An Image is Worth 16x16 Words

Transformers for Image Recognition at Scale

Deep Learning Sessions Lisboa (DLSL)
Reading Group
28th June 2021

Meet & Greet!

(name, background, motivation/interest in paper)

The paper

(summary of the paper)

TLDR;

Published as a conference paper at ICLR 2021

AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

Alexey Dosovitskiy[*,†], **Lucas Beyer**[*], **Alexander Kolesnikov**[*], **Dirk Weissenborn**[*], **Xiaohua Zhai**[*], **Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby**[*,†]

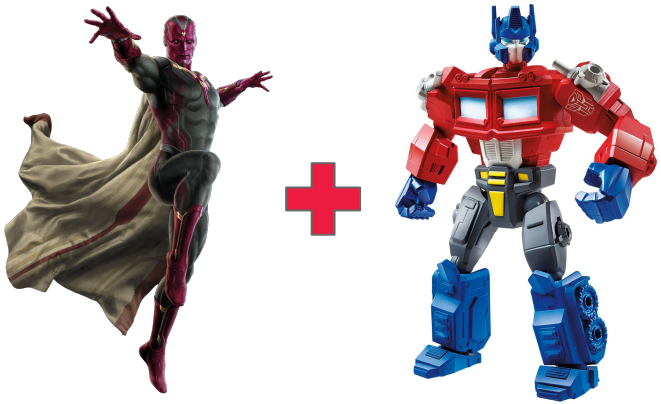
[*]equal technical contribution, [†]equal advising

Google Research, Brain Team

{adosovitskiy, neilhoulsby}@google.com

ABSTRACT

While the Transformer architecture has become the de-facto standard for natural language processing tasks, its applications to computer vision remain limited. In vision, attention is either applied in conjunction with convolutional networks, or used to replace certain components of convolutional networks while keeping their overall structure in place. We show that this reliance on CNNs is not necessary and a pure transformer applied directly to sequences of image patches can perform very well on image classification tasks. When pre-trained on large amounts of data and transferred to multiple mid-sized or small image recognition benchmarks (ImageNet, CIFAR-100, VTAB, etc.), Vision Transformer (ViT) attains excellent results compared to state-of-the-art convolutional networks while requiring substantially fewer computational resources to train.[1]



Keywords

- Transformer
- Computer Vision
- CNN
- Attention
- Vision Transformer (ViT)

Research Question

If Transformers are so successful, can we tackle CV problems using them?

→ **Approach:**

"(...) experiment with applying a standard Transformer directly to images, with the fewest possible modifications."

Research Question

If Transformers are so successful, can we tackle CV problems using them?

→ **Approach:**

"(...) experiment with applying a standard Transformer directly to images, with the fewest possible modifications."

→ **Motivation:**

Achieve better results than CNN's by removing the translation equivariance and locality biases

Research Question

If Transformers are so successful, can we tackle CV problems using them?

→ **Approach:**

"(...) experiment with applying a standard Transformer directly to images, with the fewest possible modifications."

