大规模预训练

消融1：不同backbone

消融2：embedding dimensions、number of random negatives in the batch、softmax temperature

消融3：pretrain dataset

发展

NLP领域

- 不需要人工标注
- Autoregressive and masked language modeling 任务无关
- 在computer、model capacity、data方面发展
- Text-to-text 标准的input-output接口

如：

GPT-3

Devlin 2019, Radford 2020, Yang 2019, Liu 2019, Raffel 2020

Dai 2015 Peters 2018 Howard 2018

关于自然语言监督

之前nlp是使用topic model 和 n-gram不太好做跨模态训练，transformer和bert后具有Deep contextual representation learning支持

可扩展，标注简单

数据大

因为它不仅“只是”学习一种表示，而且还将这种表示与语言联系起来，可以实现zero-shot

视觉领域

显式标注数据：ImageNet 2009和openImages 2020， JFT300M 2017

Visual representation 通常在大数据集上进行预训练 2018,2020,2021

Self-supervised 2019， 2020， 2021

Semi-supervised 2019. 2020

问题：但是对下游任务迁移有限 2020

视觉语言领域

利用image and natural language caption是学习视觉表示的另一个方向

数据集：Conceptual Captions， MSCOCO， CLIP 需要清洗数据

预训练：Lu 2019，Chen 2020, Li 2020

预训练数据集：Conceptual Captions 2018、Visual Genome Dense Captions (2016) 和 ImageBERT (2020) , 需要人工标注，语义解析，清理和平衡，1000万个样本

发展：

Mori 1999

探索了通过训练模型来预测文本文档中与图像配对的名词和形容词来改进基于内容的图像检索。

Quattoni 2007

证明了通过在分类器的权重空间中的流形学习来学习更多数据高效的图像表示是可能的，这些分类器被训练成预测与图像相关联的字幕中的单词

Srivastava 2012

探索了通过在低级图像和文本标签特征之上训练多模Deep Boltzmann机器来进行深度表示学习。

Joulin 2016 【Flickr数据集】

使这一工作路线现代化，并证明了CNN被训练来预测图像标题中的单词，学习有用的图像表示

Li 2017【Flickr数据集】

将这种方法扩展到预测单个单词之外的短语ngram，并展示了他们的系统的零镜头转移到其他图像分类数据集的能力，方法是基于他们学习的视觉n-gram字典对目标类进行评分，并预测得分最高的一个。

基于transformer

VirTex Desai 2020 transformer-based language modeling 自回归

ICMLM Bulent 2020 masked language modeling 完型填空

ConVIRT Zhang 2020 contrastive objectives 医疗

缺点：规模太小

弱监督：

相反，范围更窄但目标明确的弱监督使用提高了性能。Mahajan2018表明，在35 亿 Instagram 图像上预测与 ImageNet 相关的hashtag，是一项有效的预训练任务。

Kolesnikov2019 Dosovitskiy 2020还通过预训练模型来预测嘈杂标记的 JFT–300M 数据集的类别，在更广泛的传输基准上展示了巨大的收益。

缺点：规模太小

Gomez 2017

Desai 2020 【Coco】

Sariyildiz 2020 【Coco】

问题：但数据集太小, 跨模态检索能力不足

Visual–semantic embedding VSE Frome 2013, Faghri 2018

改进版 (leveraging object detector or dense feature maps) socher 2014, karpathy 2014, kiros li 2019 chen 2020

跨模态注意力层 lu2019 chen2020

问题：慢

比较像的工作是CLIP Radford 2021

不同的encode

不同的数据集

而 CLIP 通过首先从英语维基百科构建高频视觉概念的许可列表来收集数据集。

以前数据集：

MSCOCO 高质量的人群标记数据集100,000 张训练照片

Visual Genome 高质量的人群标记数据集100,000 张训练照片

YFCC100M 1亿 每个图像的元数据都很稀疏且质量参差不齐

在过滤以仅保留带有自然语言标题和/或英文描述的图像后，数据集缩小了 6 倍，仅包含 1500 万张照片。

关于representation learning capabilities

无监督学习更关注representation learning capabilities （学一种泛化能力好的特征，但下游任务还要调，还有distribution shift问题）

我们鼓励研究zero-shot作为衡量机器学习系统task learning能力的一种方式，是收到NLP领域启发，Liu2018

数据集评估特定分布上任务的性能，但是现在研究界都是通用的

所以一些数据集上，CLIP测试的是distribution shift 和 domain generalization 而不是task generalization

对比学习

moco 正样本第一位

clip正样本对角线

对称loss

simCLR，BYOL， moco v3, DINO

非线性投射层比线性投射层提升10个点？只是适配纯图像?

