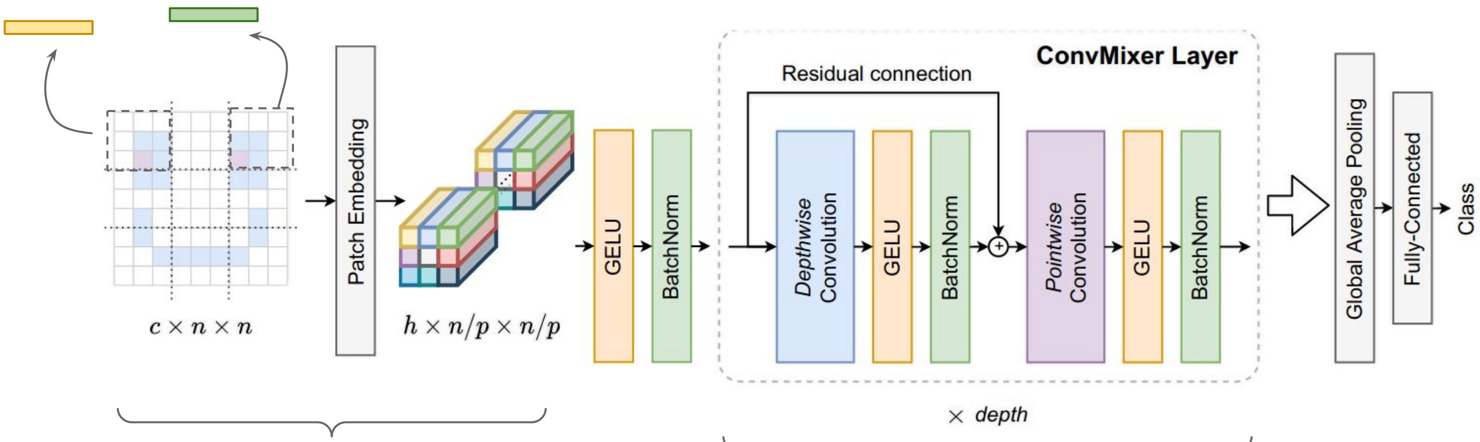


Note:
Interested in knowing more about patch-based vs full-convolution networks? Check [this CVPR 2015 paper](#).

Source: *An image is worth 16x16 words: Transformers for recognition at scale* (ICLR 2021) - [Paper](#), [ViT Explained](#)

ConvMixer Architecture

1. Create patch embeddings;
2. Apply d depth-wise separable convolution blocks.

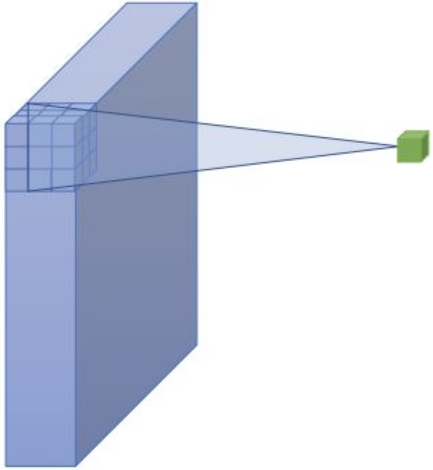


Patch Embedding

d convolution blocks

Note on the convolution block

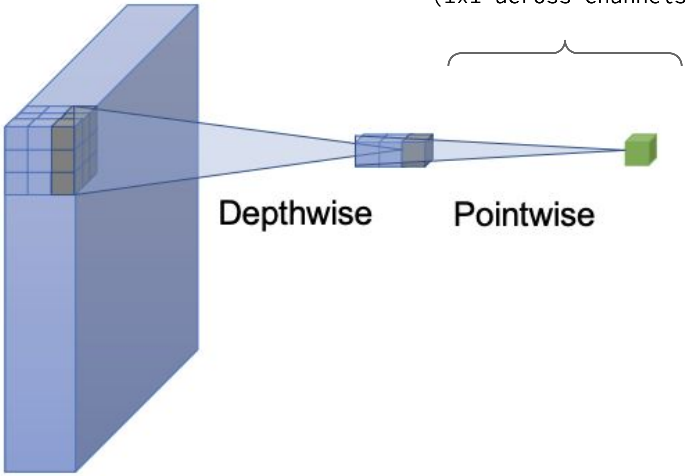
Deep Learning Sessions Portugal



Standard Convolution

Depthwise convolution
(pxp 1 channel)

Pointwise convolution
(1x1 across channels)



Depth-wise separable convolution

Source: Image retrieved from papers with code, online accessed 8th November 2021. Also check this blogpost for more information on the differences.