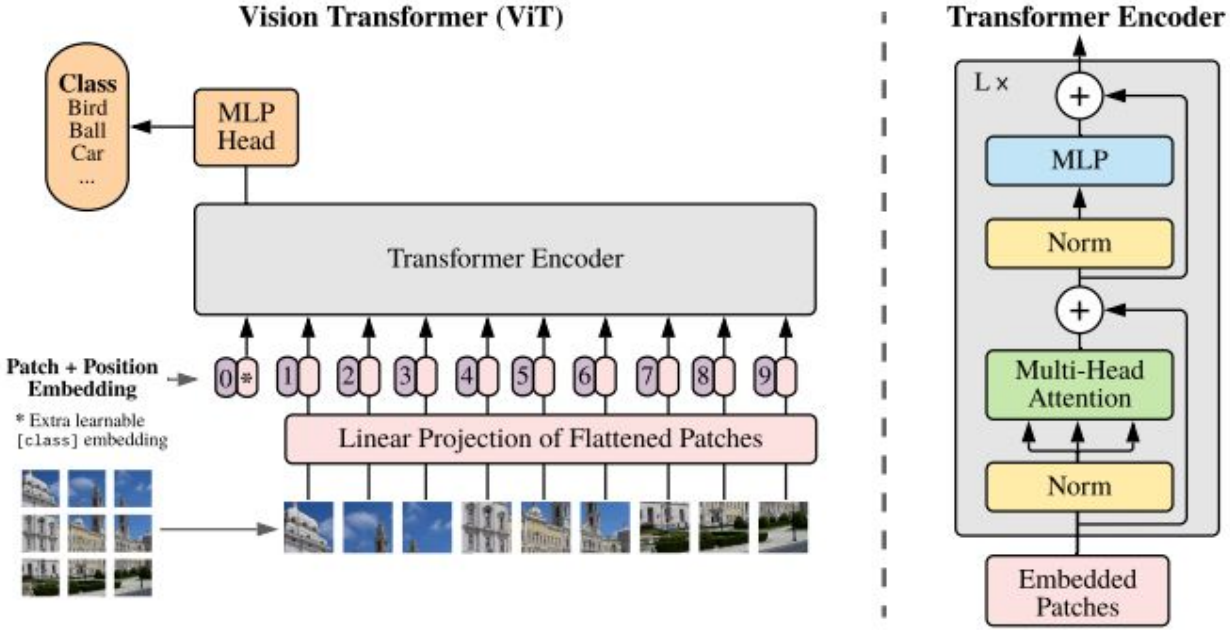
→ **Motivation:**

Achieve better results than CNN's by removing the translation equivariance and locality biases

→ **Setting:**

"(...) train the [Transformer] model on image classification in supervised fashion."

Today's paper



Reflection

A few questions

? Do we really want to exchange bias for data? Does this mean “bye bye Convs”? Data sometimes is not available and train with a lot of data takes more computational resources.

? Can a small company train and use these? Or do these mean more faith on the good will of big companies to share pretrained models?

? Are Tranformers the new “one-architecture-fits-all”? Why are transformers so powerful?

? Are these really computationally more efficient (both in training and prediction time)? What about space efficiency are these suited for edge computing?

Archeologist

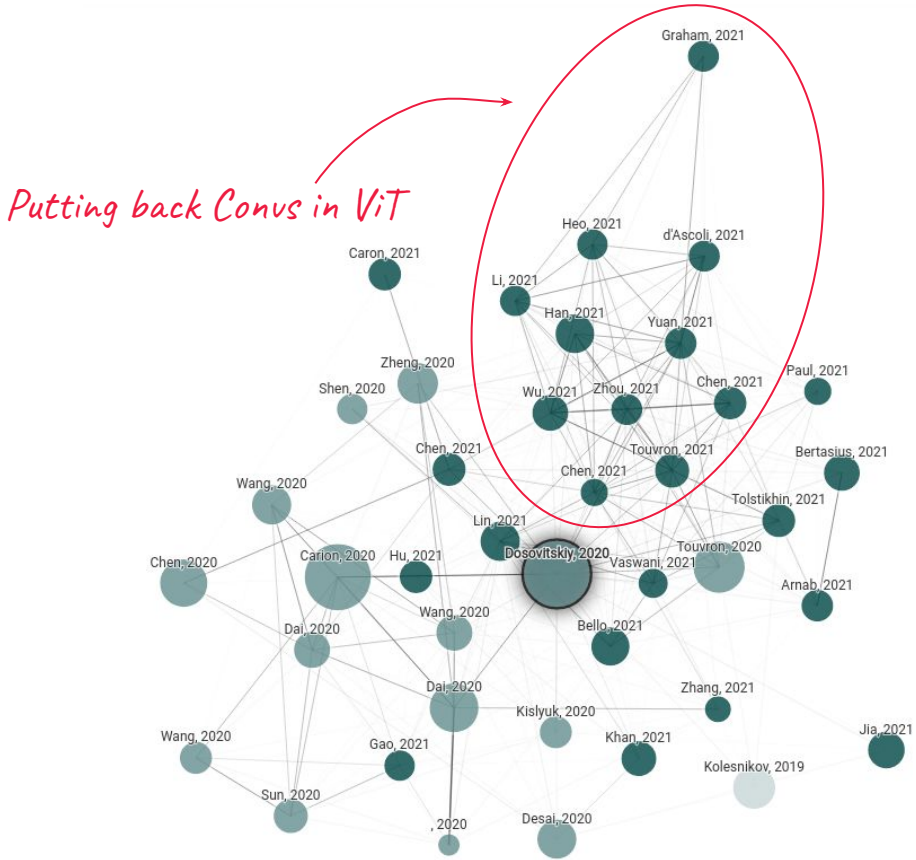
(extract knowledge out of the references...)

