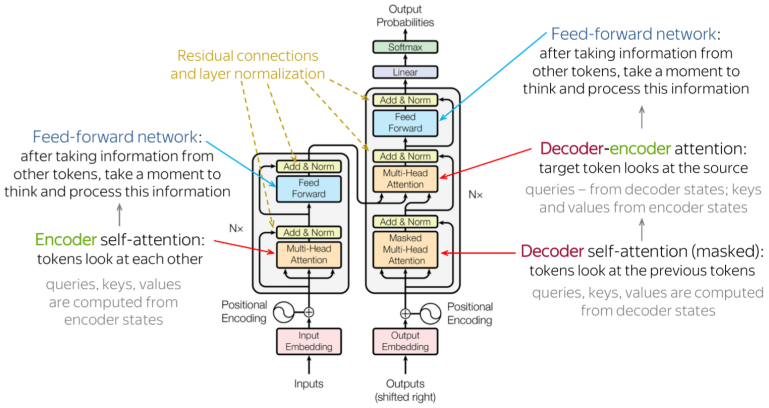
DEEP LEARNING PROJECT
FINAL REPORT

ABSTRACT

We have focused on upgrading the Vision Transformer (ViT) model, which is commonly used in NLP, for image classification. In order to enhance the performance of the ViT model, we plan to implement an additional preprocessing step using the SSD model before the Input Embedding process.

1 BACKGROUND

1.1 ViT



(a)

The MLP contains two layers with a GELU non-linearity.

$$\mathbf{z}_0 = [\mathbf{x}_{\text{class}}; \mathbf{x}_p^1\mathbf{E}; \mathbf{x}_p^2\mathbf{E}; \cdots; \mathbf{x}_p^N\mathbf{E}] + \mathbf{E}_{pos}, \quad \mathbf{E} \in \mathbb{R}^{(P^2 \cdot C) \times D}, \ \mathbf{E}_{pos} \in \mathbb{R}^{(N+1) \times D} \tag{1}$$

$$\mathbf{z}'_\ell = \text{MSA}(\text{LN}(\mathbf{z}_{\ell-1})) + \mathbf{z}_{\ell-1}, \quad \ell = 1 \ldots L \tag{2}$$

$$\mathbf{z}_\ell = \text{MLP}(\text{LN}(\mathbf{z}'_\ell)) + \mathbf{z}'_\ell, \quad \ell = 1 \ldots L \tag{3}$$

$$\mathbf{y} = \text{LN}(\mathbf{z}_L^0) \tag{4}$$

(b)

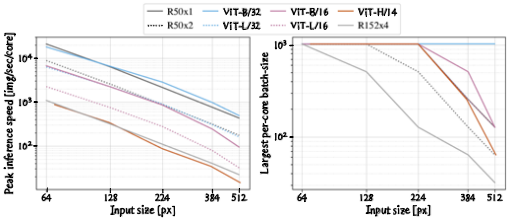
As illustrated in the figure above, the ViT model is composed of an encoder and decoder structure. In this process, we can see that the input, which has been through the embedding process, is used as the input for the Multihead Attention (MSA).

When utilizing a large-sized input image, the self-Attention operation must be performed on all patches. This means the inference is based on the entire image, rather than focusing on a specific object. This approach can be seen as observing the overall context rather than focusing on a specific point, especially when dealing with images requiring high accuracy. However, there are instances where we want to closely observe a specific point. For example, in a photo of a crowd, we might want to identify a 'Person-Name' rather than just confirming the presence of 'Person'. Similarly, when interpreting medical images (CT, MRI, etc.), we might want to focus on a particular abnormality to diagnose a disease. In these special scenarios, rather than inferring from the whole picture through the ViT, we would want to intensely analyze the content within a specific bounding box. For these reasons, we plan to use SSD as a preprocessing step to implement a higher-level classifier.

1.2 *Attention Map*

Through the Attention map, we can visually identify the areas being attended to in the photo. Let me explain using the above photo as an example.
Unlike in the bird photos where Attention functioned well, it can be visually observed that Attention is not properly functioning in the baby photos.

1.3 *Comparison of Inference Speed Depending on Input Size in ViT Model*

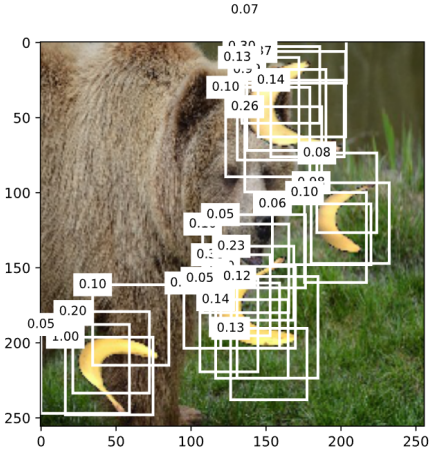


This graph represents the correlation between input size and inference speed in several models, including ViT. It shows a high dependency on the input size. Therefore, if smaller images are inputted through single, the inference speed of ViT will improve.