Differences to MLP Mixer and ViT

Deep Learning Sessions Portugal

ConvMixer is similar to **MLP-Mixer**. MLP-Mixer separates mixing of spatial and channel dimensions, by applying an MLP across spatial dimension and then an MLP across the channel dimension (spatial MLP replaces the ViT attention and channel MLP is the FFN of ViT).

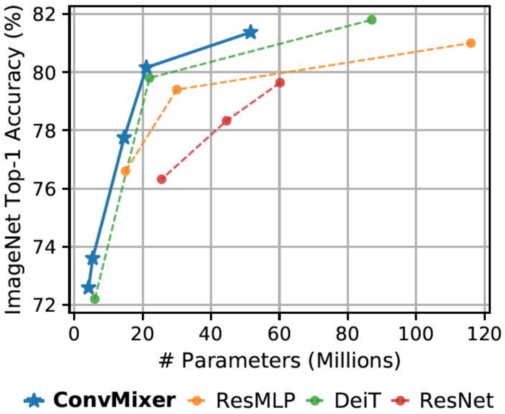
ConvMixer uses a 1x1 convolution for channel mixing and a depth-wise convolution for spatial mixing. Since it's a convolution instead of a full MLP across the space, it **mixes only the nearby batches** in contrast to ViT or MLP-Mixer. Also, the MLP-mixer uses MLPs of two layers for each mixing and **ConvMixer uses a single layer for each mixing**.

The paper recommends **removing the residual connection** across the channel mixing (point-wise convolution) and having **only a residual connection over the spatial mixing** (depth-wise convolution). They also use **Batch normalization** instead of Layer normalization.

Source: Summary retrieved from github, online accessed on the 8th November 2021.

Results @ ImageNet 1k

Deep Learning Sessions Portugal



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	89	G	150	81.37
ConvMixer-768/32	7	7	21.1	203	R	300	80.16
ResNet-152	–	3	60.2	872	R	150	79.64
DeiT-B	16	–	86	703	G	300	81.8
ResMLP-B24/8	8	–	129	140	G	400	81.0

**ConvMixer-h/d* represents the ConvMixer model trained with path embedding dimension h and d convolution blocks.

Trained using timm framework with several data augmentation strategies.
(e.g., RandAugment, mixup, CutMix, random erasing, gradient norm clipping, timm augmentation)

Good Peer Reviewer

Deep Learning Sessions Portugal

Good Peer Reviewer

- Simple, well written, straight to the point;
- Evaluate some of the best SotA models;
- Relevant research topic;

- Raises awareness to the importance of isolating the benefits of different components of complex architectures.

Bad Peer Reviewer

Bad Peer Reviewer

- Writing style may come across as *rude*;
- Main design decisions based on CIFAR-10;
- Questionable performance comparison:
  - Throughput is a concern -- Limited applicability in practice.
  - No Free Lunch Theorem (NFLT) -- No guarantees optimal training for ConvMixer is also optimal for ResNET or DeIT (more on this this paper and this one).
- Hyperparameter Optimization could be improved (why don't they use Hyperband?)

Developer

Developer

PyTorch Code:
https://github.com/tmp-iclr/convmixer

Is this Ross Wightman's work?

Papers w/ code:
https://paperswithcode.com/paper/patches-are-all-you-need

LabAI.ml implementation:
https://nn.labml.ai/conv_mixer/experiment.html

Torch2TF:
https://github.com/Rishit-dagli/ConvMixer-torch2tf

Journalist

Journalist

IntelArtiGen · 1m

Is the paper only interesting because of the new method and its simplicity? It's more parameter efficient (so is efficientnet) but the throughput is terrible

10 Reply Share Report Save

Imnimo · 1m

I think it's interesting as an ablation experiment - I don't think the goal of the paper is to say "this is a good new architecture that people should use", it's "the fact that this simple architecture works helps us narrow down what features of other models are most valuable".

14 Reply Share Report Save

machinelearner77 · 1m

I think you are right with the ablation, but not fully correct with the rest.

Instead, I think the paper questions whether ViTs are so strong *because* they are ViTs or *because* they simply see other input...

2 Reply Share Report Save

Source: In subreddit MachineLearning, online accessed 8 November 2021.

