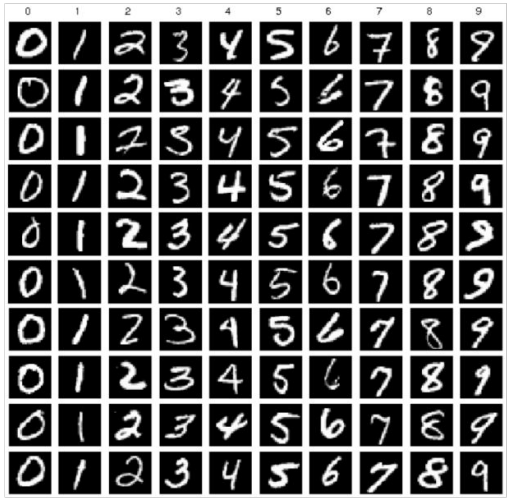
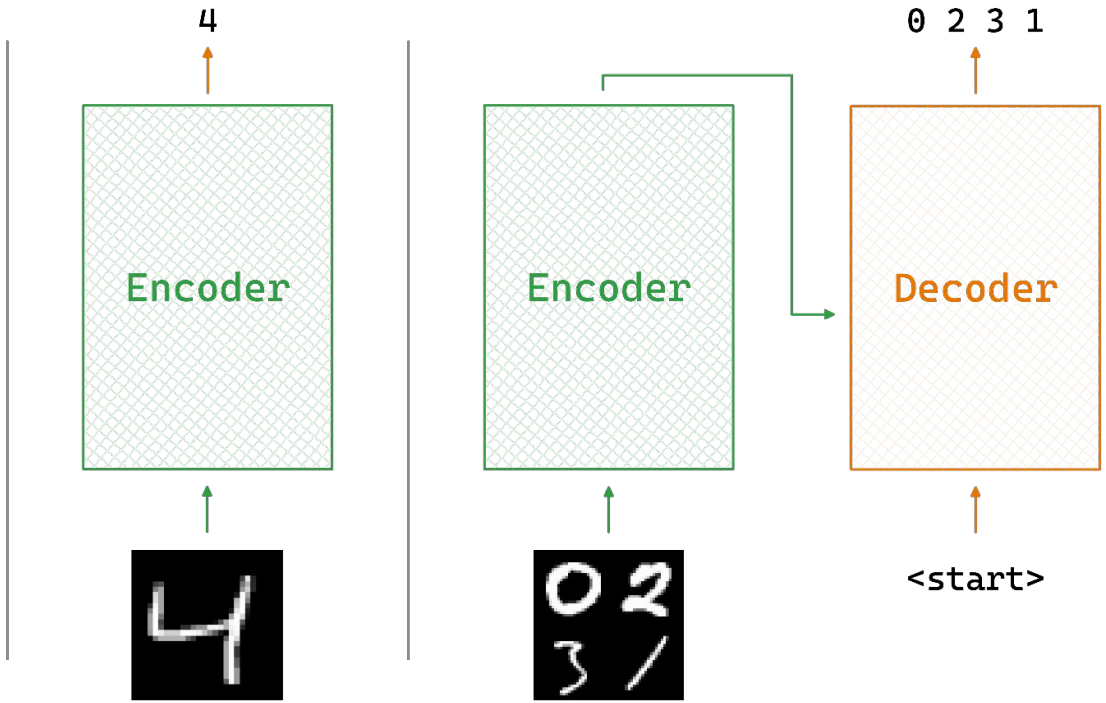
Multimodal transfer learning

