Project DeiT presentation

DeiT: Data-efficient Image Transformers

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, Hervé Jégou

Sébastien Meyer

08 mars 2022

ToC

1. Available packages
Introducing Vision Transformers

- **Attention mechanisms for NLP**
  - Greatly improved in 2017 by Vaswani et al.
  - Focused on NLP tasks (language translation, parsing)

- **Attention mechanisms for Computer Vision**
  - Dosovitskiy et al. (2021) have adapted Transformers for vision tasks
  - Models have to be trained on large datasets to achieve good results

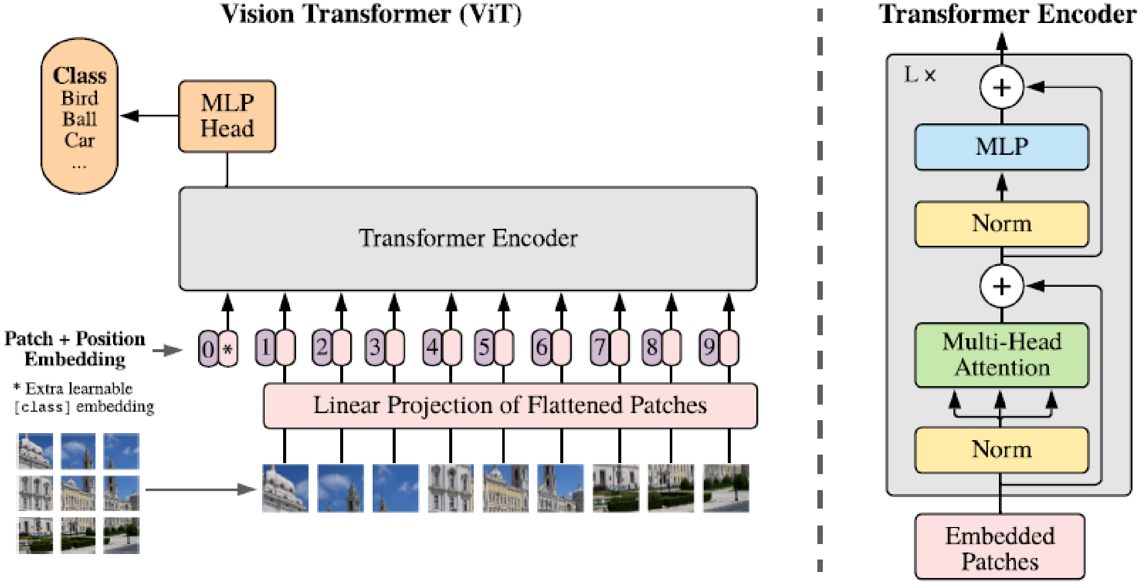


Fig. 1: Vision Transformer architecture

1. Available packages

- **Data-efficient Vision Transformers (2021)**
  - Small changes to ViT (lots of data aug) to allow training on smaller datasets
  - Introduced teacher-student distillation (concatenated token & loss)

Model	ImageNet	CIFAR-10	CIFAR-100	Flowers	Cars	iNat-18	iNat-19	im/sec
Grafit ResNet-50 [49]	79.6	–	–	98.2	92.5	69.8	75.9	1226.1
Grafit RegNetY-8GF [49]	–	–	–	99.0	94.0	76.8	80.0	591.6
ResNet-152 [10]	–	–	–	–	–	69.1	–	526.3
EfficientNet-B7 [48]	84.3	98.9	91.7	98.8	94.7	–	–	55.1
ViT-B/32 [15]	73.4	97.8	86.3	85.4	–	–	–	394.5
ViT-B/16 [15]	77.9	98.1	87.1	89.5	–	–	–	85.9
ViT-L/32 [15]	71.2	97.9	87.1	86.4	–	–	–	124.1
ViT-L/16 [15]	76.5	97.9	86.4	89.7	–	–	–	27.3
DeiT-B	81.8	99.1	90.8	98.4	92.1	73.2	77.7	292.3
DeiT-B↑384	83.1	99.1	90.8	98.5	93.3	79.5	81.4	85.9
DeiT-B⚗	83.4	99.1	91.3	98.8	92.9	73.7	78.4	290.9
DeiT-B⚗↑384	84.4	99.2	91.4	98.9	93.9	80.1	83.0	85.9