对称交叉熵优化目标演变过程

Sohn2016提出度量学习的multi-class N-pair loss

Oord 作为InfoNCE loss对比表示学习

Zhang 2020等人改编用于医学成像领域的对比（文本、图像）表示学习

大规模预训练

https://lilianweng.github.io/posts/2021-09-25-train-large/

http://km.vivo.xyz/pages/viewpage.action?pageId=676707050

mobileVIT

http://km.vivo.xyz/pages/viewpage.action?pageId=558680481

MAE

nlp

transformer, GPT, BERT

Auto-regressive GPT

掩码模型/denosing autoencoder去噪自编码 bert

视觉

Auto-regressive iGPT 上半像素逐渐预测下半像素

自编码方式 MAE

patch整个掩码

encoder 只输入未被掩码部分

decoder 编码后的特征、被掩码的标志特征、1-D位置编码

轻量网络

mobile-VIT

https://zhuanlan.zhihu.com/p/421338571

distribution shift

演变

Taori 2020 研究对ImageNet模型的量化和理解

natural distribution shift【只看】

研究在对自然分布变化进行评估时，ImageNet 模型的性能如何变化。他们测量了一组 7 个分布变化的性能：ImageNetV2（Recht 等人，2019）、ImageNet Sketch（Wang 等人，2019）、Youtube-BB 和

ImageNet–Vid（Shankar 等人，2019）、ObjectNet（ Barbu 等人，2019）、ImageNet Adversarial（Hendrycks 等人，2019）和 ImageNet Rendition（Hendrycks 等人，2020a）。

Synthetic distribution shift

ImageNet–C、Stylized ImageNet、adversarial attacks

Resnet101在自然分布的数据集是imagenet验证的5倍错误

robustness analysis提出

effective robustness：distribution shift情况下提升

relative robustness：分布外的任意改进

厉害的网站

预训练模型

https://zhuanlan.zhihu.com/p/240007128

CLIP发展

clip解读

http://km.vivo.xyz/display/VNF/CLIP

http://km.vivo.xyz/pages/viewpage.action?pageId=680234637http://km.vivo.xyz/display/VNF/OpenAI–CLIP

代码

https://github.com/mlfoundations/open_clip

https://github.com/openai/CLIP/issues/83

clip官方zeroshot模型

指标计算方式

ViT-B/32

clip生成

StyleCLIP

CLIPDraw

clip检测

《open-vocabulary object detection via vision and language knowledge distillation》clifs?

