Self-supervised Learning

Yu Xueqing

2022 年 7 月 27 日

1 name

- ELMo(Embeddings from Language Models):94M
- BERT(Bidirectional Encoder Representations from Transformers):big model,340M parameters
- ERNIE(Enhanced Representation through Knowledge Integration)
- Big Bird:Transformers for Longer Sequences
- GPT-2: 1542M
- GPT-3: 175B

2 BERT

- transformer encoder
  masked token prediction:randomly masking some tokens-换成特殊符号/随机换成另一个字
  盖住部分的输出 vector-linear-softmax-cross entropy 和 ground truth 算 loss
- next sentence prediction
- benchmark:GLUE 微调得到九个模型，执行九个任务
- downstream tasks
- how to use BERT
  - input:sequence,output:class
    example:sentiment analysis
    BERT 的参数 pre-trained 得到，具体任务的参数随机初始化，下游任务需要少量标注的资料，总体是 semi-supervised
    pre-trained+fine-tune
  - input:sequence output:same as input
    example:POS tagging(词性标注)
  - input:two sequences output:a class
    example:NLI(前提 + 假设，判断两句的逻辑关系)

  - extraction-based Question Answering(QA) 答案在文章中的问答
    input:Document,Query;output:two integers, 表示答案的范围

- training BERT is challenging,training data 大. 训练时间长
- why does BERT work: 输出词 token 的 embedding，近义词间距离近；同一个字根据上下文不同输出结果不同。学到语义，理解 token 之间的关联

3 GPT

能力：类似 tranformer 的 decoder，预测下一个 token，生成文章
how to use? few-shot learning, 举一反三，给例子，让 GPT 补完——in-context learning, 不做 training

4 Auto-Encoder

- self-supervised learning pre-trained 的一种方法 (unlabeled data)
- reconstruction:input->NN Encoder->vector->NN Decoder->output，使 output 和 input 越接近越好，类似 cycleGAN
- 使用 encoder 的输出向量 (bottleneck), 把高维向量转成低维向量 (dimension reduction)
- 高维向量的信息往往有限，可以用低维向量涵盖
- de-noising auto-encoder——将加杂讯的图片还原为原始图片
- feature disentanglement：将 encoder 输出的向量每一维对应到输入信息的特征
- 对 vector 加限制，binary/one-hot vector（classification)
- VQVAE(vector quantized variational auto-encoder)
- anomaly detection 异常检测：given a set of training data,detecting input x is similar to training data or not.
  reconstruciton loss 大——> 可能出现异常