Fig. 2: Comparison of accuracy between ViT and DeiT models

- **Expediting Vision Transformers via Token Reorganization (2022)**
  - Merged inattentive image patches between each Transformerto speed up both training & testing

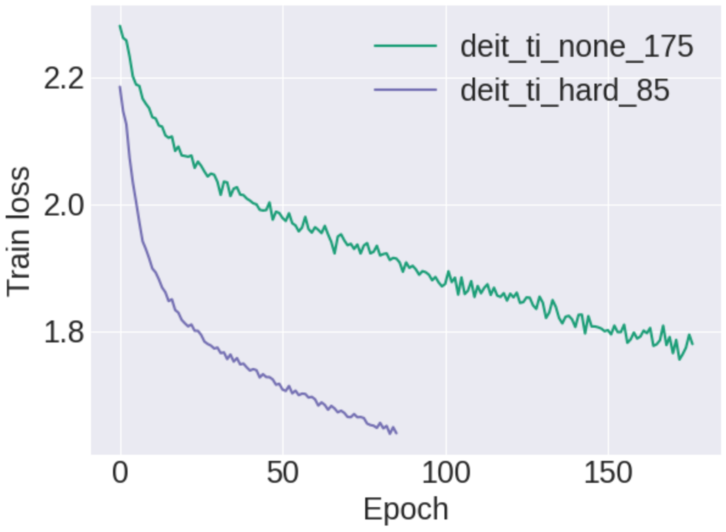
2. My modifications

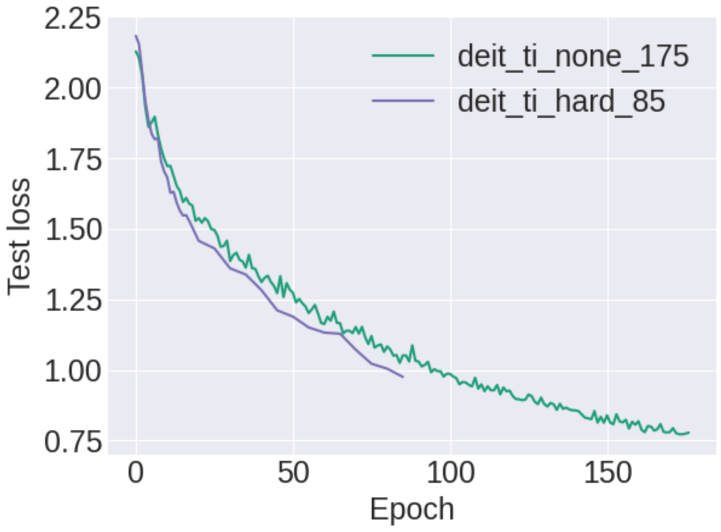
- **Forked repo**
  - See code: https://github.com/sebastienmeyer2/Project-deit

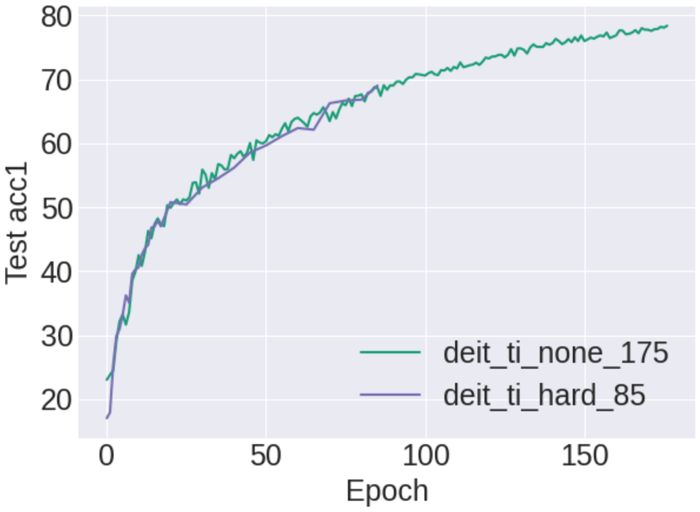
- **Support for CIFAR10**
  - Added CIFAR10 to the datasets preparation
  - Transfer learning of a teacher RegNetY for distillation