clip字幕

ClipCap

http://km.vivo.xyz/pages/viewpage.action?pageId=680234637

http://km.vivo.xyz/pages/viewpage.action?pageId=650910484

代码：

https://github.com/DtYXs/Chinese-CLIP.git

clip改进

CoOp

http://km.vivo.xyz/pages/viewpage.action?pageId=684244159

clip-adapter

http://km.vivo.xyz/display/VNF/CLIP-Adapter

CLIP论文

动机

- 之前大多数方法都是通过固定类别的数据集进行训练，CLIP则利用自然语言作监督
- 之前类似方法数据规模太小，效果不好

优点

- 数据不再固定类别，可扩展，易获取，规模大
- 动态创建分类器，不需要重新训练
- 可以学到多模态特征，容易做zero-shot迁移

数据集WebImageText （WIT）

4亿互联网的文本图像对

500,000 个查询文本

基本查询列表是wiki中至少出现100次的单词，用bi-grams进行数据增强

每个查询中包含多达20,000对(图像、文本)对来对结果进行分类平衡

总单词数和训练GPT-2的WebText相近

模型

给定一批 N（图像，文本）对，训练 CLIP 以预测批次中 N × N 可能的（图像，文本）对中的哪一个实际发生。

学习通过联合训练图像编码器和文本编码器来最大化批次中 N 个实数对的图像和文本嵌入的余弦相似度，同时最小化 N 2 – N 个不正确配对嵌入的余弦相似度。

在这些相似性分数上优化了对称交叉熵损失

演变过程

VirTex，从头开始联合训练图像 CNN 和文本转换器来预测图像的标题

问题：计算量大，速度慢，预测太难

对比表示学习表明对比目标比预测更容易 Tian 2019

生成模型可以学到更好表示，但是需要比对比学习多一个数量级计算 chen2020

所以：

仅预测整个文本与哪个图像配对，而不是该文本的确切单词。

模型：ConVIRT简化版

- 从头训练
- representation和embedding space之间使用线性映射
- 删除文本转换函数
- 简化了图像转换函数，训练唯一数据增强：调整图像大小后的随机方形裁剪
- loss: 对称交叉熵损失
    - 温度变量τ，控制softmax中logits范围， 可学习，不作为超参数
- Backbone
    - Image encoder
        - 改进的resnet-50
            - Resnet-D He 2019
            - antialiased rect-2 blur pooling zhang 2019
            - Attention pooling 机制替换了全局平均池化层
                - 是一个单层transformer-style multi-head QKV attention
                - Query的条件是全局平均池化表示
            - 采用了Tan & Le(2019)的方法，在所有宽度、深度和分辨率上平均分配额外的计算比只将其分配到模型的一个维度要好。
        - Vision Transformer
            - 在transfomer之前对 combined patch 和 position embeddings 添加额外的归一化层
            - 略有不同的初始化方案
    - Text encoder
        - Transformer Radford 2019改的
            - 63M-parameter 12 layer 512-wide model with 8 attention heads
            - 对文本的小写字节对编码 (BPE) 表示进行操作
            - 词汇大小为 49,152
            - 为了计算效率，最大序列长度上限为 76
            - 文本序列用 [SOS] 和 [EOS] 标记括起来
            - 转换器的最高层的激活在 [EOS] 标记处被视为文本的特征表示，它被层归一化，然后

线性投影到多-模态嵌入空间。

- 对于文本编码器，我们只缩放模型的宽度，使其与计算的 ResNet 宽度增加成正比，根本不缩放深度，因为我们发现 CLIP 的性能对文本编码器的容量不太敏感

训练设置

- Backbone
  - Resnet-50
  - Resnet-101
  - Resmet-50 4x RN50x4 【efficientNet把channel、模型深度、输入大小调整】
  - Resmet-50 16x RN50x16
  - Resmet-50 64x RN50x64
  - ViT-B/32 【patch大小】
  - ViT-B/16
  - ViT-L/14
  - ViT-L/14@336px 在336尺寸finetune的ViT-L/14
- 32epoch
- Batchsize 32768
- Adam 优化器
- 解耦权重衰减正则化 (Loshchilov & Hutter, 2017) 应用于所有非增益或偏差的权重
- 余弦调度 (Loshchilov & Hutter, 2016) 衰减学习率
- 使用网格搜索、随机搜索和手动调整的组合，去训练最小模型ResNet50 1 个 epoch
- 设置初始超参数, 由于计算限制，超参数被启发式地适应更大的模型
- 可学习的温度参数τ从（Wu et al., 2018）初始化为相当于 0.07，并被剪裁以防止将 logits 缩放超过 100，我们认为这是防止训练不稳定所必需的。
- 省内存
  - 混合精度 (Micikevicius et al., 2017) 用于加速训练和节省内存
  - 为了节省额外的内存，使用了gradient checkpointing (Griewank & Walther, 2000; Chen et al., 2016)、half-precision Adam statistics (Dhariwal et al., 2020) 和半精度随机四舍五入的文本编码器权重

PROMPT ENGINEERING AND ENSEMBLING

- 数据集存在没有描述、多义词、只是一个词不是完整句子的问题
- 每个任务定制prompt text，如："A photo of a {label}."
- 集成多个分类器，节省开销