Journalist

IntelArtiGen · 1m

Is the paper only interesting because of the new method and its simplicity? It's more parameter efficient (so is efficientnet) but the throughput is terrible

10 Reply Share Report Save

Imnimo · 1m

I think it's interesting as an ablation experimen
is a good new architecture that people should u
works helps us narrow down what features of c

14 Reply Share Report Save

machinelearner77 · 1m

I think you are right with the ablation, but n

Instead, I think the paper questions whethe
because they simply see other input...

2 Reply Share Report Save

CyberDainz · 1m

like for "all you need" in the title.

7 Reply Share Report Save

EyedMoon · 1m

All you need papers are all you need

13 Reply Share Report Save

Competitive_Dog_6639 · 1m

"All you need" is actually not all you need anymore, if you check the pdf you will see you now need emojis too 🙃

3 Reply Share Report Save

ai_who_found_love · 27d

tbf if you want to make some noise in this space, you need to have some attitude ;)

4 Reply Share Report Save

Competitive_Dog_6639 · 27d

Agreed, although memes alone are probably a shallow local minimum in the loss surface of good ideas lol... no shade to this paper, just speaking in general

1 Reply Share Report Save

Source: In subreddit MachineLearning, online acces

Journalist

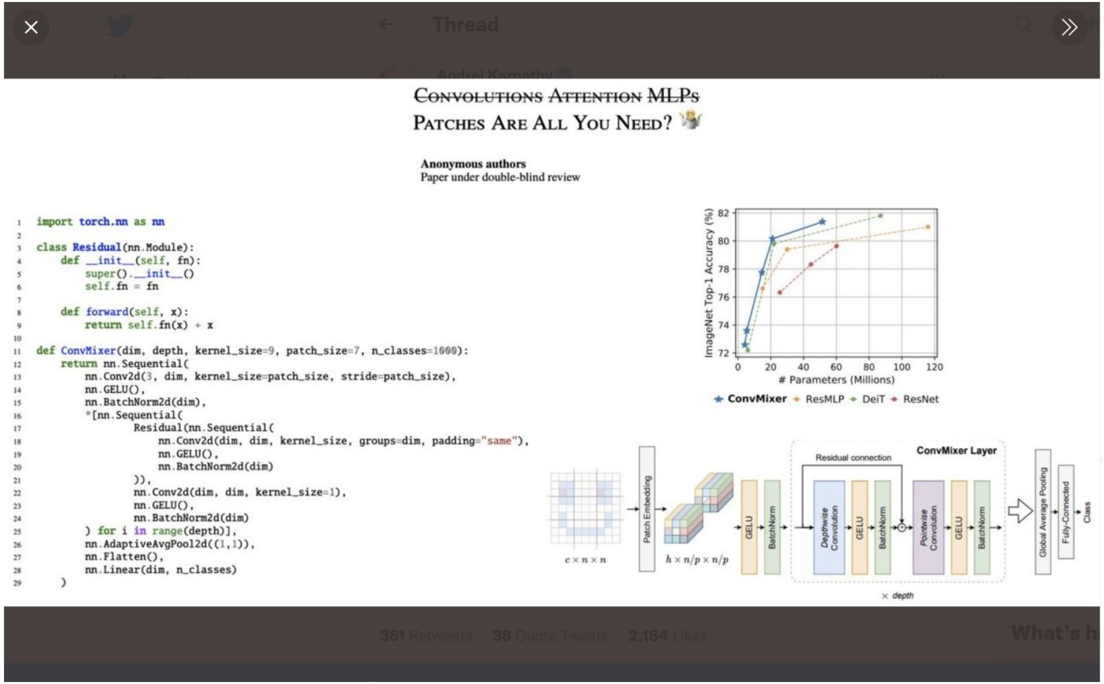
Deep Learning Sessions Portugal

This trend of exploring MLPs and the connection between the strengths of ViT and convolutions dates back to (at least) NeuRIPS 2021...

Source: In Twitter labml.ai, online accessed 8 November 2021.

Journalist

Deep Learning Sessions Portugal



Errr ok wow, I am shook by the new ConvMixer architecture openreview.net/forum?id=TVHS5... "the first model that achieves the elusive dual goals of 80%+ ImageNet top-1 accuracy while also fitting into a tweet" 😐

5:53 pm · 6 Oct 2021 · Twitter Web App

361 Retweets **38** Quote Tweets

2,184 Likes

Andrej Karpathy @ · 6 Oct

Replying to @karpathy

The simplicity and isotropy of the model is aesthetically appealing, by crushing space all at once at start. A bit like refactoring all of the pooling operations in a MobileNet by sliding them all the way down.