Merge models for “novel” tasks

Previous Task





Task

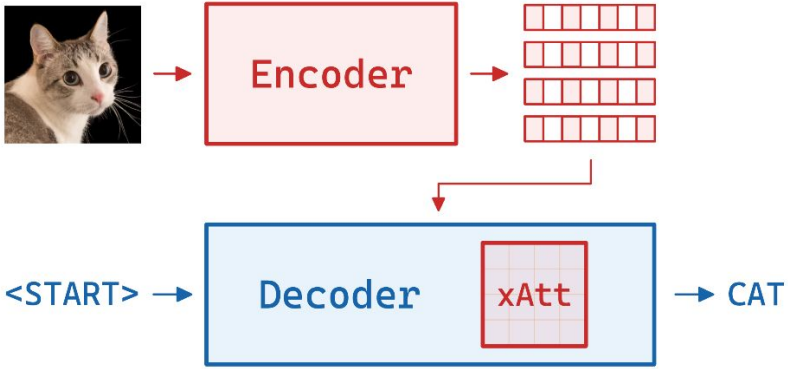
Magic

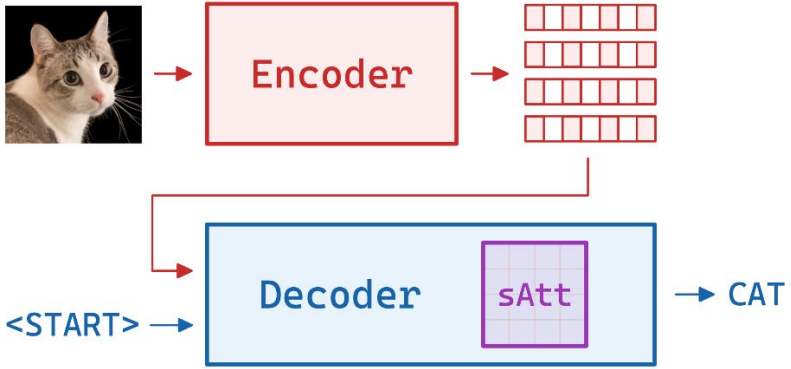
A little girl climbing into a wooden playhouse.

FOUNDERS AND CODERS

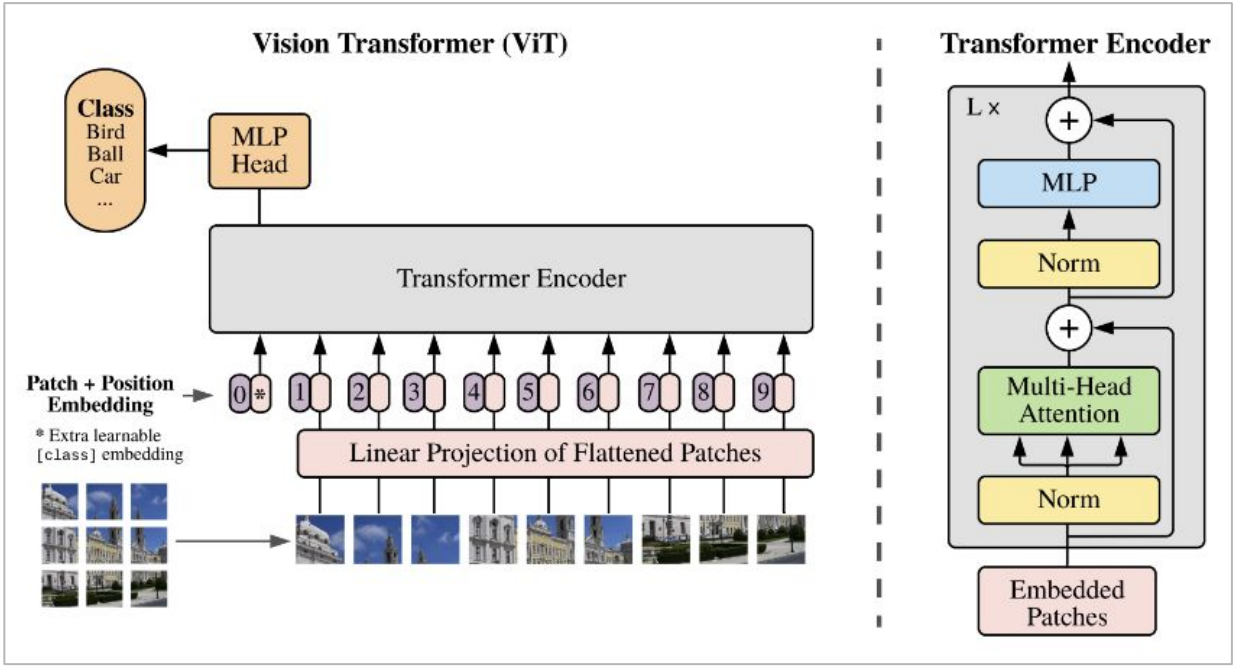
Transfer Learning

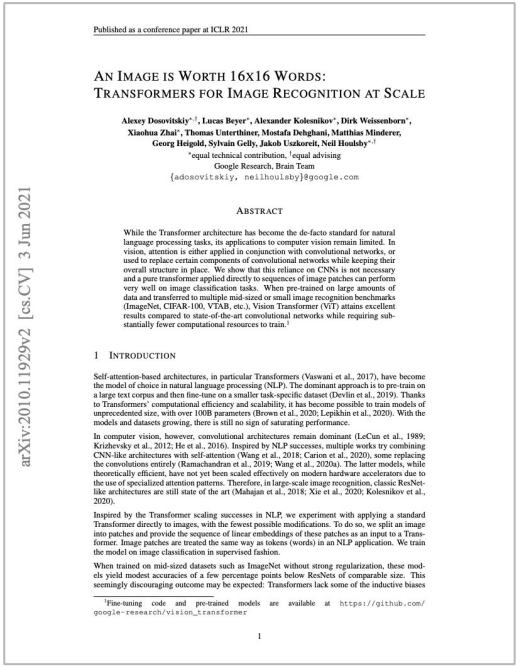
Machine Learning Institute





Vision Transformer (ViT)





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]

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 into patches and provide the sequence of linear embeddings of these patches as an input to a Transformer. Image patches are treated the same way as tokens (words) in an NLP application. We train the model on image classification in supervised fashion.