Zero-shot

步骤：计算嵌入表示，计算余弦相似度，通过温度系数缩放，通过softmax归一化为概率分布

L2-normalized input, L2-normalized weights, no bias, temperature scaling

实验

实验1：和zero-shot模型对比

数据集ImageNet、Yahoo、SUN

比较

Visual N-grams、CLIP最好模型、从头训练在YFCC100M 训练CLIP ResNet-50

VisualN-Grams 唯一一个 zero-shot 通用到标准图像分类，但在imagenet zero-shot很低

GPT-1专注于训练前作为一种迁移学习方法，提升监督微调

还有一项额外的消融实验four heuristic zero-shot transfer

该分析是 GPT-2 的基础，专注于通过零样本迁移研究语言模型的任务学习能力。

结论

不是公平的对比

实验2：prompt

实验3：zero-shot clip和linear probe resnet-50对比

数据集：27个数据集

模型：clip和imagenet预训练的linear probe 只训练分类头的resnet-50相比，16个比resnet-50好

结论：对特定任务，WIT包含每个任务不一样，效果参差；通用任务都还行；动作识别好，多了动词概念；复杂任务效果差

实验4：和few-shot方法比较

数据集：20个数据集？【因为训练样本不足16】

结论

比较了不同预训练的BiT-M、SimCLR、ResNet50

Clip zero-shot和4 shot相当，和16shot的在ImageNet-21k训练的Resnet-152x2训练的BiT-M相当

结论：对于难数据集，有一些训练样本还是有必要的

zero-shot clip 和 few-shot clip

26个数据集又逐一测了

ImageNet Clip zero-shot 和 16shot相当

实验5：zero-shot clip 和linear probe clip

27个模型

Zero-shot clip 和 完全 Linear Proble Clip比较

监督训练是zero-shot的上限，还有10-25的提升

有5个数据集表现相当

实验6：是否像GPT遵循性能-计算量/数据缩放

5个clip模型，在36个数据集，39次评估

总体遵循缩放，但不同任务混乱

实验7：特征学习

representation learning capabilities分析 全部数据

linear proble一般是固定representation extracted 只训练线性分类层

端到端地微调，比上一个好，但会掩盖在预训练的失败的学习泛化和鲁棒表示

比较线性分类层，更小的超参、标准操作和评估流程

学习66个不同的模型在27个不同数据集？？？？

12个数据集，和imagenet关联很大

27个数据集

计算效率好

结论：在12个数据集不好是因为这些数据集和imagenet关系大，很多模型都是在imagenet上预训练的

和EffcientNet比

21/27高，在imagenet，低分辨率

可能是clip没有scale-based data augmentation

实验8：Robustness to Natural Distribution Shift

深度学习模型非常擅长寻找贯穿其训练数据集的相关性和模式，从而提高分布性能。

7个自然分布数据

- 实验8-1：Zero-shot CLIP 与现有 ImageNet 模型在自然分布变化上比较【ImageNet Zero-shot CLIP】

（虽然这些结果表明零样本模型可以更加稳健，但它们并不一定意味着 ImageNet 上的监督学习会导致稳健性差距。 CLIP 的其他细节，例如其庞大而多样的预训练数据集或自然语言监督的使用也可能导致）

（与在 ImageNet 上预训练的模型相比，CLIP 的特征对任务转移更加稳健。在 ImageNet 上训练的模型的表示对他们的任务有些过拟合。）

结论：与在 ImageNet 上预训练的模型相比，CLIP 的特征对任务转移更加稳健

- 实验8-2：通过 L2 正则化逻辑回归分类器适应 ImageNet 分布后【Logistic Regression CLIP】

（尽管将 CLIP 适应 ImageNet 分布使其 ImageNet 准确度总体提高了 ，分布偏移下的平均准确度略有下降）

Imagenet和imagenetv2好

结论：CLIP对ImageNet的监督适应将ImageNet准确度提高了，但降低了平均鲁棒性

- 实验8-3：Adapt to class shift 【Adaptive Zero-shot CLIP】

(使用 CLIP，我们可以直接根据每个数据集的类名为每个数据集生成一个自定义的零样本分类器)

- 实验8-4：我们研究了从零样本到完全监督的连续统一体上的有效鲁棒性如何变化

(与现有 ImageNet 模型相比，Few-shot CLIP 还增加了有效的鲁棒性，但不如 zero-shot CLIP 鲁棒性)