Digging the references

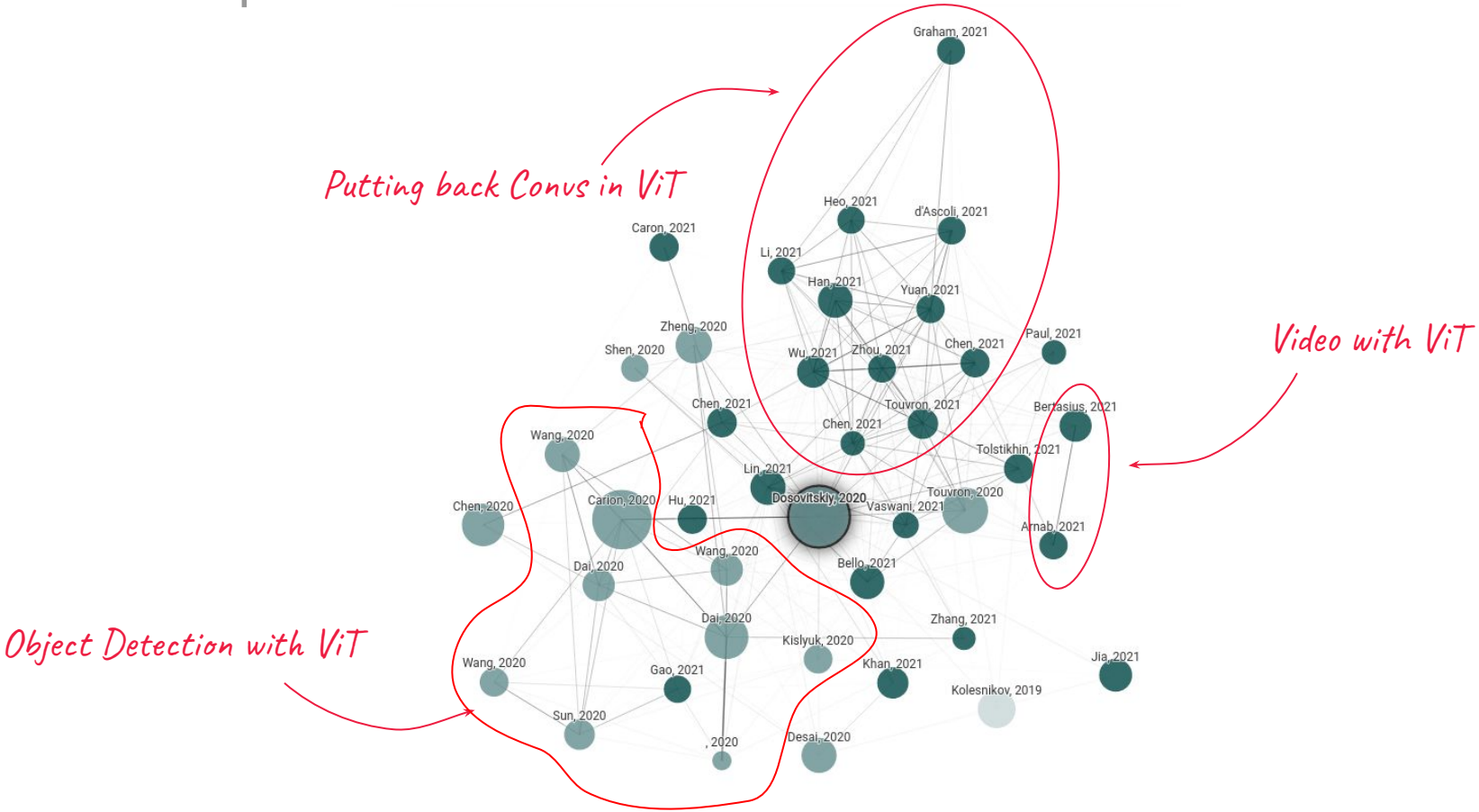
- Technical paper
  (dates back to 22nd of October of 2020);
- Comprises 54 references and 401 citation.
- References spread mostly across 2 different topics:

Transformers	Computer Vision (CV)
Seminal paper on Transformers (Vaswani *et al.*, 2017)	Classical Achitecture for CV
	Attention in CNN's
Applications in NLP domain (text generation, language models, language understanding)	Datasets for Image Recognition