When trained on mid-sized datasets such as ImageNet without strong regularization, these models yield modest accuracies of a few percentage points below ResNets of comparable size. This seemingly discouraging outcome may be expected: Transformers lack some of the inductive biases

[1]Fine-tuning code and pre-trained models are available at https://github.com/google-research/vision_transformer

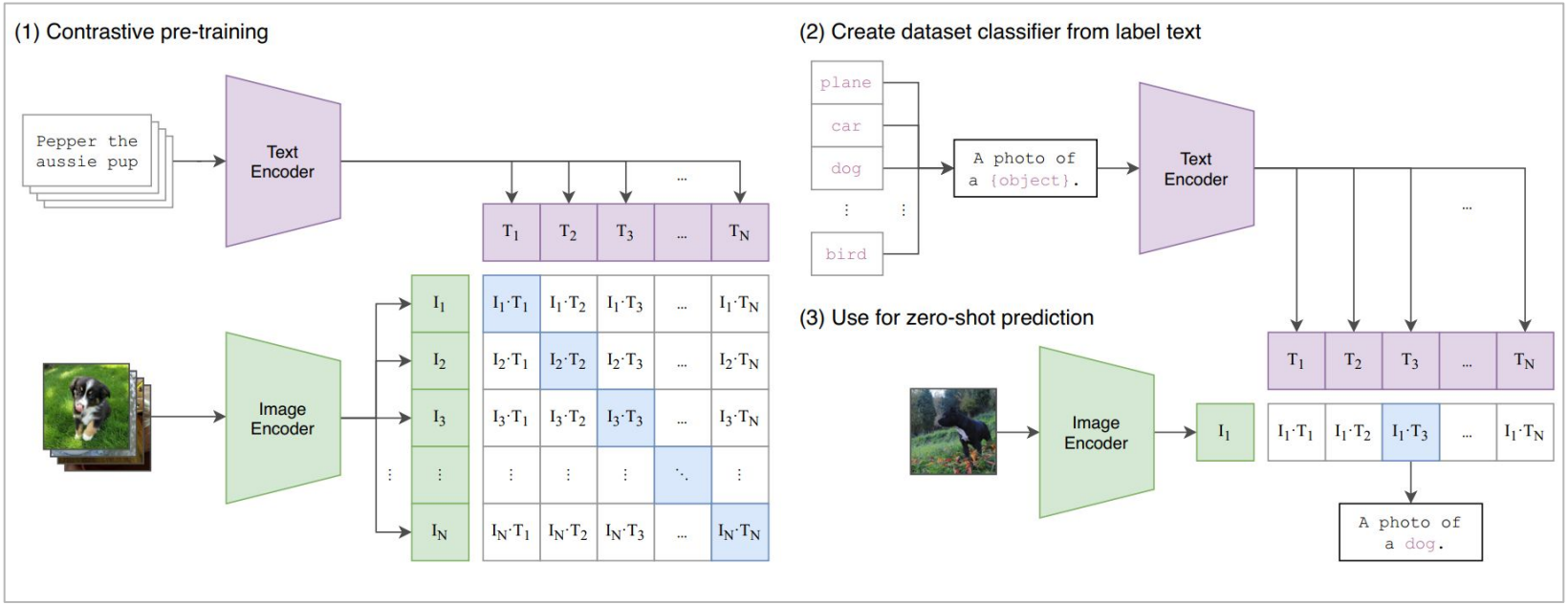
1

Dosovitskiy et al. (ICLR 2021)

FOUNDERS AND CODERS

CLIP

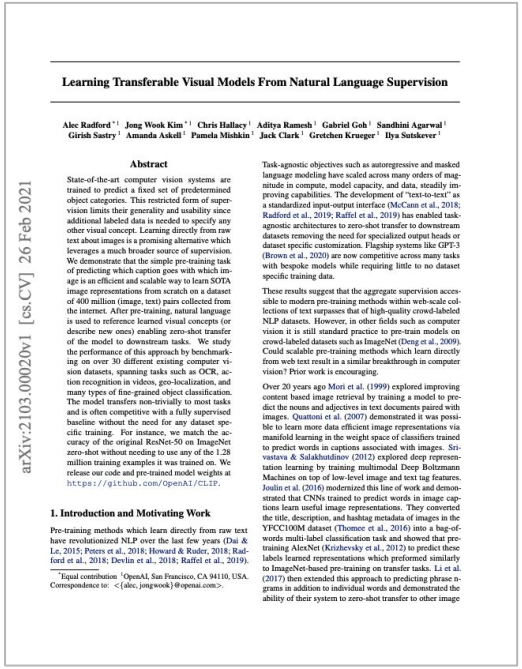
Machine Learning Institute



FOUNDERS AND CODERS

CLIP

Machine Learning Institute



Learning Transferable Visual Models From Natural Language Supervision

Alec Radford [*1] Jong Wook Kim [*1] Chris Hallacy [1] Aditya Ramesh [1] Gabriel Goh [1] Sandhini Agarwal [1]
Girish Sastry [1] Amanda Askell [1] Pamela Mishkin [1] Jack Clark [1] Gretchen Krueger [1] Ilya Sutskever [1]