实验9：和人比较

Zero-shot、Few-shot

Oxford IIT Pets dataset

5个人看3669个图，在37种猫或狗选择满足图像的类别

Clip 没有像人用到先验知识

错误：错误标注噪声数据、分布外的数据

结论：人可以利用先验知识，人难分别的CLIP也难分别

实验10：数据重叠分析

是否数据太大导致下游任务假泛化

训练前剔除 需要知道所有测试集、重新训练的代价

记录多少重叠，步骤

1、重复检测器，设置阈值，最大化召回，创建overlap clean子集

2、计算三个集合的准确率，报告ALL-Clean

3、看不懂

结论：还是CLIP本身泛化性好

实验11：social biases

bias probes

实验12：Surveillance 监控视频

局限

- 比不上sota方法，扩大规模可以弥补，但是要Scaling 1000x
- 几种细粒度分类不好，难任务不好
- out of distribution的数据泛化不好
    - 这表明 CLIP 对解决深度学习模型泛化脆弱的潜在问题几乎没有作用
- 没办法像image caption生成新颖字幕，对比和生成结合
- 没办法解决数据少问题，和自监督、自训练有前途 伪标签
- 现有数据容易引入偏见，创建新的验证基准测试zero-shot的迁移能力
- 数据集没经过过滤和清洗，会有social bias
- Few-shot学习倒退的问题

ALIGN论文

动机

- 预训练很重要
    - 视觉任务通常在大数据进行预训练，但对下游任务迁移有限
    - 视觉语言是学习视觉表示的另一个方向，但任务数据规模小
- 视觉任务依赖精选数据集
- 复杂的数据获取过程限制了数据大小，阻碍了模型的扩展（包括CLIP）

优点

- 大规模嘈杂数据集，没有昂贵的过滤或后处理步骤
- 语料库的规模也可以弥补其噪音并实现优秀性能

数据集

Conceptual Captions 数据(sharma 2018)，简单过滤后，得到的18亿图像文本数据

比cc数据集大两个数量级:1.8 bilion

简单的frequency-based filtering 过滤原则：

- Image-based filtering
  - 删除色情图片
  - 保留较短边大于200像素且纵横比小于3的图片
  - 删除具有超过1000个替代文本的图片
  - 删除了下游评估数据集的接近重复的图像（ILSVRC-2012, Flickr30K, and MSCOCO）
- Text-based filtering
  - 删除超过10个替代图像的文本
  - 删除稀有标注（原始数据集中 1 亿个最常见的一元组和二元组之外）的文本
  - 删除太短（<3 个一元组）或太长（>20 个一元组）的替代文本
  - 删除过于嘈杂文本
  - 删除过于通用而无用的文本

优点：

- 可以获取原始文本数据的自然分布
- 不需要专家知识来学习

模型

dual-encoder结构

图像和文本encoder共享潜在嵌入空间，将匹配的图像-文本对的嵌入推到一起，同时将不匹配的图像-文本对的嵌入推开

相似的物体会学visual-semantic embedding（VSE）（2013，2018）

最简单的vse形式

将配对文本视为图像的细粒度标签，我们的图像到文本对比损失类似于传统的基于标签的分类目标，text encoder生成label weight

- image encoder：efficientNet-L2
  - global pooling

    - 没有训练1x1的classification head
    - 289x289的尺寸训练
    - 先裁剪成346x346, 然后训练执行随机裁剪和随机水平翻转，测试是中心裁剪
    - embedding dimensions: 1376
- text encoder：BERT-Large
    - Bert-Large [CLS] token
    - 从训练集生成100k的词条词汇
    - 在 BERT 编码器的顶部添加了一个具有线性激活的全连接层，以匹配图像的尺寸
    - 我们使用最多64个符号的单词片段序列，因为输入文本不超过20个单字
- contrastive loss（公式化为归一化的 softmax. normalized softmax loss ）（zhai 2019， chen 2020, Musgrave2020 自监督和监督最有效的损失函数）cosine-similarity combination function
    - normalized softmax loss
    - Image-to-text 分类损失
    - Text-to-image 分类损失
    - 温度变量【 温度变量至关重要】
        - 初始为1.0
        - 可以学习
        - 在text-to-image 和 image-to-text共享
        - Label smoothing parameter 0.1
        - 因为图像和文本嵌入都是 L2 归一化的

训练设置

- 算力：1024 Cloud TPUv3 cores
- 图像和文本编码器都是从头开始训练
- Batch_size
    - 使批量负样本更有效，将来自所有计算核心的嵌入连接起来以形成更大的批量
    - 1024 个 cores， 每个cores16个正样本， 有效batchszie是16384
    - 配对的是正样本，其它随机对是负样本
- LAMB optimizer
    - weight decay ratio1e-5