5 6 177

Source: In Twitter Andrej Karpathy, online accessed 8 November 2021.

Journalist

Nicholas Guttenberg @ngutten · 6 Oct
Replying to @karpathy
It looks like a big part of this comes from the data augmentation methods and gradient norm clipping (Table 3). The difference between patches and ResNet looks smaller than patches with/without random scaling...

10

Dmitry @RespectToX · 6 Oct
Replying to @karpathy
Resnets: downsampling is effective, but adds noise and breaks shift equivariance. So, get rid of downsampling, except the first layer (where it is least harmful). Cool, but it is 10x slower than resnet. Next step is 100% strideless model, magnitudes slower, but even simpler

3

xhlulu @xhluu · 6 Oct
Replying to @karpathy
I'm curious how much of the improvement is in the architecture vs the improved training tools thanks to timm, considering Resnet-50 can achieve 80%+ without any architecture change: arxiv.org/pdf/2110.00476...

2 16

Source: In Twitter Andrej Karpathy, online accessed 8 November 2021.

Journalist

Deep Learning Sessions Portugal

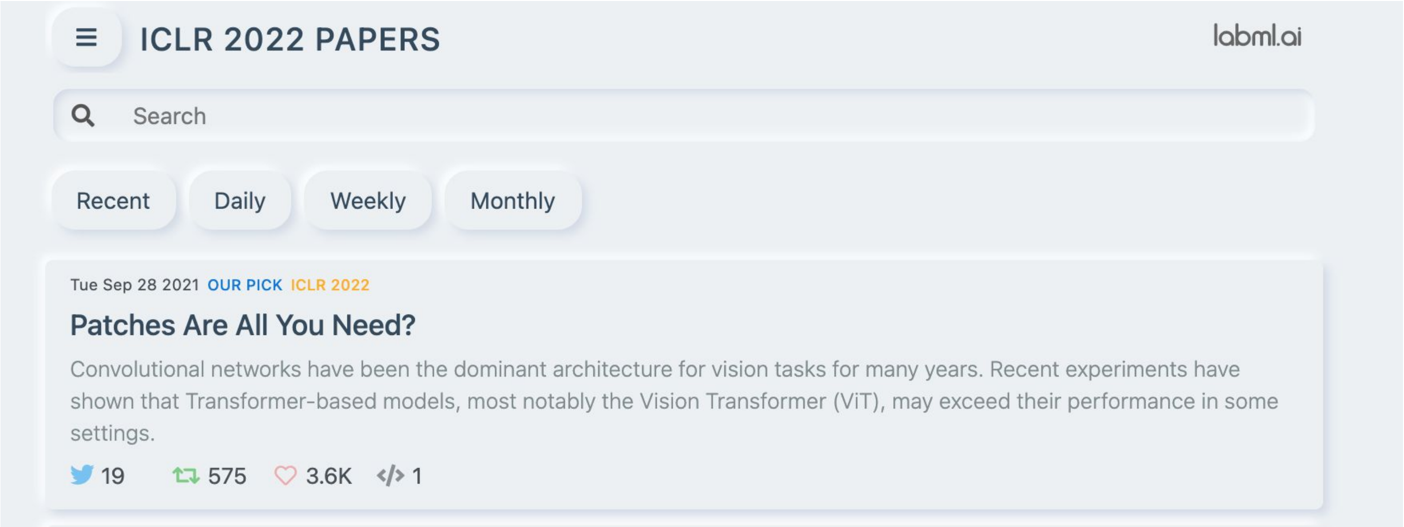
There are so many tweets… Here are some examples:

- Horace He @ Facebook (cHHillee)

Source: In Twitter Andrej Karpathy, online accessed 8 November 2021.

Journalist

Deep Learning Sessions Portugal



Source: In labml.ai selection of ICLR 2022 papers, online accessed 8 November 2021.

Archeologist

Archeologist

Deep Learning Sessions Portugal

Background:

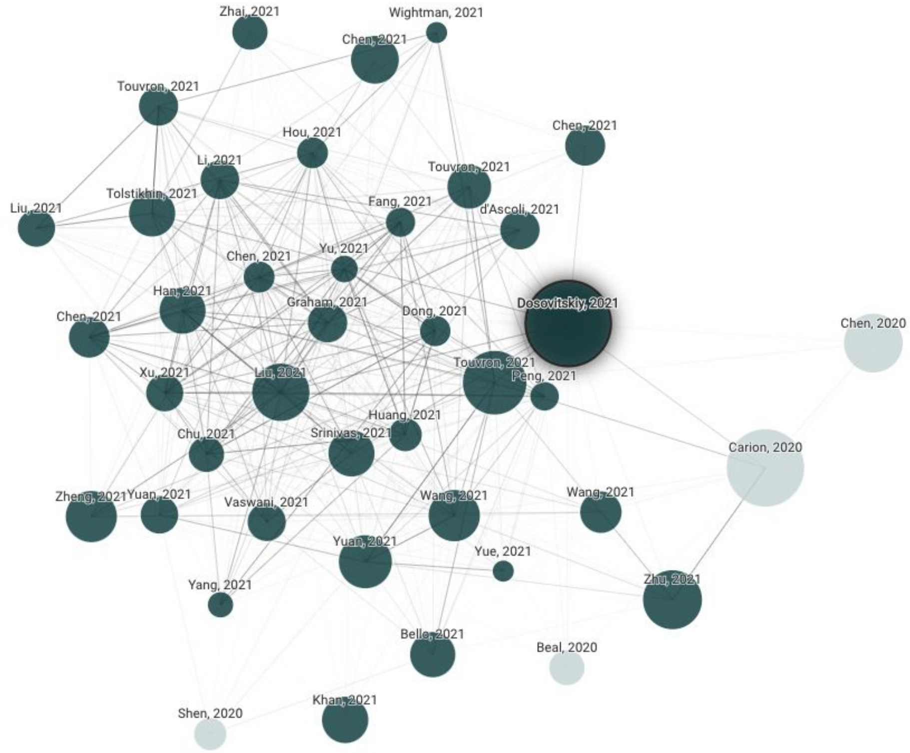
Attention is all you need (2017) - Intro to Transformers

Image is worth 16x16 words (2021) - Intro to ViTs

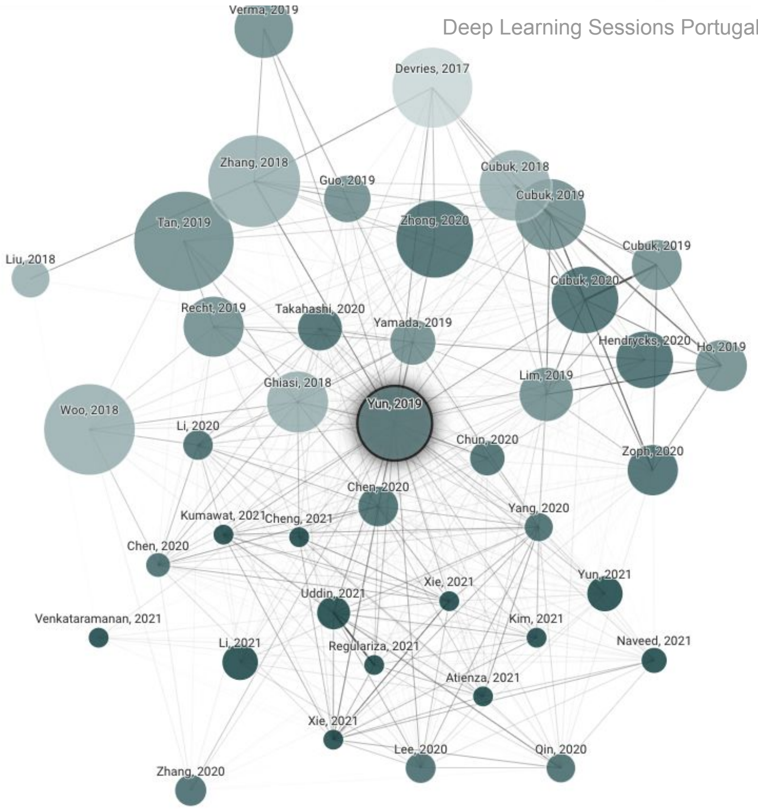
- **2019 and 2020 references:** Attention + Convolution, or training strategies;
- **2021 references** focused on deconstructing ViT, *e.g.*, MLP-Mixer, ResMLP, PVT, CycleMLP, SwinTransformer, etc.

Archeologist

Deep Learning Sessions Portugal



ViT (@ICLR 2021) similarity graph
@ ConnectedPapers.



CutMix (@CVPR 2019) similarity graph
@ ConnectedPapers.

Thank you!

Deep Learning Sessions Portugal

Follow us on Twitter, Meetup, and LinkedIn!

If you liked this event, feel free to **bring a friend** with you next time :)

Please share with us your opinion on papers you'd love to discuss in upcoming sessions!