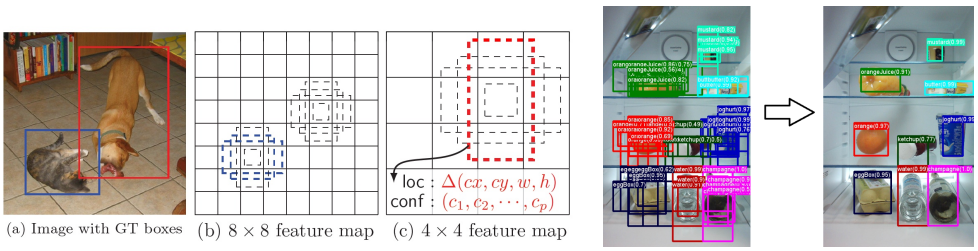
However, there exists a trade-off between the increased time consumption due to single and the reduced ViT inference time due to the decrease in input size.

1.4 *Single Shot multibox Detection (SSD)*

SSD differs from traditional predefined anchor boxes by generating anchor box centers only at reasonable locations. This approach prevents the creation of anchor boxes in unnecessary locations, thereby reducing unnecessary computations. The figure below illustrates an example of how anchor boxes are created in SSD.



We have also trained the model with objects other than bananas and confirmed that the prediction bounding box is accurately generated at the corresponding location.



(a) Image with GT boxes (b) 8×8 feature map (c) 4×4 feature map

(c) (d)

Figure 1: (c) Multiscale object (d) NMS

1.5 *NMS(Non-Maximum Suppression) Algorithm*

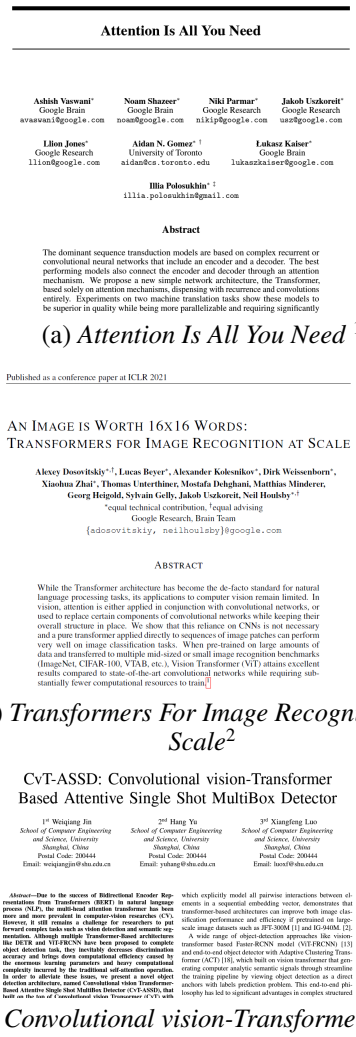
Before understanding the NMS algorithm, it's important to understand confidence scores.
When we say Confidence Score, we mean that we have n% confidence that the network has produced the correct answer.
means that when the network comes up with the correct answer, it has n% confidence in that answer.
Every bounding box carries a confidence score that indicates how well it captures the corresponding object. Bounding boxes with confidence scores below a certain threshold are eliminated, which is a preliminary filtering process for boxes with low-level certainty.
The remaining bounding boxes are then sorted in descending order based on their confidence scores.

Algorithm 1 Non-Max Suppression

```
procedure NMS(B,c)
    B_nms ← ∅
    for b_i ∈ B do
        discard ← False
        for b_j ∈ B do
            if same(b_i, b_j) > λ_nms then
                if score(c, b_j) > score(c, b_i) then
                    discard ← True
        if not discard then
            B_nms ← B_nms ∪ b_i
    return B_nms
```

Taking the bounding box at the front as a standard, the Intersection over Union (IoU) values with other bounding boxes are calculated. Boxes with an IoU exceeding a certain threshold are removed. This is because the higher the IoU between bounding boxes, the more likely they are detecting the same object.
This process is sequentially executed to compare and eliminate all bounding boxes. The higher the confidence threshold and the lower the IoU threshold, the more bounding boxes are removed.

