Dino

Bert

MOCO V3

Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

Caron M, Touvron H, Misra I, et al. Emerging properties in self-supervised vision transformers[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 9650-9660.

Slides cover:

Topic title, list of paper, names

Problem statement

Motivation

Key contributions

Technical approach overview

Experimental set ups and results

Strengths and weaknesses of the approaches and the experiments

Open research questions, possible applications, extensions

Extra credit: run a simple experimental demo

RNN:

Hard to parallel

Transformer:

一种NLP经典模型

Sequence to sequence model

使用了self-attention机制 (大量使用)

不采用RNN顺序结构，可以并行化

Bert (Transformer + Detection):

基于Transformer

Bidirectional: 双向transformer的encoder, 人们想要的肯定不止某个词左边的信息，而是左右两边的信息

DINO:

无需对数据进行标记

不指定特定目标的情况下，发现和分割图像/视频中的对象

省去以往传统的数据标注才能识别，减少计算的过程，降低训练时间，更加高效便捷地识别人或物，软硬件协作更为紧密

MAE

BEIT

BEIT v2

ViT Vision Transformer

Motivation

Key contributions

Self-supervised learning

BERT 模型成功用在 image 领域的首创

## BERT（在NLP任务中使用）

Transformer中的encoder

Transformer和Attention（带权求和）得到广泛应用

Attention解决了RNN结构的局限

SpanBERT

GPT

BERT方法用到图像

困难：

视觉任务没有一个大的词汇表，而是patch

## BEIT

组成：

BEIT encoder, 类似于 Transformer Encoder，是对输入的 image patches 进行编码的过程

dVAE, 类似于 VAE，也是对输入的 image patches 进行编码的过程

# Presentation

Hey guys, I am Jiayun. They are my teammate Lingyu and Yihao. Today, we’ll be speaking about BEiT. The full name is Bidirectional Encoder representation from Image Transformers. The three papers published recently we use are the different version of BEiT. BEiT is all you need is our topic.

First thing first, why BEiT is important, why we choose BEiT as our topic?

As the slide shows, there are two motivations.

To solve data-hungry issue because vision Transformers require more training data than convolutional neural networks

BEiT used self-supervised pre-training to leverage large-scale image data.

And, BERT has achieved great success in natural language processing.

Now the question is, why don’t we apply BERT-style pre-training for image data.

I guess that’s the reason why the BEiT model was proposed.

Comparing these two models,

BERT models are pre-trained to recover the corrupted text

BEiT models are pre-trained to predict masked visual tokens.

~~The propose of BEiT, it aims to pre-train vision transformers and relieve the dependency on labeled image data because each image has two views in the pre-training, image patches and visual tokens to train the model. BEiT is a self-supervised vision representation model.~~

This page shows the three main contributions of BEiT.

1. It is the first paper makes self-supervised pretraining of Vision Transformers (ViTs) outperform supervised pre-training.
2. It proposed a masked image modeling task to pretrain vision transformers in a self-supervised manner.
3. It presents the self-attention mechanism of self-supervised BEiT learning to distinguish semantic regions and object boundaries.

Let’s Leo to introduce the methodology.

## Strength and weakness

The pretraining model before BEiT proposed was implemented in CV by reducing image resolution

and used regression model but discretization

So that BEiT has better training effect

## Extension

Depending on the weakness, we raised some extensions that we believe can improve the BEiT model.

Masked Image Modeling can be promoted from pixel-level to semantic level.

How if we combine vision and semantic pretraining.

These two extensions lead us to explore next two papers, BEiT version 2 and version 3.

## V2

As we can see, this is the process of training VQ-KD.

VQ-KD is a Vector-Quantized Knowledge Distillation (VQ-KD) algorithm to discretize a semantic space.

Use a trained model, such as CLIP or DINO as a Teacher to guide the learning of visual tokenizer.

Version 2 inherits the masked image modeling framework defined by BEiT

The difference is that it proposed a semantic-rich visual tokenizer as the reconstruction target for masked prediction

And introduce a patch aggregation strategy which explicitly encourages the [CLS] token to associate all patches.

patch aggregation strategy which explicitly encourages the [CLS] token to associate all patches (Gao & Callan, 2021). Such a strategy resolves the issue that MIM put patch reconstruction the first place which diminishes learning global image representations

it conducted more experiments on downstream tasks than BEiT

and improves performance across model sizes, training steps, and downstream tasks with a higher accuracy results.

introducing the

The tokenizer is consist of a vision Transformer encoder, and a quantizer. The tokenizer first encodes the input image to vectors. Then, the vector quantizer looks up the nearest neighbor in the codebook for each patch representation hi . Let {v1 , v2 , · · · , vK } denote the codebook embeddings.

After training VQ-KD, its encoder is used as a semantic visual tokenizer for BEIT pretraining

Codebook

This is the overall framework of BEiT version 2.

## V3

The image is an overview of BEiT version 3 pretraining.

As the image shows, BEIT-3 is pretrained by masked data modeling on monomodal and multimodal data, using a shared Multiway Transformer network. The model can be transferred to various vision and vision-language downstream tasks.

This model makes the extension of version 1 come true.

Actually, it is a trend toward the big convergence of language, vision and multimodal pretraining, so that various downstream tasks can be transfer easily.

That concludes our presentation. Thank you for listening. If you have any question, feel free to ask.