Abstract

State-of-the-art computer vision systems are trained to predict a fixed set of predetermined object categories. This restricted form of supervision limits their generality and usability since additional labeled data is needed to specify any other visual concept. Learning directly from raw text about images is a promising alternative which leverages a much broader source of supervision. We demonstrate that the simple pre-training task of predicting which caption goes with which image is an efficient and scalable way to learn SOTA image representations from scratch on a dataset of 400 million (image, text) pairs collected from the internet. After pre-training, natural language is used to reference learned visual concepts (or describe new ones) enabling zero-shot transfer of the model to downstream tasks. We study the performance of this approach by benchmarking on over 30 different existing computer vision datasets, spanning tasks such as OCR, action recognition in videos, geo-localization, and many types of fine-grained object classification. The model transfers non-trivially to most tasks and is often competitive with a fully supervised baseline without the need for any dataset specific training. For instance, we match the accuracy of the original ResNet-50 on ImageNet zero-shot without needing to use any of the 1.28 million training examples it was trained on. We release our code and pre-trained model weights at https://github.com/OpenAI/CLIP.

1. Introduction and Motivating Work

Pre-training methods which learn directly from raw text have revolutionized NLP over the last few years (Dai & Le, 2015; Peters et al., 2018; Howard & Ruder, 2018; Radford et al., 2018; Devlin et al., 2018; Raffel et al., 2019). Task-agnostic objectives such as autoregressive and masked language modeling have scaled across many orders of magnitude in compute, model capacity, and data, steadily improving capabilities. The development of "text-to-text" as a standardized input-output interface (McCann et al., 2018; Radford et al., 2019; Raffel et al., 2019) has enabled task-agnostic architectures to zero-shot transfer to downstream datasets removing the need for specialized output heads or dataset specific customization. Flagship systems like GPT-3 (Brown et al., 2020) are now competitive across many tasks with bespoke models while requiring little to no dataset specific training data.

These results suggest that the aggregate supervision accessible to modern pre-training methods within web-scale collections of text surpasses that of high-quality crowd-labeled NLP datasets. However, in other fields such as computer vision it is still standard practice to pre-train models on crowd-labeled datasets such as ImageNet (Deng et al., 2009). Could scalable pre-training methods which learn directly from web text result in a similar breakthrough in computer vision? Prior work is encouraging.

Over 20 years ago Mori et al. (1999) explored improving content based image retrieval by training a model to predict the nouns and adjectives in text documents paired with images. Quattoni et al. (2007) demonstrated it was possible to learn more data efficient image representations via manifold learning in the weight space of classifiers trained to predict words in captions associated with images. Srivastava & Salakhutdinov (2012) explored deep representation learning by training multimodal Deep Boltzmann Machines on top of low-level image and text tag features. Joulin et al. (2016) modernized this line of work and demonstrated that CNNs trained to predict words in image captions learn useful image representations. They converted the title, description, and hashtag metadata of images in the YFCC100M dataset (Thomee et al., 2016) into a bag-of-words multi-label classification task and showed that pre-training AlexNet (Krizhevsky et al., 2012) to predict these labels learned representations which preformed similarly to ImageNet-based pre-training on transfer tasks. Li et al. (2017) then extended this approach to predicting phrase n-grams in addition to individual words and demonstrated the ability of their system to zero-shot transfer to other image

[*]Equal contribution [1]OpenAI, San Francisco, CA 94110, USA. Correspondence to: <{alec, jongwook}@openai.com>.

Radford et al. (ICML 2021)

Preprint

DEMYSTIFYING CLIP DATA

Hu Xu [1] Saining Xie [2] Xiaoqing Ellen Tan [1] Po-Yao Huang [1] Russell Howes [1] Vasu Sharma [1]
Shang-Wen Li [3] Gargi Ghosh [1] Luke Zettlemoyer [1,3] Christoph Feichtenhofer [1]
[1]FAIR, Meta AI [2]New York University [3]University of Washington

ABSTRACT

Contrastive Language-Image Pre-training (CLIP) is an approach that has advanced research and applications in computer vision, fueling modern recognition systems and generative models. We believe that the main ingredient to the success of CLIP is its *data* and *not* the *model* architecture or pre-training objective. However, CLIP only provides very limited information about its data and how it has been collected, leading to works that aim to reproduce CLIP's data by filtering with its model parameters. In this work, we intend to reveal CLIP's data curation approach and in our pursuit of making it open to the community introduce Metadata-Curated Language-Image Pre-training (MetaCLIP). MetaCLIP takes a raw data pool and metadata (derived from CLIP's concepts) and yields a balanced subset over the metadata distribution. Our experimental study rigorously isolates the model and training settings, concentrating solely on data. MetaCLIP applied to CommonCrawl with 400M image-text data pairs outperforms CLIP's data on multiple standard benchmarks. In zero-shot ImageNet classification, MetaCLIP achieves 70.8% accuracy, surpassing CLIP's 68.3% on ViT-B models. Scaling to 1B data, while maintaining the same training budget, attains 72.4%. Our observations hold across various model sizes, exemplified by ViT-bigG producing 82.1%. Curation code and training data distribution over metadata is available at https://github.com/facebookresearch/MetaCLIP.

