جامعة حلوان
Helwan University

كلية الحاسبات والذكاء الإصطناعي
Faculty of Computers & Artificial Intelligence

Faculty of Computers and Artificial Intelligence, Helwan University

Computer Science Department

2021/2022

CS 396 Selected Topics in CS-2

Research Project

Team ID No.46

ID	Name
20180086	أحمد هشام عطية
20180113	اسلام محمد محمود زكى
20180244	روان رفعت عبد الرازق
20180243	رنا جمال محمد محمد
20180233	رانا احمد على
20180394	عمر يحيى ابراهيم محمد
20180392	عمر وائل محمد خطاب

Delivered to:

Dr. Wessam El-Behaidy

Eng. Mai Mokhtar

Paper Details

- **Paper Name:** An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale
- **Authors:** Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov Dirk Weissenborn, Xiaohua Zhai, Thomas Unterhiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvian Gelly, Jakob Uszkoreit, Neil (Google Research, Brain Team)
- **Paper Link:** [2010.11929] An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (arxiv.org)
- **Conference:** ICLR 2021
- **Publication Year:** Submitted on 22 Oct 2020 (v1), last revised 3 Jun 2021 (this version, v2)]
- **Datasets used:**

For pretraining: the ILSVRC-2012 ImageNet dataset with 1k classes and 1.3M images, its superset ImageNet-21k with 21k classes and 14M images, and JFT Dataset with 18k classes and 303M high-resolution images.

For Fine-tuning: ImageNet, CIFAR-10/100, Oxford-IIIT Pets, Oxford Flowers-102, the 19-task VTAB classification suite(1000 training examples per task.)

- **Implemented Approach:**

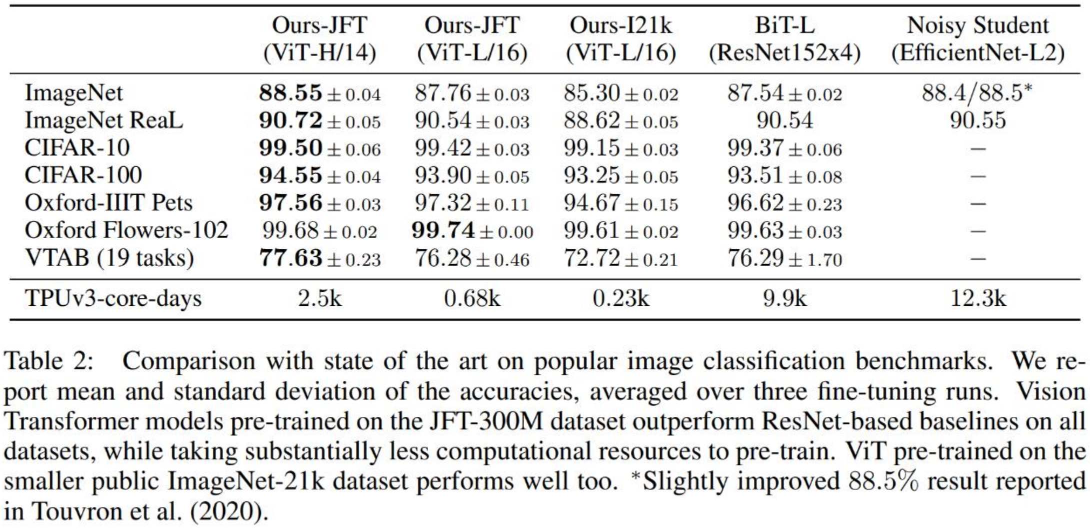
Vision Transformer is a visual model based on the architecture of a transformer originally designed for NLP tasks. It follows the original Transformer architecture as closely as possible. The ViT model represents an input image as a series of image patches, like the series of word embeddings used with transformers, and directly predicts class labels for the image. The model also learns on training data to encode the relative location of the image patches to reconstruct the structure of the image.
The transformer encoder includes:

- Multi-Head Self Attention Layer: This layer concatenates all the attention outputs linearly to the right dimensions. The many attention heads help train local and global dependencies in an image.
- Multi-Layer Perceptrons Layer: This layer contains a two-layer with GELU.
- Layer Norm: This is added prior to each block as it does not include any new dependencies between the training images. This thereby helps improve the training time and overall performance.

