CLIP

Anurat, Abhinandar

Abstract

Model Diagram

Architecture

Pretraining

Pros & Cons

Reference

Learning Transferable Visual Models from Natural Language Supervision

Anurat Bhattacharya Abhinandan De

IIT Kharagpur

Paper Presentation March 29, 2022



Presentation Overview

CLIP

Anurat, Abhinandan

Abstract

Model

Diagram

Architecture

Train/Tes

Pretraining Testing

Pros & Cons Pros Cons 1 Abstract

- 2 Model Diagram Architecture
- 3 Train/Test
 Pretraining
 Testing
- 4 Pros & Cons Pros Cons

Abstract

CLIP

Anurat, Abhinandan

Abstract

Model Diagram Architectur

Train/Test
Pretraining
Testing

Pros & Cons Pros Cons

Reference

- Current SOTA systems predict fixed categories.
- Our model CLIP learns from captions of images.
- Zero shot transfer after pretraining.
- Applied to > 30 CV datasets spanning common tasks.

Fun facts

- CLIP matches accuracy of original ResNet-50 on ImageNet without using any of the 1.28 M examples.
- It was developed in conjunction with DALL-E by Open-Al to evaluate the latter's performance.

Diagram

CLIP

Anurat, Abbinandai

Abstract

Model

Diagram Architectur

Train/Tes
Pretraining

Testing
Pros & Co

D 6

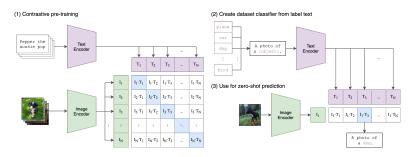


Figure: The CLIP model

Architecture

CLIF

Anurat, Abhinandar

Abstract

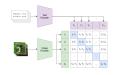
Model Diagram Architecture

Train/Tes

Pros & Cons Pros Cons

Reference:

- Text encoder is a Transformer[4] with modifications[3].
- 63 M parameters, 12-layers, 512-wide model.
- Image encoder had 2 possible architectures: ResNet[2], ViT[1].
- 5 ResNets and 3 ViT models were explored



Pretraining

CLIP

Anurat,

Abstract

Model Diagram

Architecture

Pretraining

Pros & Cons

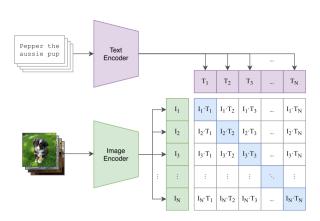


Figure: Contrastive Pretraining.

Pretraining

CLIP

Anurat, Abhinanda

Abstract

Model Diagram Architectun

Train/Tes
Pretraining
Testing

Pros & Con: Pros Cons

- Changing the caption from just a number to text
- Enabled by large amounts of publicly available data of this form
- Dataset of 400 M (img,text) pairs and trained using a simplified version of ConVIRT
- Scalability judged by training 8 models spanning 2 orders of magnitude

Testing

CLIP

Anurat, Abbinandar

Abstract

Model

Diagram Architectur

Train/Test

Pros & Cons

References

(2) Create dataset classifier from label text

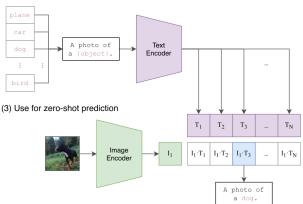


Figure: Testing phase

Zero Shot Transfer

CLIP

Anurat, Abhinanda

Abstrac

Model
Diagram

Train/Tes

Pros & Cons

- Comparison with Visual N Grams on aYahoo, ImageNet and SUN
- Comparison with a fully supervised linear classifier fitted on ResNet-50 features on 27 datasets
- A look at where Zero shot CLIP underperforms

Zero Shot Transfer



Abstract

Model Diagram

Train/Tes

Pretraining Testing

Pros & Con

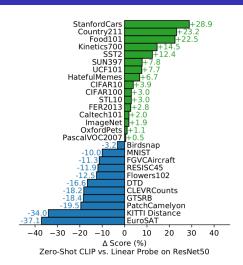


Figure: Evaluation on Different Datasets compared to fully supervised linear classifier fitted on ResNet-50 features.

Few Shot Transfer

CLIP

Anurat, Abbinandar

Abstract

Model

Diagram Architectur

Train/Tes
Pretraining

Pros & Cons

References

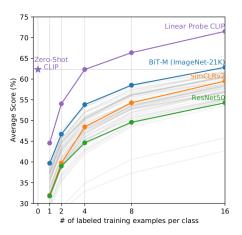


Figure: Few Shot Tranfer Comparison with different models.

Few Shot Transfer

CLIP

Anurat, Abhinandan

Ahstract

Model

Diagram Architectur

Train/Tes
Pretraining

Pros & Con

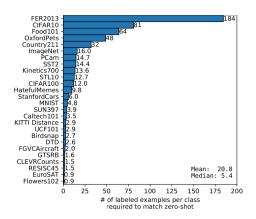


Figure: No. of labelled examples to match zero shot.

Pros

CLIP

Anurat, Abhinanda

Abstrac

Model
Diagram

Train/Tes Pretraining

Pros & Cons

- Wide Range of capabilities
- Significant Benefit for tasks that have low data
- Suitable for tasks like Image Retrieval/Search from a Database

Cons

CLIP

Anurat, Abhinanda

Abstrac

Model
Diagram

Train/Tes

Pros & Cons Pros Cons Need to improve scalability

- Poor performance on fine grained classification tasks
- Performs poorly on tasks like object detection and semantic segmentation

References

CLIF

Anurat, Abhinandar

Abstract

Model Diagram Architecture

Train/Tes
Pretraining
Testing

Pros & Con
Pros
Cons

- [1] Alexey Dosovitskiy et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv* preprint arXiv:2010.11929 (2020).
- [2] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.
- [3] Tong He et al. "Bag of tricks for image classification with convolutional neural networks". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 558–567.
- [4] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

CLIP

Anurat, Abhinandai

Abstract

Model

)iagram .rchitoctur

Train/Tes

Destruision

Pretraining

Pros & Cons

Pros Cons References Thank You!

Questions? Comments?