1 INTRODUCTION

Deep learning has revolutionized the field of artificial intelligence, and pre-trained models have played a pivotal role in democratizing access to cutting-edge AI capabilities. However, the training data used to create these models is often concealed from the public eye, shrouded in secrecy.

The increasing availability of pre-trained models for public use contrasts sharply with the lack of transparency regarding their training data. Further, proprietary concerns, such as copyright issues, often limit access to the original data sources. Consequently, the need to explore novel approaches for curating high-quality training data that can be shared openly arises.

In the vision-language domain, the dominant model and learning approach is Contrastive Language-Image Pre-training (CLIP) (Radford et al., 2021), a simple technique to learn from image-text pairs. We believe that the secret to the dominance of CLIP models is attributed to its high-quality WIT400M *dataset* which is curated from the web. Despite its popularity, the specifics of CLIP's curation process have remained a mystery, captivating the research community for years.

Follow-up works (Schuhmann et al., 2022; 2021) have attempted to replicate CLIP's data, but with a notable difference in their curation method. While CLIP generates data based on its unknown data source and curation methodology, these approaches remove noise by applying the CLIP model as a hard blackbox filter which in turn is a form of distilling WIT400M information captured in CLIP.

The advantages of CLIP's curation are apparent. First, it starts *from scratch*, avoiding the introduction of biases through filters. Second, CLIP's curation process *balances* the data distribution over metadata, maximizing signal preservation while mitigating, rather than removing, noise in the data[1]. Such distribution lays the groundwork for task-agnostic data, a crucial part of foundation models.

[1]For example, a filter on digits can remove noise from date or id strings but remove signal for tasks that involve OCR (e.g., MNIST), or a filter removing text with less than 5 characters can remove signal "dog".

Hu Xu et al. (ICLR 2024)

FOUNDERS AND CODERS

Transfer Learning

Machine Learning Institute

```python
#
#
#
import transformers


#
#
#
c = transformers.CLIPModel.from_pretrained('openai/clip-vit-base-patch32')
v = transformers.ViTModel.from_pretrained('google/vit-base-patch16-224-in21k')
params = lambda m: sum(p.numel() for p in m.parameters())


#
#
#
print("CLIP:", params(c)) # 151,277,313
print("ViT:", params(v)) #   86,389,248


#
#
#
print("CLIP:", c)
print("ViT:", v)

```

FOUNDERS AND CODERS

ViT

```
ViTModel(
  (embeddings): ViTEmbeddings(
    (patch_embeddings): ViTPatchEmbeddings(
      (projection): Conv2d(3, 768, kernel_size=(16, 16), stride=(16, 16))
    )
    (dropout): Dropout(p=0.0, inplace=False)
  )
  (encoder): ViTEncoder(
    (layer): ModuleList(
      (0-11): 12 x ViTLayer(
        (attention): ViTSdpaAttention(
          (attention): ViTSdpaSelfAttention(
            (query): Linear(in_features=768, out_features=768, bias=True)
            (key): Linear(in_features=768, out_features=768, bias=True)
            (value): Linear(in_features=768, out_features=768, bias=True)
            (dropout): Dropout(p=0.0, inplace=False)
          )
          (output): ViTSelfOutput(
            (dense): Linear(in_features=768, out_features=768, bias=True)
            (dropout): Dropout(p=0.0, inplace=False)
          )
        )
        (intermediate): ViTIntermediate(
          (dense): Linear(in_features=768, out_features=3072, bias=True)
          (intermediate_act_fn): GELUActivation()
        )
        (output): ViTOutput(
          (dense): Linear(in_features=3072, out_features=768, bias=True)
          (dropout): Dropout(p=0.0, inplace=False)
        )
        (layernorm_before): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
        (layernorm_after): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
      )
    )
  )
  (layernorm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
  (pooler): ViTPooler(
    (dense): Linear(in_features=768, out_features=768, bias=True)
    (activation): Tanh()
  )
)
```

