

# Loading and Processing Data

## Instructor

Dipanjan Sarkar

Head of Community & Principal AI Scientist at Analytics Vidhya

Google Developer Expert - ML & Cloud Champion Innovator

Published Author



# Key Objectives of the Project

Build a Search Engine on  
Wikipedia Articles & Research Papers



Process text and PDF  
documents



Create document and  
contextual chunks



Index chunks and  
embeddings in a vector  
database



Experiment with different  
retrievers

Loading and Processing Data

# Data Sources



Text Article JSON

Deep Residual Learning for Image Recognition

Kaiming HeXiangyu ZhangShaoqing RenJian Sun

Microsoft Research

*(This work was done while Kaiming He was a student at the University of California, Berkeley.)*

Attention Is All You Need

Ashish VaswaniGoogle

Llion JonesGoogle

Deeper neural nets present a residual learn of networks that are as previously. We explicitly ing residual functions instead of learning unrepresentative empirical networks are easier to considerably increased evaluate residual nets deeper than VGG nets; its. An ensemble of them on the ImageNet test set ILSVRC 2015 classifies on CIFAR-10 with 100.

The depth of resnet for many visual recog tremely deep represents provement on the COCO residual nets are found & COCO 2015 compes places on the tasks of l cation, COCO detection

### 1. Introduction

Deep convolutional to a series of blocks 50, 40). Deep network level features [50] and layer fusion, and the by the number of stack [4], 44) reveals that net and the leading results ImageNet dataset [36] with a depth of sixteen trivial visual recognition

<https://arxiv.org/abs/1607.08022v1>

### TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

Alexey Dosovitskiy<sup>1</sup>, Lucas Beyer<sup>2</sup>, Alexander Kolesnikov<sup>2</sup>, Dirk Weissenborn<sup>2</sup>, Xiaohua Zhai<sup>1</sup>, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelb, Jakob Uszkoreit, Neil Houlsby<sup>1</sup>

<sup>1</sup>equal technical contribution, <sup>2</sup>equal advising

Google Research, Brain Team

{adosovitskiy, nellhoulsby}@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 INTRODUCTION

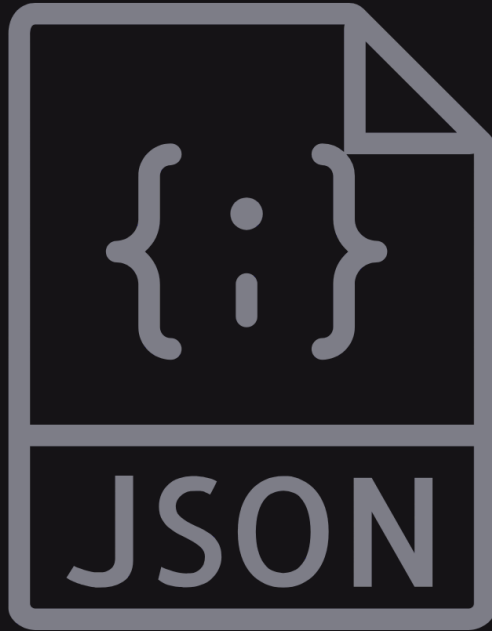
Self-attention-based architectures, in particular Transformers (Vaswani et al., 2017), have become the model of choice in natural language processing (NLP). The dominant approach is to pre-train on a large text corpus and then fine-tune on a smaller task-specific dataset (Devlin et al., 2019). Thanks to Transformers' computational efficiency and scalability, it has become possible to train models of unprecedented size, with over 100B parameters (Brown et al., 2020; Lepikhin et al., 2020). With the models and datasets growing, there is still no sign of saturating performance.

In computer vision, however, convolutional architectures remain dominant (LeCun et al., 1989; Krizhevsky et al., 2012; He et al., 2016). Inspired by NLP successes, multiple works try combining CNN-like architectures with self-attention (Wang et al., 2018; Carion et al., 2020), some replacing the convolutions entirely (Ramachandran et al., 2019; Wang et al., 2020a). The latter models, while theoretically efficient, have not yet been scaled effectively on modern hardware accelerators due to the use of specialized attention patterns. Therefore, in large-scale image recognition, classic ResNet-like architectures are still state of the art (Mahajan et al., 2018; Xie et al., 2020; Kolesnikov et al., 2020).