The approach followed:

- Each image is split into fixed-size patches, then the image patches are flattened.
- Lower-dimensional linear embeddings from these flattened image patches are created
- 1D Position embeddings are added to the patch embeddings to retain positional information
- An extra learnable embedding, classification token, is added to the sequence whose state at the output of the Transformer serves as the image representation class.
- The resulting sequence of embedding vectors is fed to a to a state-of-the-art transformer encoder
- ViT is pretrained with image labels, which is then fully supervised on a big dataset
- Fine-tune on the downstream dataset for image classification

○ **Results:**



	Ours-JFT (ViT-H/14)	Ours-JFT (ViT-L/16)	Ours-I21k (ViT-L/16)	BiT-L (ResNet152x4)	Noisy Student (EfficientNet-L2)
ImageNet	**88.55** ± 0.04	87.76 ± 0.03	85.30 ± 0.02	87.54 ± 0.02	88.4/88.5*
ImageNet ReaL	**90.72** ± 0.05	90.54 ± 0.03	88.62 ± 0.05	90.54	90.55
CIFAR-10	**99.50** ± 0.06	99.42 ± 0.03	99.15 ± 0.03	99.37 ± 0.06	–
CIFAR-100	**94.55** ± 0.04	93.90 ± 0.05	93.25 ± 0.05	93.51 ± 0.08	–
Oxford-IIIT Pets	**97.56** ± 0.03	97.32 ± 0.11	94.67 ± 0.15	96.62 ± 0.23	–
Oxford Flowers-102	99.68 ± 0.02	**99.74** ± 0.00	99.61 ± 0.02	99.63 ± 0.03	–
VTAB (19 tasks)	**77.63** ± 0.23	76.28 ± 0.46	72.72 ± 0.21	76.29 ± 1.70	–
TPUv3-core-days	2.5k	0.68k	0.23k	9.9k	12.3k

Table 2: Comparison with state of the art on popular image classification benchmarks. We report mean and standard deviation of the accuracies, averaged over three fine-tuning runs. Vision Transformer models pre-trained on the JFT-300M dataset outperform ResNet-based baselines on all datasets, while taking substantially less computational resources to pre-train. ViT pre-trained on the smaller public ImageNet-21k dataset performs well too. *Slightly improved 88.5% result reported in Touvron et al. (2020).

Project Description

Multi-class classification model using ViT pre-trained model to classify images in CIFAR Dataset

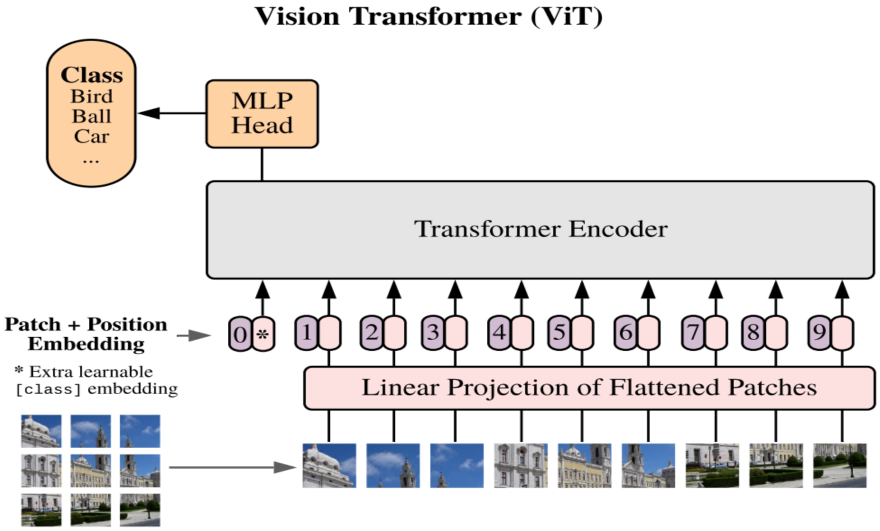
1. **Dataset:**
   **Name: CIFAR-100
   Link: cifar100 · Datasets at Hugging Face
   Total number of samples: The CIFAR-100 dataset consists of 60000 images.
   Dimension of images: 32x32x3 (3 color)

Number of classes and their labels: 100 classes, with 600 images per class. There are 500 training images and 100 testing images per class. The 100 classes are grouped into 20 super classes. There are two labels per image - fine label (actual class) and coarse label (superclass).

**Name: FOOD-101
Link: food101 · Datasets at Hugging Face
Total number of samples: The food-101 dataset consists of 101000 images.
Dimension of images: All images were rescaled to have a maximum side length of 512 pixels.
Number of classes and their labels: This dataset consists of 101 food categories, with 101'000 images. For each class, 250 manually reviewed test images are provided as well as 750 training images.