- **Integration of EviT for comparison**
  - Largely taken from https://github.com/youweiliang/evit
  - Combine both distillation and token pooling?

3. Some results

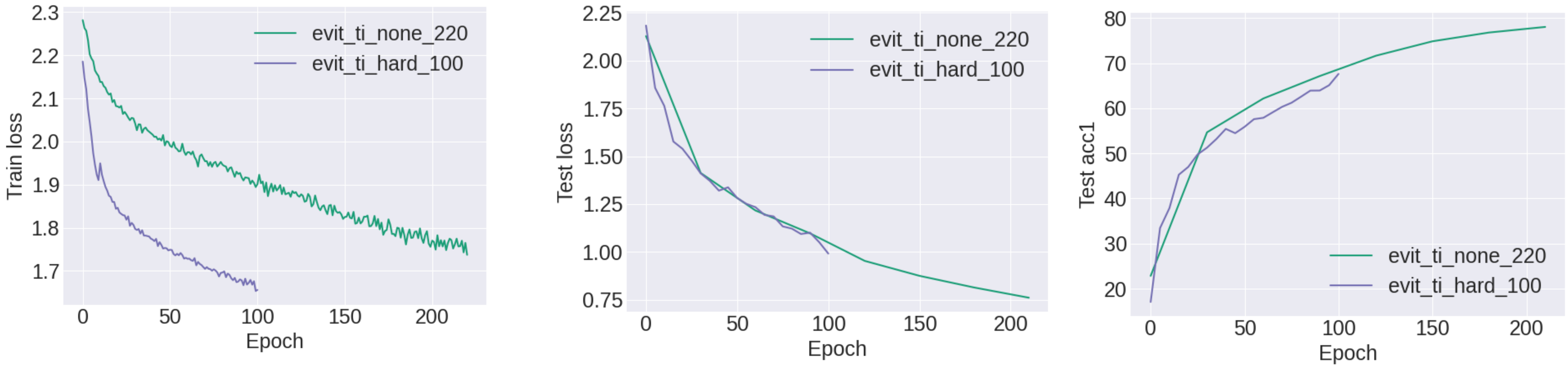






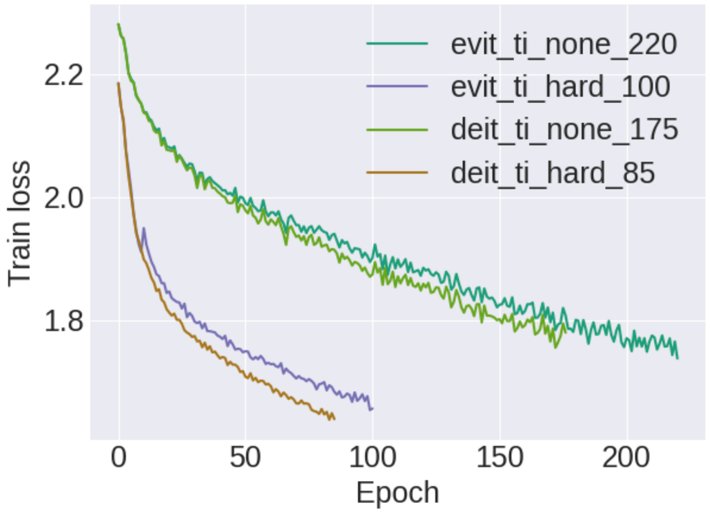
- **Comparison of training with and without distillation of DeiT (tiny, patch 16, size 224)**
  - Only the last classification layer of the teacher was modified
  - Approx. accuracy on CIFAR10 is 78% for the teacher after 100 epochs (can do much better!)