1.6 *Related Works*



Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly

(a) *Attention Is All You Need* [1]

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]

(b) *Transformers For Image Recognition At Scale*[2]

CvT-ASSD: Convolutional vision-Transformer Based Attentive Single Shot MultiBox Detector

1st Weiqiang Jin School of Computer Engineering and Science, University Shanghai, China Postal Code: 200444 Email: weiqiangjin@shu.edu.cn
2nd Hang Yu School of Computer Engineering and Science, University Shanghai, China Postal Code: 200444 Email: yuhang@shu.edu.cn
3rd Xiangfeng Luo School of Computer Engineering and Science, University Shanghai, China Postal Code: 200444 Email: luoxf@shu.edu.cn

Abstract—Due to the success of Bidirectional Encoder Representations from Transformers (BERT) in natural language process (NLP), the multi-head attention transformer has been more and more prevalent in computer-vision researches (CV). However, it still remains a challenge for researchers to put forward complex tasks such as vision detection and semantic segmentation. Although multiple Transformer-Based architectures like DETR and ViT-FRCNN have been proposed to complete object detection task, they inevitably decreases discrimination accuracy and brings down computational efficiency caused by the enormous learning parameters and heavy computational complexity incurred by the traditional self-attention operation. In order to alleviate these issues, we present a novel object detection architecture, named Convolutional vision Transformer-Based Attentive Single Shot MultiBox Detector (CvT-ASSD), that built on the top of Convolutional vision Transformer (CvT) with which explicitly model all pairwise interactions between elements in a sequential embedding vector, demonstrates that transformer-based architectures can improve both image classification performance and efficiency if pretrained on large-scale image datasets such as JFT-300M [1] and IG-940M. [2].

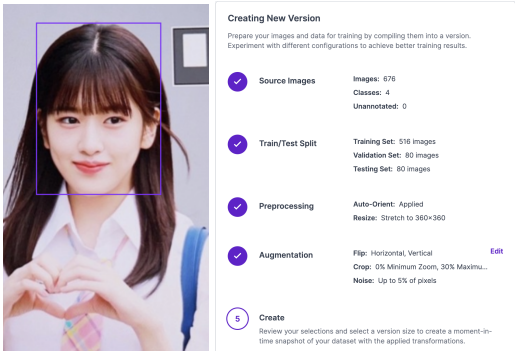
A wide range of object-detection approaches like vision-transformer based Faster-RCNN model (ViT-FRCNN) [13] and end-to-end object detector with Adaptive Clustering Transformer (ACT) [18], which built on vision transformer that generating computer analytic semantic signals through streamline the training pipeline by viewing object detection as a direct anchors with labels prediction problem. This end-to-end philosophy has led to significant advantages in complex structured

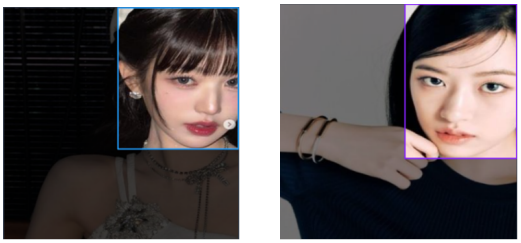
(c) *Convolutional vision-Transformer Based Attentive Single Shot MultiBox Detector* [3]

[1] Attention Is All You Need
[2] Transformers For Image Recognition Scale
[3] Convolutional vision-Transformer Based Attentive Single Shot MultiBox Detector

2 *Dataset and method*

2.1 *Dataset*

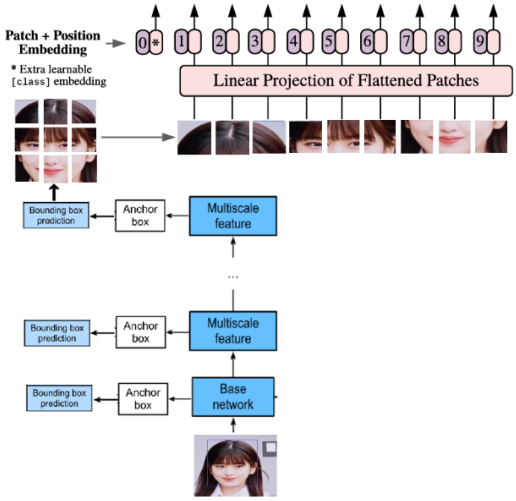




Jang Won-young, An Yu-jin, Kwon Eun-bi and other members picture. We used Roboflow to classify and set the Ground Truth for 671 original images.

After augmenting these images, the dataset was expanded to include 1548 training, 80 validation, and 80 testing images.