2. Implementation details (Fine-Tuning)

- Library: PyTorch
- Model: ViT (base-sized model)
- Image Transformations: a ViT feature extractor is initialized with a previously saved configuration to apply appropriate transformations to images before being passed to ViT
- The output tensors contain pixel values, which are the numeric representations of the image that we pass to the model.
- Data collator to form batches
- Accuracy as evaluation metric from HuggingFace
- The number of classes in our dataset (101 classes) is specified to create the appropriate classification head
- A pretrained checkpoint is loaded and configured for training
- **Data ratio:** There are 50000 training images and 10000 test images. Each class has 500 training images and 100 testing images.
- **Training Hyperparameters:** Training epochs=4, Training batch size= 16, Evaluation Batch Size = 8, Seed= 42, Learning Rate= 0.0002, Optimizer: Adam with betas= (0.9,0.999) and epsilon=1e-08, Mixed Precision Training= Native AMP
- **Model:**


Vision Transformer (ViT)
Class
Bird
Ball
Car
...
MLP Head
Transformer Encoder
Patch + Position Embedding
* Extra learnable [class] embedding
0
*
1
2
3
4
5
6
7
8
9
Linear Projection of Flattened Patches

For ViT Pretraining:

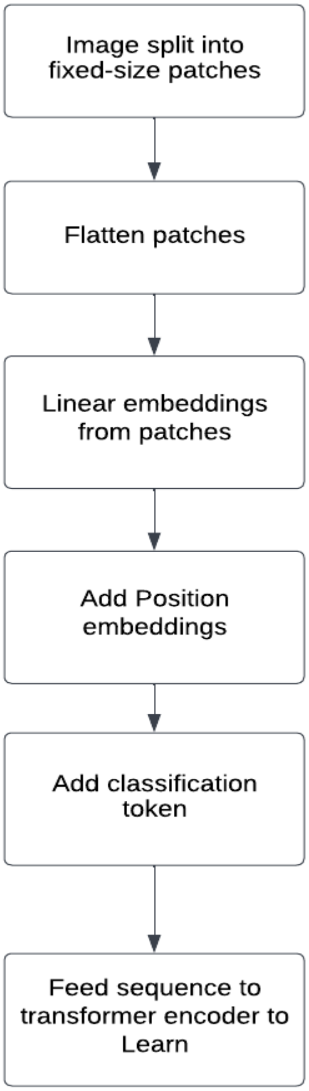


Image split into fixed-size patches
Flatten patches
Linear embeddings from patches
Add Position embeddings
Add classification token
Feed sequence to transformer encoder to Learn

3. Results

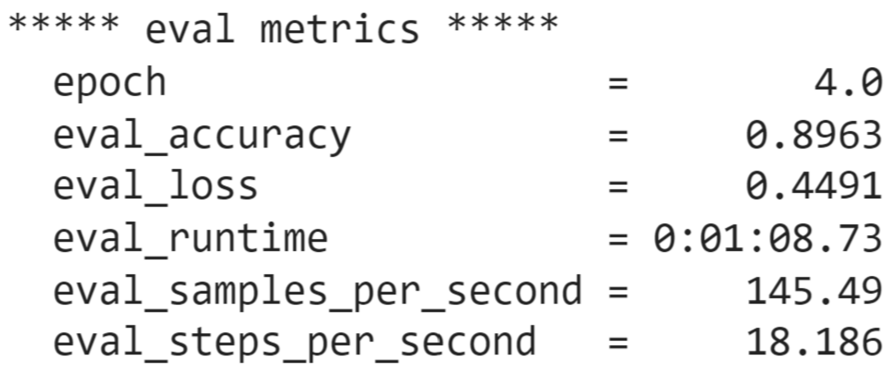
**Cifar-100:

Number of Samples: 10000 samples

Batch Size: 8

Accuracy: 0.8985

Loss: 0.442



```
***** eval metrics *****
  epoch                   =        4.0
  eval_accuracy           =     0.8963
  eval_loss               =     0.4491
  eval_runtime            = 0:01:08.73
  eval_samples_per_second =     145.49
  eval_steps_per_second   =     18.186
```

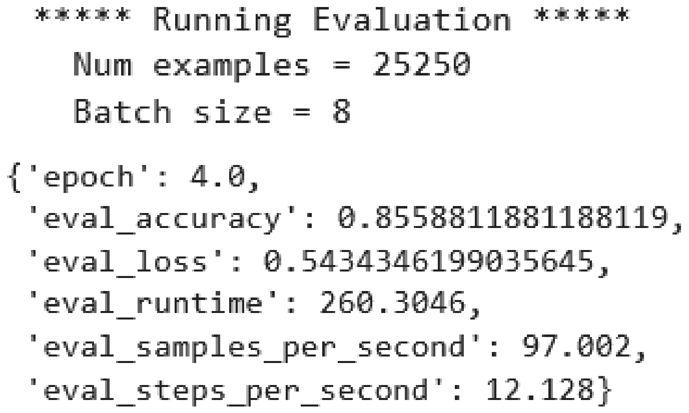
**Food-101:

Number of Samples: 10000 samples

Batch Size: 8

Accuracy: 0. 8559

Loss: 0.5434



```
 ***** Running Evaluation *****
   Num examples = 25250
   Batch size = 8
{'epoch': 4.0,
 'eval_accuracy': 0.8558811881188119,
 'eval_loss': 0.5434346199035645,
 'eval_runtime': 260.3046,
 'eval_samples_per_second': 97.002,
 'eval_steps_per_second': 12.128}
```