Machine Learning Institute

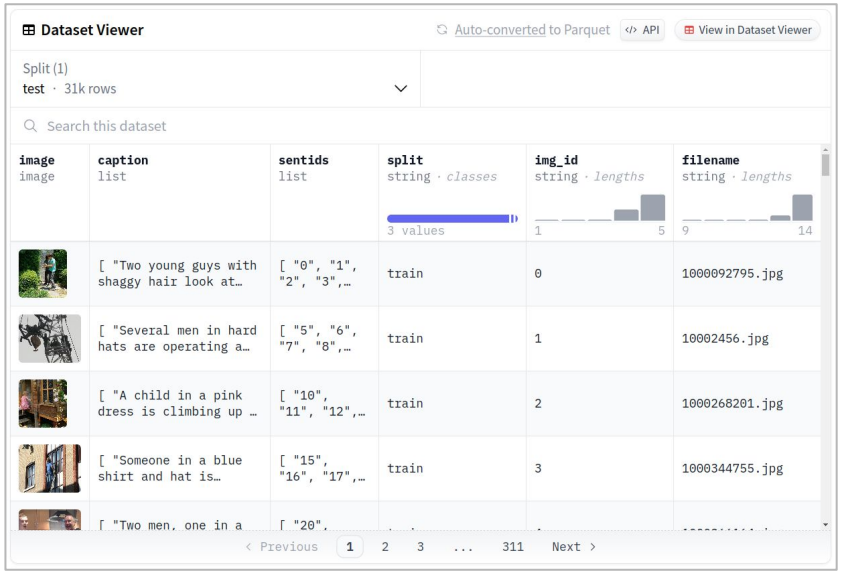
FOUNDERS AND CODERS

CLIP

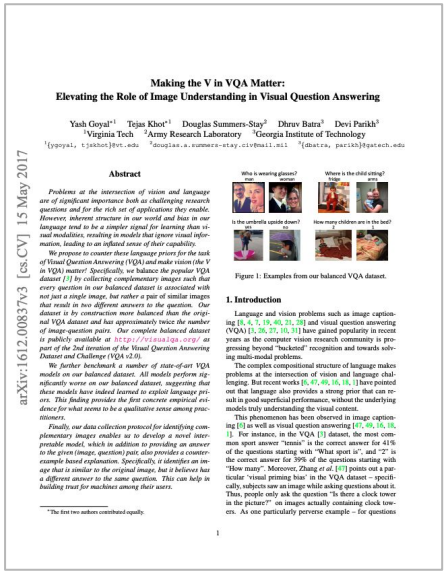
Machine Learning Institute

```
CLIPModel(
  (text_model): CLIPTextTransformer(
    (embeddings): CLIPTextEmbeddings(
      (token_embedding): Embedding(49408, 512)
      (position_embedding): Embedding(77, 512)
    )
    (encoder): CLIPEncoder(
      (layers): ModuleList(
        (0-11): 12 x CLIPEncoderLayer(
          (self_attn): CLIPSdpaAttention(
            (k_proj): Linear(in_features=512, out_features=512, bias=True)
            (v_proj): Linear(in_features=512, out_features=512, bias=True)
            (q_proj): Linear(in_features=512, out_features=512, bias=True)
            (out_proj): Linear(in_features=512, out_features=512, bias=True)
          )
          (layer_norm1): LayerNorm((512,), eps=1e-05, elementwise_affine=True)
          (mlp): CLIPMLP(
            (activation_fn): QuickGELUActivation()
            (fc1): Linear(in_features=512, out_features=2048, bias=True)
            (fc2): Linear(in_features=2048, out_features=512, bias=True)
          )
          (layer_norm2): LayerNorm((512,), eps=1e-05, elementwise_affine=True)
        )
      )
    )
    (final_layer_norm): LayerNorm((512,), eps=1e-05, elementwise_affine=True)
  )
  (vision_model): CLIPVisionTransformer(
    (embeddings): CLIPVisionEmbeddings(
      (patch_embedding): Conv2d(3, 768, kernel_size=(32, 32), stride=(32, 32), bias=False)
      (position_embedding): Embedding(50, 768)
    )
    (pre_layrnorm): LayerNorm((768,), eps=1e-05, elementwise_affine=True)
    (encoder): CLIPEncoder(
      (layers): ModuleList(
        (0-11): 12 x CLIPEncoderLayer(
          (self_attn): CLIPSdpaAttention(
            (k_proj): Linear(in_features=768, out_features=768, bias=True)
            (v_proj): Linear(in_features=768, out_features=768, bias=True)
            (q_proj): Linear(in_features=768, out_features=768, bias=True)
            (out_proj): Linear(in_features=768, out_features=768, bias=True)
          )
          (layer_norm1): LayerNorm((768,), eps=1e-05, elementwise_affine=True)
          (mlp): CLIPMLP(
            (activation_fn): QuickGELUActivation()
            (fc1): Linear(in_features=768, out_features=3072, bias=True)
            (fc2): Linear(in_features=3072, out_features=768, bias=True)
          )
          (layer_norm2): LayerNorm((768,), eps=1e-05, elementwise_affine=True)
        )
      )
    )
    (post_layernorm): LayerNorm((768,), eps=1e-05, elementwise_affine=True)
  )
  (visual_projection): Linear(in_features=768, out_features=512, bias=False)
  (text_projection): Linear(in_features=512, out_features=512, bias=False)
)
```