3 *Method*



1. Use SSD to crop the image area that will be input to ViT.
2. Insert the bounding box produced by SSD as an input to ViT.

Anticipating the location of the objects, the image is passed as input to the Vision Transformer. it is akin to providing ViT with images containing a single object.
This is expected to have a similar effect as noise reduction.
The aim of this process is to enhance the prediction accuracy of the ViT.

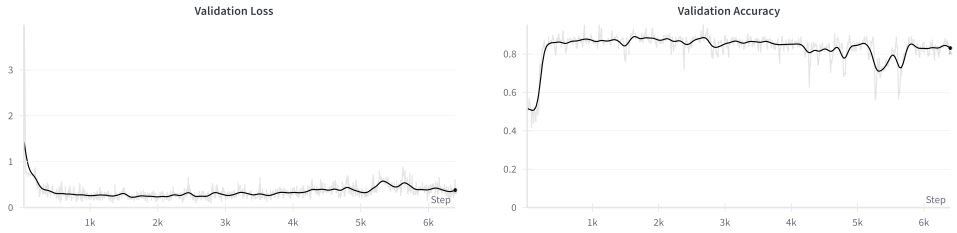
ViT model Parameter	Parameter Value
n_patches	16x16 (256)
model_dim	768
hidden_dim	3072
n_class	4
n_heads	12
n_layers	12

4 *Result*

4.1 *Environment*

Google Colab	Pro version
System RAM	50GB
GPU	NVIDIA A100, V100, T4

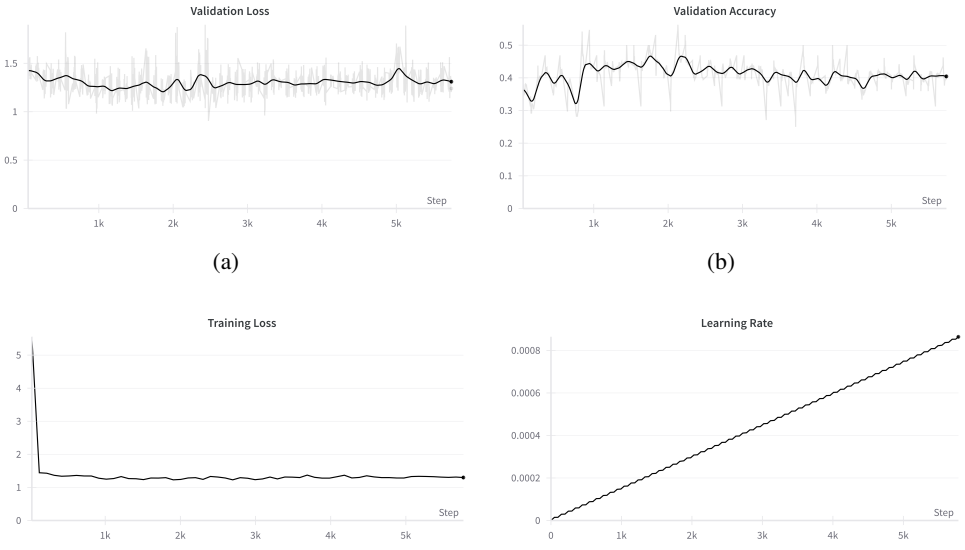
4.2 *SSD*



(a) (b)

Figure 2: (a) Validation loss (b) Validation Accuracy

4.3 *ViT*



(a) (b)

(c) (d)

Figure 3: (a) Validation loss (b) Validation Accuracy
(c) Training loss (d) learning rate

5 *Discussion*

The anticipated benefit of using the ViT model and SSD's prediction bounding box together is an increase in ViT's inference speed due to a decrease in input size. However, as we are using the SSD's prediction bounding box, the speed is expected to slow down.

The primary goal is to accurately analyze a specific object, hence the design isn't focused on real-time inference.

Let's first consider why the result accuracy did not come out as well as expected. One possibility is that the dataset we manually classified did not become a good quality dataset because the criteria for setting the ground truth on the face were not consistent. Secondly, a large amount of training data is required to train the ViT model, but the dataset we created only contains just over 1800 images, which I think is a problem. Considering that it requires more data for training compared to other models, we even proceeded with augmentation, but it was insufficient. Lastly, I believe not being able to utilize a pre-trained model was a problem. As introduced in the paper, it is recommended to use a pre-trained model.

We've realized that the computing resources required to train the ViT are crucial. The long training time of up to 30 hours for ViT makes it hard to run experiments and check results, this is a regrettable aspect.

6 *Reference*

6.1 *Baseline*

ViT baseline [4]

SSD baseline [5]

[4] ViT baseline
[5] SSD baseline