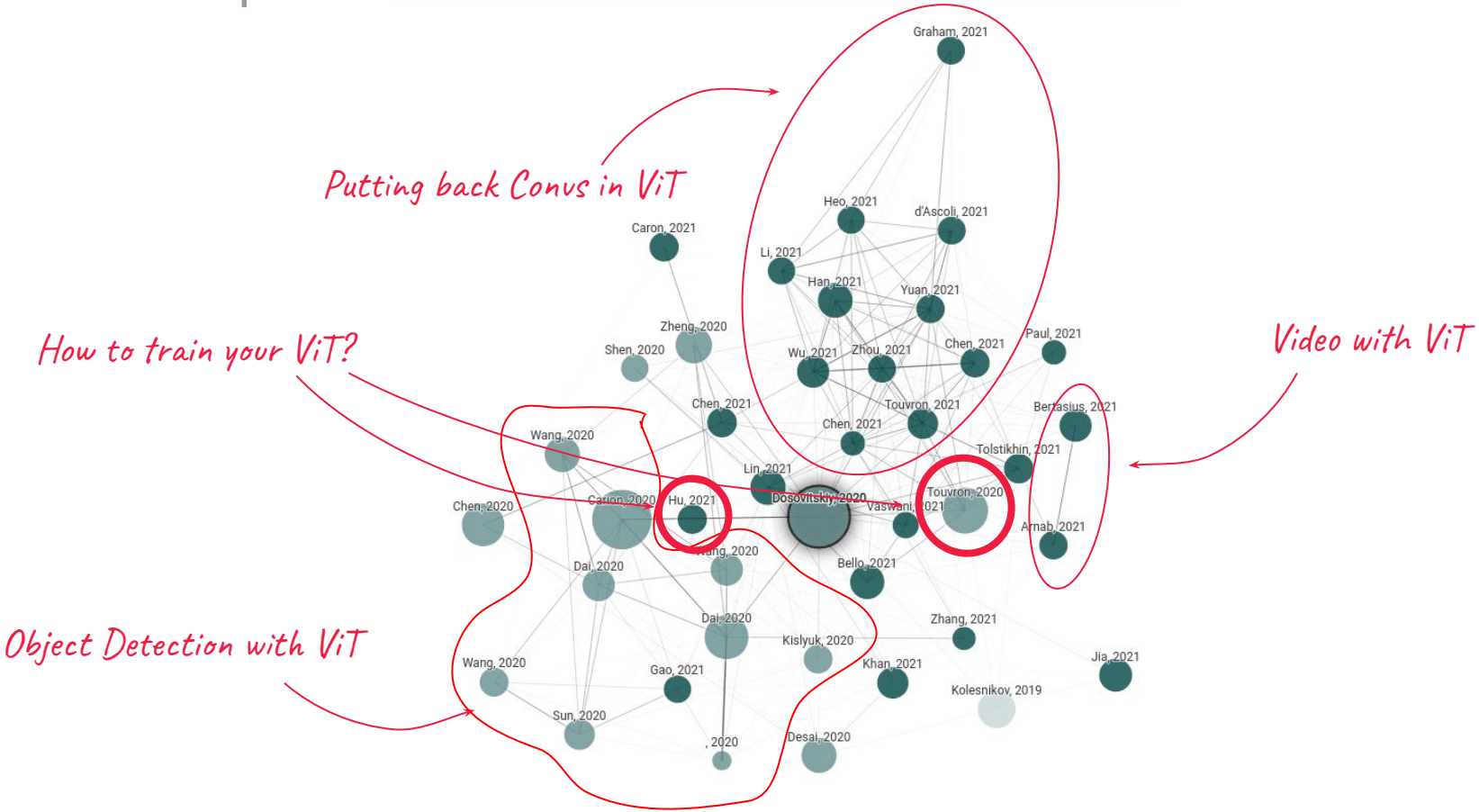
ConnectedPapers - Vision Transformer [1]



ConnectedPapers - Vision Transformer [1]



ConnectedPapers - Vision Transformer [1]



Digging the paper's references - Most influential

- **Deep Residual Learning for Image Recognition** ***(2015)***
  Deep Residual Networks' (ResNets) seminal paper. They present the Residual Block layer which consists of an usual convolutional layer with an additional skip (or residual) connection, which allow shortcuts and makes training lighter.

- **Attention is all you need** ***(2017)***
  Transformers' seminal paper. Despite not being fundamental for reading today's paper, reading it will provide you with the details on the architecture and the overall mechanisms associated.

- **Generative Pretraining from Pixels** ***(2020)***
  Applying GPT (a tranformer) model to image completion task. An example of using transformers for image related tasks before the paper hereby discussed.

Developer

(check for code, tutorials, implementations, ...)

Related Code Resources

- Paper repo: repo provided by the original authors
- Papers with Code Page: collection of related code resources
- Colab: example of how to use pre-trained ViT
- HuggingFace ViT Docs: docs about ViT HF implementation
- Implementing ViT from scratch: video on how to implement ViT in Pytorch
- Paper on “How to train your ViT?”: paper describing methods to train a ViT

Entrepreneur

(come up w/ ideas for applications/products and pitch them)

ViT in the Industry

Open source ViT

```
>>> from transformers import ViTFeatureExtractor, ViTForImageClassification
>>> from PIL import Image
>>> import requests

>>> url = 'http://images.cocodataset.org/val2017/000000039769.jpg'
>>> image = Image.open(requests.get(url, stream=True).raw)

>>> feature_extractor = ViTFeatureExtractor.from_pretrained('google/vit-base-patch16-224')
>>> model = ViTForImageClassification.from_pretrained('google/vit-base-patch16-224')

>>> inputs = feature_extractor(images=image, return_tensors="pt")
>>> outputs = model(**inputs)
>>> logits = outputs.logits
>>> # model predicts one of the 1000 ImageNet classes
>>> predicted_class_idx = logits.argmax(-1).item()
>>> print("Predicted class:", model.config.id2label[predicted_class_idx])
```