Datasets



Flickr30k



Making the V in VQA Matter:
Elevating the Role of Image Understanding in Visual Question Answering

Yash Goyal[*1] Tejas Khot[*1] Douglas Summers-Stay[2] Dhruv Batra[3] Devi Parikh[3]
[1]Virginia Tech [2]Army Research Laboratory [3]Georgia Institute of Technology
[1]{ygoyal, tjskhot}@vt.edu [2]douglas.a.summers-stay.civ@mail.mil [3]{dbatra, parikh}@gatech.edu

Abstract

Problems at the intersection of vision and language are of significant importance both as challenging research questions and for the rich set of applications they enable. However, inherent structure in our world and bias in our language tend to be a simpler signal for learning than visual modalities, resulting in models that ignore visual information, leading to an inflated sense of their capability.

We propose to counter these language priors for the task of Visual Question Answering (VQA) and make vision (the V in VQA) matter! Specifically, we balance the popular VQA dataset [3] by collecting complementary images such that every question in our balanced dataset is associated with not just a single image, but rather a pair of similar images that result in two different answers to the question. Our dataset is by construction more balanced than the original VQA dataset and has approximately twice the number of image-question pairs. Our complete balanced dataset is publicly available at http://visualqa.org/ as part of the 2nd iteration of the Visual Question Answering Dataset and Challenge (VQA v2.0).

We further benchmark a number of state-of-art VQA models on our balanced dataset. All models perform significantly worse on our balanced dataset, suggesting that these models have indeed learned to exploit language priors. This finding provides the first concrete empirical evidence for what seems to be a qualitative sense among practitioners.

Finally, our data collection protocol for identifying complementary images enables us to develop a novel interpretable model, which in addition to providing an answer to the given (image, question) pair, also provides a counter-example based explanation. Specifically, it identifies an image that is similar to the original image, but it believes has a different answer to the same question. This can help in building trust for machines among their users.

*The first two authors contributed equally.

Figure 1: Examples from our balanced VQA dataset.

1. Introduction

Language and vision problems such as image captioning [8, 6, 7, 19, 40, 21, 28] and visual question answering (VQA) [3, 26, 27, 16, 31] have gained popularity in recent years as the computer vision research community is progressing beyond "bucketed" recognition and towards solving multi-modal problems.

The complex compositional structure of language makes problems at the intersection of vision and language challenging. But recent works [6, 47, 40, 16, 18, 1] have pointed out that language also provides a strong prior that can result in good superficial performance, without the underlying models truly understanding the visual content.

This phenomenon has been observed in image captioning [8] as well as visual question answering [47, 49, 16, 18, 1]. For instance, in the VQA [3] dataset, the most common sport answer "tennis" is the correct answer for 41% of the questions starting with "What sport is", and "2" is the correct answer for 39% of the questions starting with "How many". Moreover, Zhang *et al.* [47] points out a particular 'visual priming bias' in the VQA dataset – specifically, subjects saw an image while asking questions about it. Thus, people only ask the question "Is there a clock tower in the picture?" on images actually containing clock towers. As one particularly perverse example – for questions

VQAv2

SlideVQA: A Dataset for Document Visual Question Answering on Multiple Images

Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, Kuniko Saito
NTT Human Informatics Laboratories, NTT Corporation
{ryouta.tanaka.rg, kyosuke.nishida.rx, kosuke.nishida.ap, taku.hasegawa.ps, itsumi.saito.df, kuniko.saito.ku}@hco.ntt.co.jp

Abstract

Visual question answering on document images that contain textual, visual, and layout information, called document VQA, has received much attention recently. Although many datasets have been proposed for developing document VQA systems, most of the existing datasets focus on understanding the content relationships within a single image and not across multiple images. In this study, we propose a new multi-image document VQA dataset, SlideVQA, containing 2.6k+ slide decks composed of 52k+ slide images and 14.5k questions about a slide deck. SlideVQA requires complex reasoning, including single-hop, multi-hop, and numerical reasoning, and also provides annotated arithmetic expressions of numerical answers for enhancing the ability of numerical reasoning. Moreover, we developed a new end-to-end document VQA model that treats evidence selection and question answering in a unified sequence-to-sequence format. Experiments on SlideVQA show that our model outperformed existing state-of-the-art QA models, but that it still has a large gap behind human performance. We believe that our dataset will facilitate research on document VQA.