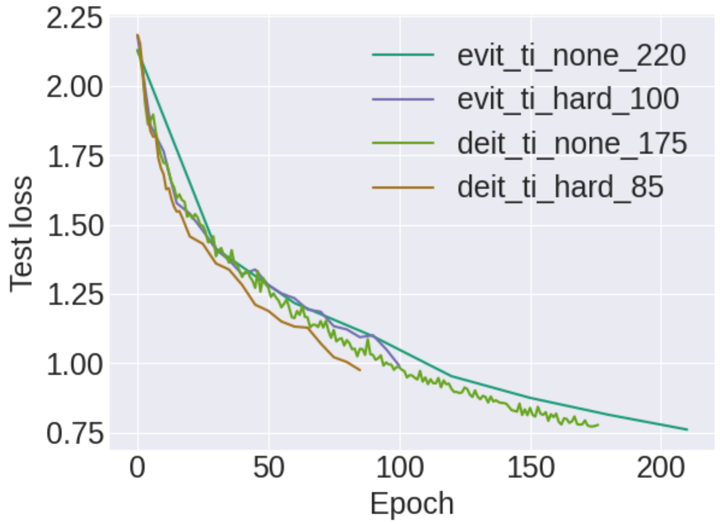
3. Some results

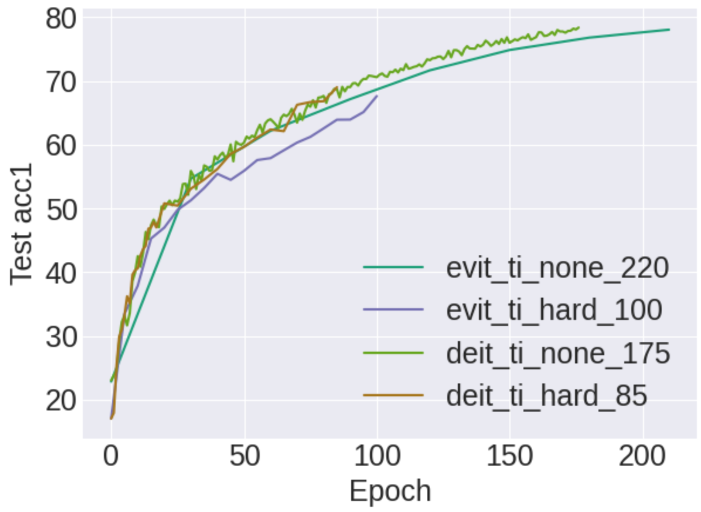


- **Comparison of training with and without distillation of EViT (tiny, patch 16, size 224)**
  - Same teacher as for DeiT - does not seem to help as much

3. Some results







- **Comparison of DeiT with EViT**
  - As specified by the authors, EViT is slightly less accurate than DeiT
  - No speed difference for CIFAR10: EViT 785 img/s & DeiT 785 img/s

Concluding remarks

- **Easy-to-use models but very long to train -> use timm!**
- **Speed up training by precomputing teacher's predictions?**

References

- Ashish Vaswani et al. *Attention Is All You Need*. December 2017. (Available at: https://arxiv.org/abs/1706.03762)
- Alexey Dosovitskiy et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. ICLR 2021. (Available at: https://arxiv.org/abs/2010.11929)
- Touvron et al. *Training data-efficient image transformers & distillation through attention*. January 2021. (Available at: https://arxiv.org/abs/2012.12877)
- Liang et al. *Not all Patches are What You Need: Expediting Vision Transformers via Token Reorganizations*. ICLR 2022. (Available at: https://arxiv.org/abs/2202.07800)