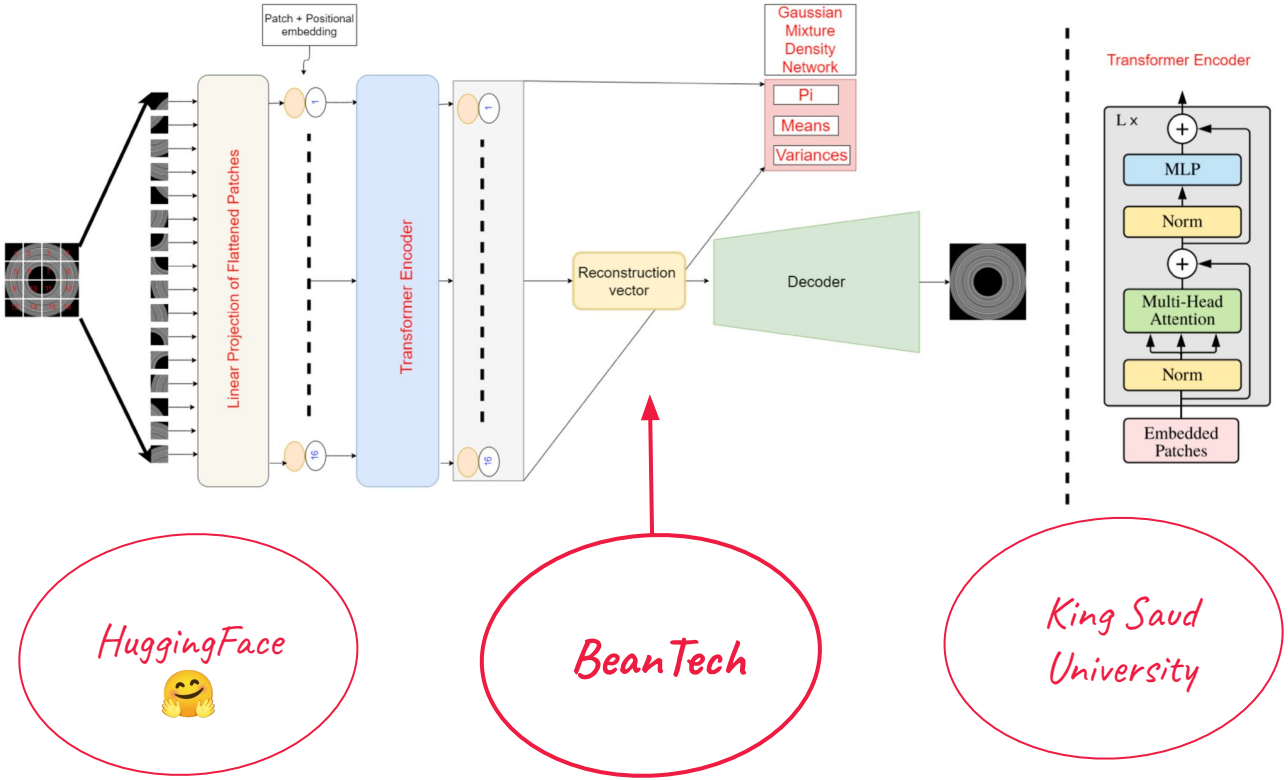
HuggingFace

BeanTech

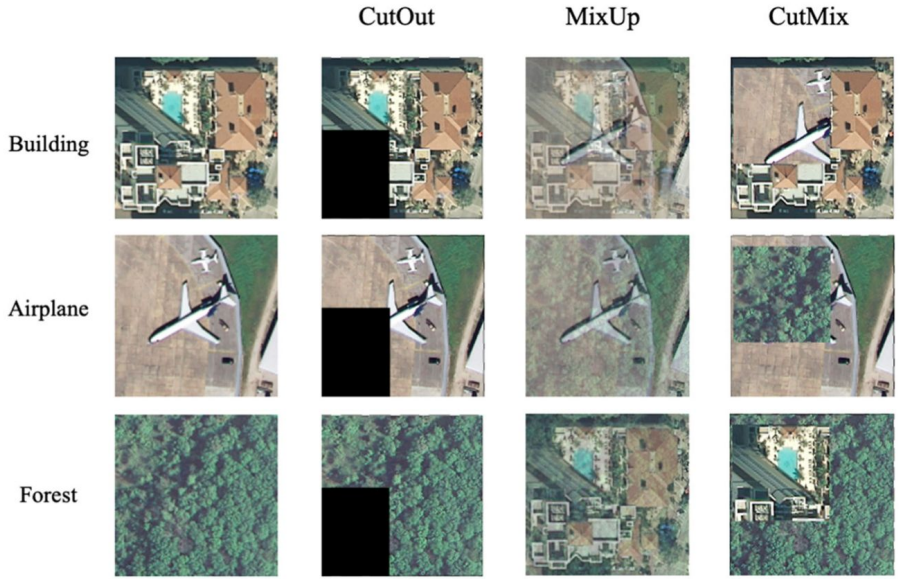
King Saud University

ViT in the Industry

Anomaly detection



ViT in the Industry



[Remote-sensing scene-classification]