Inspired by the Transformer scaling successes in NLP, we experiment with applying a standard Transformer directly to images, with the fewest possible modifications. To do so, we split an image

Research Paper PDFs

# Data Loader

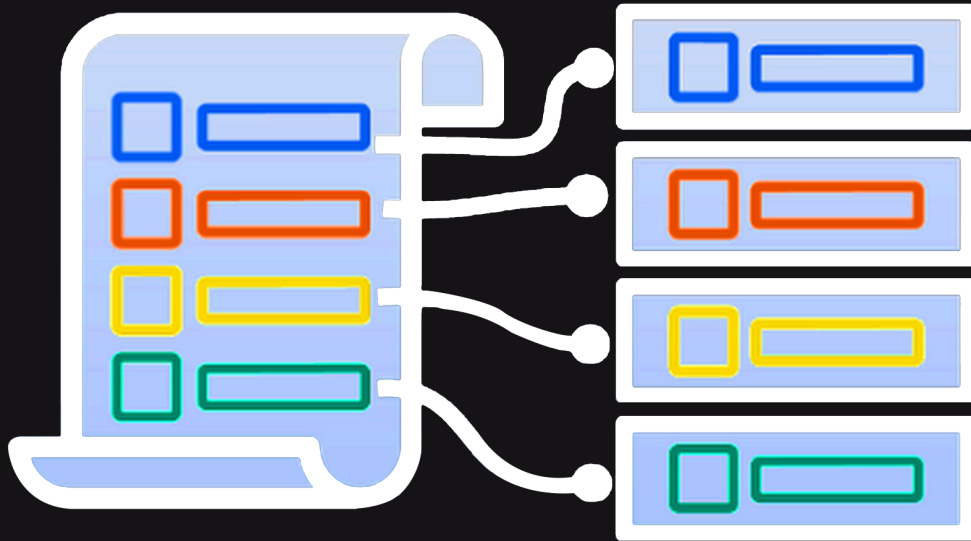


JSON Loader

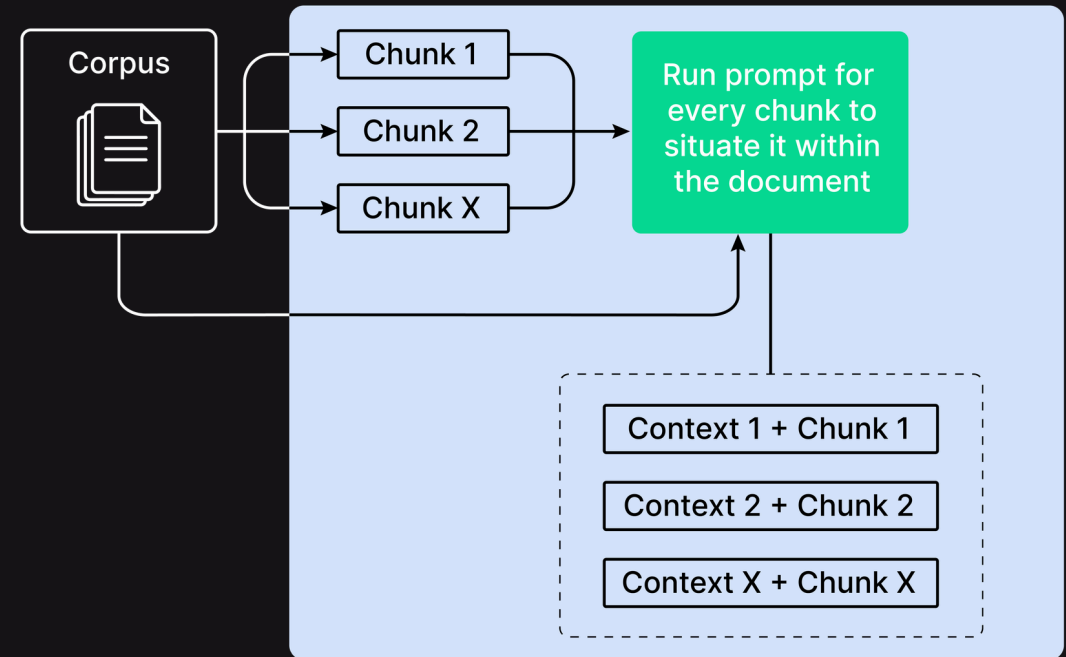


PDF Loader

# Chunking Strategies



Recursive Character  
Text Splitting & Chunking



Contextual Chunking

# Thank You

---