Introduction

Building intelligent agents that can read and comprehend real-world documents, such as webpages, office documents, lecture slides, etc., has been a long-standing goal of artificial intelligence. To achieve this goal, machine reading comprehension (MRC), a central task in natural language understanding, has been intensively studied. The typical definition of the MRC task is quite simple, wherein given a short natural language text as a context and a question about it, a machine reads the text and then answers the question by extracting a span from the text (Rajpurkar et al. 2016; Rajpurkar, Jia, and Liang 2018). However, this definition is far from real-world applications, such as customer service chatbots on e-commerce websites (Cui et al. 2017) and assistant systems for reading professional literature (Hong et al. 2019), in that the context is composed entirely of text, with no graphical elements.

To this end, visual question answering on document images (document VQA) has received much attention. It is a challenging vision and language task that requires methods to reason about document layout, textual content, and visual elements (Mathew, Karatzas, and Jawahar 2021; Tanaka, Nishida, and Yoshida 2021; Mathew et al. 2022). When the primary content in a document is text (e.g., e-mails and forms) and the task is to understand it on the basis of its layout information, state-of-the-art models have already achieved nearly human-level performance (Xu et al. 2021; Powalski et al. 2021). On the other hand, challenges remain when it comes to handling diverse real-world documents. First and foremost is that current models are not capable of performing reasoning across multiple images since the existing datasets focus on testing reasoning ability on a single image. Moreover, compared with humans, document VQA models still have trouble understanding documents that contain visual elements and understanding questions that require numerical reasoning (Mathew et al. 2022).

To address the above challenges, we introduce a new document VQA dataset[1], SlideVQA, for tasks wherein given a slide deck composed of multiple slide images and a corresponding question, a system selects a set of evidence images and answers the question. Slide decks are one of the most efficient document types that arrange visual and textual elements for communication. As shown in Figure 1, SlideVQA requires complex reasoning over slide images, including single-hop, multi-hop, and numerical reasoning. These reasoning skills play essential roles in MRC tasks (Yang et al. 2018; Dua et al. 2019).

Our main contributions are summarized as follows:

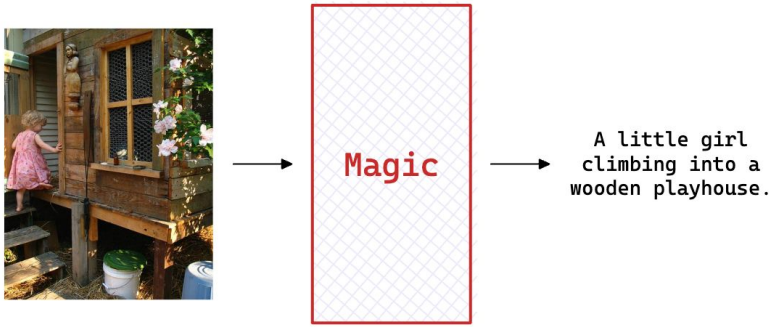
- We introduce a novel task and dataset, SlideVQA, wherein to answer its questions, a machine has to read and comprehend a slide deck. It is the largest multi-image document VQA dataset containing 2.6k+ slide decks (each consisting of 20 slides) and 14.5k questions. It also provides bounding boxes around textual and visual elements for understanding document layout and arithmetic expressions for numerical reasoning.
- We developed a Multi-Modal Multi-image Document VQA model, M3D, to jointly perform evidence selection and question answering tasks and to enhance numerical reasoning by generating arithmetic expressions.

[1]Our dataset and codes are publicly available at https://github.com/nttmdlab-nlp/SlideVQA

SlideVQA

Task

1. learn how to use the hidden state from ViT or CLIP
2. code from scratch the decoder
3. create synthetic datasets and save them on Hugging Face
   a. **get used to it!**
   b. use Qwen2.5-VL-3B-Instruct
4. train and experiment with datasets and alignment
5. **use “Qwen/Qwen3-0.6B-Base” pre-trained as decoder**



FOUNDERS AND CODERS

Good luck!

Machine Learning Institute

Recurrent Rebels

Ben Liong
Kadriye Turkcan
Rosh Beed
Joao Esteves

Gradient Gigglers

Jingyan Chen
Umut Sagir
Tyrone Nicholas
James Yan

Overfitting Overlords

Tao Zamorano
Andrew
Melanie Wong
Felipe Lavratti

Hyperparameter Hippies

Miguel Parracho
Prima Gouse
Esperanza Shi
Rasched Haidari

Perceptron Party

Yali Pan
Andrei Zhirnov
Dan Goss
Adam Beedell

Backprop Bunch

Ben Bethell
Arjuna James
Helen Zhou
Maria Sharif

Dropout Disco

Hikaru Tsujimura
Ewan Beattie
Ethan Edwards
Nikolas Kuhn

Kernel Kittens

Jacob Jenner
Marcin Tolysz
Tomas Krajcoviech
Aparna Pillai

Bayesian Buccaneers

Anton Dergunov
David Edev
James Carter
Peter O'Keeffe

Feature Fiestas

Ben Williams
Charles Cai
Clement Ha