HuggingFace 🤗

BeanTech

King Saud University

Journalist

(check authors backgrounds, impact of the paper in the media, gossip, ...)

Who are the authors? (1)

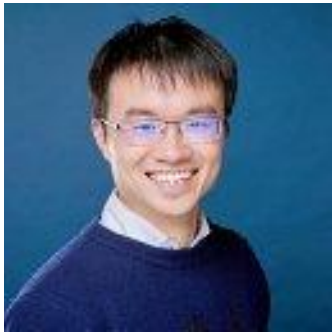
* advising

Alexey Dosovitskiy*
(Google Brain)
Alexey Dosovitskiy - Google Scholar

Lucas Beyer
(Google Brain)
Lucas Beyer (@giffmana) / Twiter
Lucas Beyer - Google Scholar

Alexander Kolesnikov
(Google Brain)
Alexander K. (@__kolesnikov__) / Twitter
Alexander Kolesnikov - Google Scholar

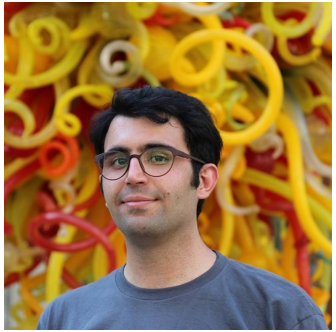
Dirk Weissenborn
(Google)
Dirk W. (@dirkweissenborn) / Github
Dirk W. - Google Scholar



Xiaohua Zhai
(Google Brain)
Xiaohua Zhai - Google Scholar

Who are the authors? (2)

Thomas Unterthiner
(Google Brain)
Thomas_Unterthiner - Google Scholar



Mostafa Dehghani
(Google Brain)
Thomas Unterthiner - Google Scholar

Matthias Minderer
(Google Brain)
Matthias Minderer - Google Scholar

Georg Heigold
(Google Brain)
Georg Heigold - Google Scholar

Who are the authors? (3)

* advising

Sylvain Gelly

(Google Brain)

Sylvain Gelly - Google Scholar

Jakob Uszkoreit

(Google Brain)

Jakob Uszkoreit - Google Scholar

Neil Houlsby *

(Google Brain)

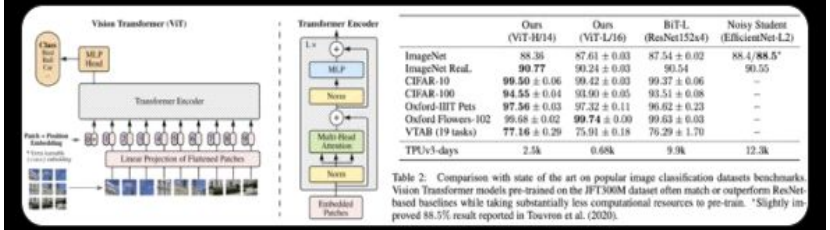
Neil Houlsby - Google Scholar

Hot Discussions on Social Media

Is this the end for convolutions?

<source>

An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale openreview.net/forum?id=YicbF... v cool. Further steps towards deprecating ConvNets with Transformers. Loving the increasing convergence of Vision/NLP and the much more efficient/flexible class of architectures.



	Ours (ViT-H/14)	Ours (ViT-L/16)	BiT-L (ResNet152x4)	Noisy Student (EfficientNet-L2)
ImageNet	88.36	87.61 ± 0.03	87.54 ± 0.02	88.4/**88.5***
ImageNet ReaL	**90.77**	90.24 ± 0.03	90.54	90.55
CIFAR-10	**99.50** ± 0.06	99.42 ± 0.03	99.37 ± 0.06	–
CIFAR-100	**94.55** ± 0.04	93.90 ± 0.05	93.51 ± 0.08	–
Oxford-IIIT Pets	**97.56** ± 0.03	97.32 ± 0.11	96.62 ± 0.23	–
Oxford Flowers-102	99.68 ± 0.02	**99.74** ± 0.00	99.63 ± 0.03	–
VTAB (19 tasks)	**77.16** ± 0.29	75.91 ± 0.18	76.29 ± 1.70	–
TPUv3-days	2.5k	0.68k	9.9k	12.3k

Table 2: Comparison with state of the art on popular image classification datasets benchmarks. Vision Transformer models pre-trained on the JFT300M dataset often match or outperform ResNet-based baselines while taking substantially less computational resources to pre-train. *Slightly improved 88.5% result reported in Touvron et al. (2020).

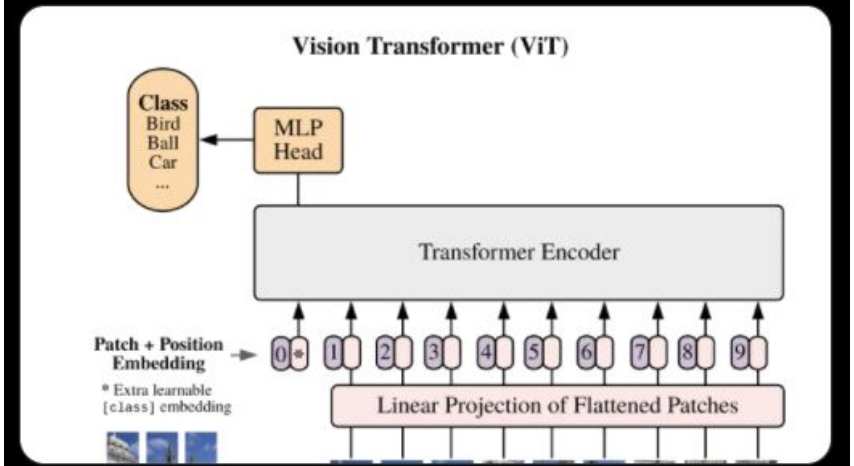
8:32 AM · Oct 3, 2020 · Twitter Web App

507 Retweets **61** Quote Tweets **2,035** Likes

<source>

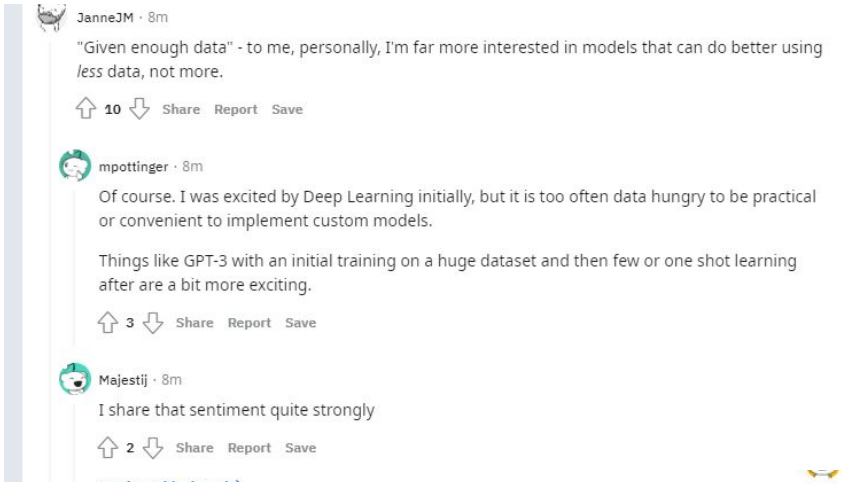
Recent conversation with a friend:

@ilyasut: what's your take on openreview.net/pdf?id=YicbFdN...?
@OriolVinyalsML: my take is: farewell convolutions :)



11 227 923

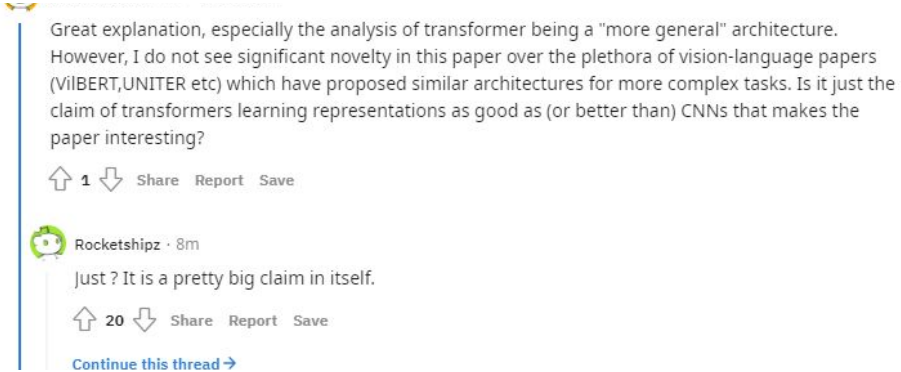
Hot Discussions on Social Media



Why do we need models that require more data?

<source>

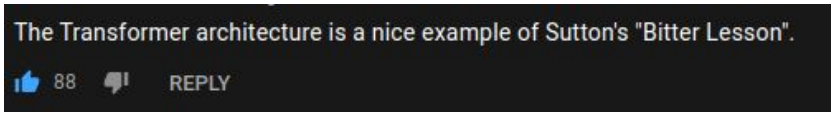
Is this paper "just" about a general model learning properties that were previously built-in?



<source>

Hot Discussions on Social Media

Is this an example of "The Bitter Lesson"[2], i.e., given more power, must we remove the bias to improve further?



<source>

Other Hot Discussions on Social Media

Will pre-trained models of this and following architectures be provided for open access?

How to implement double-blind review when resources are only available to some people?

Are Transformers becoming a universal model?

Pointers

Youtube

- An Image is Worth 16x16 Words [Yannic Kilcher] - https://www.youtube.com/watch?v=TrdevFK_am4

Medium

- https://medium.com/nerd-for-tech/an-image-is-worth-16x16-words-transformers-for-image-recognition-at-scale-paper-summary-3a387e71880a
- https://medium.com/swlh/an-image-is-worth-16x16-words-transformers-for-image-recognition-at-scale-brief-review-of-the-8770a636c6a8

Twitter

- https://twitter.com/ykilcher/status/1312718227953405952

Reddit

- https://www.reddit.com/r/MachineLearning/comments/j4xmht/d_paper_explained_an_image_is_worth_16x16_words/

Good Peer Reviewer

(find the strengths of the paper, the most valuable aspects, ...)

Positive Aspects of the Paper ↑

- ✔ In spite of using a lot of resources, this paper achieves state-of-the-art results for image classification on multiple relevant datasets, while being least expensive to train (compared to previous benchmarks).

- ✔ Great visualization of internal representations in the Vision Transformer. Very important for explainability.

Mean Peer Reviewer

(find the cons/weaknesses of the paper)

Downsides of the Paper ↓

- × Not tested in tasks other than classification, such as segmentation and object detection.
- × No mentions to the downsides of this approach, only that they don't beat CNN's when trained on little data.
- × The proposed model do not completely exclude the inductive bias of CNN's, and even proposes an hybrid model, without explaining why we might want to put convolutions back on if Transformers seem to reach better results with lower computational consumption?
- × Data thirsty models (and the paradigm of big pre-trained models) bring some economical and ethical questions about the democratization of AI, that are not even mentioned in the paper.

Wrap-up

(discussion; come up w/ ideas for titles)

Summary

Vision Transformer casts image recognition as a sequence problem, removing both the locality and translation invariance biases we have in CNN's. They also remove the 2D (height x width) structure of images.

Besides the well known Bias vs. Variance tradeoff, we see here represented new tradeoffs: Bias vs. Data / Computational Power. If we have enough representative data and enough power, then a general model probably will be able to learn the biases.

These models outperform biased models (CNN's) but only when there's enough data and computational power available. Which might mean that, if both data and computational power continue to become more and more available, then in order to make models better, we might need to remove bias (Sutton's Bitter Lesson).

Some resources…

Paper;

Paper review (by Yannick's Kilcher);

Last papers

Decision Transformer

Decision Transformer: Reinforcement Learning via Sequence Modeling

Lili Chen[*,1], **Kevin Lu**[*,1], **Aravind Rajeswaran**[2], **Kimin Lee**[1],
Aditya Grover[2], **Michael Laskin**[1], **Pieter Abbeel**[1], **Aravind Srinivas**[†,1], **Igor Mordatch**[†,3]
[*]equal contribution [†]equal advising
[1]UC Berkeley [2]Facebook AI Research [3]Google Brain
{lilichen, kzl}@berkeley.edu

Abstract

We introduce a framework that abstracts Reinforcement Learning (RL) as a sequence modeling problem. This allows us to draw upon the simplicity and scalability of the Transformer architecture, and associated advances in language modeling such as GPT-x and BERT. In particular, we present Decision Transformer, an architecture that casts the problem of RL as conditional sequence modeling. Unlike prior approaches to RL that fit value functions or compute policy gradients, Decision Transformer simply outputs the optimal actions by leveraging a causally masked Transformer. By conditioning an autoregressive model on the desired return (reward), past states, and actions, our Decision Transformer model can generate future actions that achieve the desired return. Despite the simplicity, Decision Transformer matches or exceeds the performance of state-of-the-art model-free offline RL baselines on Atari, OpenAI Gym, and Key-to-Door tasks.

Keywords

- Reinforcement Learning (RL)
- Offline RL
- Model-free RL
- Conditional sequence modeling
- Autoregressive model
- Transformer

Thank you!

Deep Learning Sessions Lisboa (deep.learning.lx@gmail.com)

Vote for the next paper @ List Suggested Papers