    - 试过SGD和ADAM
- 学习率 warmup
    - 以 10k 步从0线性到1e−3，然后以1.2M步（~12 epochs）线性衰减到0

结果

Visual representation视觉任务 (Deng et al., 2009)

ImageNet ILSVRC−2012数据集

- 冻结image encoder只训练classification head
    - 训练和测试分辨率: 289和360
    - 初始学习率 0.1
- 完全finetune
    - 训练和测试分辨率:475和600
    - 初始学习率 0.01
- 两种训练方式其它设置
    - 数据增强
        - 训练：随机裁剪和水平翻转
        - 测试：中心裁剪
    - Batchsize：1024
    - SGD opt和momentum0.9
    - Lr: 每30代 0.2 ratio下降
    - 训练100代
    - Weight decay设为0
- 比较
    - WSL、CLIP、BiT、ViT、Meta Pseudo Labels、 NoisyStudent
    - 因为只使用600尺寸，和后面两个相比节省flops

较小的细粒度分类数据集

- 数据集
  - Oxford Flowers–102 (Nilsback & Zisserman, 2008)
  - Oxford–IIIT Pets (Parkhi et al., 2012))
  - Stanford Cars (Krause 2013)
  - Food101 (Bossard et al., 2014)
- 完全finetune
  - 训练和验证分辨率: 289和360
  - 和imagenet一样的数据增强和opt
  - Batchsize 256
  - Weight decay 1e–5
  - 1e–2 and 1e–3 respectively, with cosine learning rate decay in 20k steps
- 比较
  - BiT–L (Kolesnikov et al., 2020) and SAM (Foret et al., 2021）一样的超参设计
  - 测三次取平均

Visual task adaptation benchmark数据集 VTAB zhai 2019

- 由 19 个不同的（涵盖自然、专业和结构化图像分类任务的子组）视觉分类任务组成，每个任务有 1000 个训练样本
- hyper–parameter sweep 遵循Zhai 2019
  - 每个任务测试了50次
  - 每个任务在800图像训练，在200图像测试
  - Sweep后selected hyperparameters用于参与组合训练
- 比较
  - Bit–L
  -

视觉语言任务

Flicker30数据集

- Image–to–text 检索任务、Text–to–image 检索任务
- Zero–shot
  - Test:1K
- Fine–tuned
  - Train: 30k微调
  - Test:1K
  - Learing rate: 初始为1e–5，3k step后衰减
  - batch_size:之前16384和训练样本数量相近会出问题，改成2048
  - 按照Karpathy 2015划分得到训练和测试集

MSCOCO数据集

- Image–to–text 检索任务、Text–to–image 检索任务
- Zero–shot
  - Test:5k
- Fine–tuned
  - Train:82k+30k额外的验证图像
  - Test:5k
  - Learing rate: 初始为1e–5，6k step后衰减
  - batch_size:之前16384和训练样本数量相近会出问题，改成2048
  - 按照Karpathy 2015划分得到训练和测试集

Ficker30和MSCOCO主要和以下方法比较

- CLIP（Radford 2021）
- 跨模态注意力方法
  - ImageBERT（Qi 2020）、UNITER（chen 2020）、ERNIE–ViL（Yu 2020）、VILLA（Gan 2020）、Oscar（Li 2020）、GPO（chen 2020）

Crisscrossed Captions (CxC) 数据集

- MSCOCO的拓展
- 含human semantic similarity judgments for caption–caption, image–image, and image–caption pairs
- 支持四种模式内和模式间的检索任务：

    - image–to–text
    - text–to–image
    - text–to–text
    - image–to–image
- 支持三种语义相似性任务
    - semantic textual similarity (STS)
    - semantic image similarity (SIS)
    - semantic image–text similarity (SITS)
- Fine–tuned
    - 在MSCOCOfinetune后测试
- 主要和以下方法比较
    - VSE++ (Faghri et al., 2018), VSRN (Li et al., 2019), DEI2T (Parekh et al., 2021), and DET2T+I2T (Parekh et al., 2021)
    - 在image–to–text和text–to–image任务提升大
    - 模态内任务不好

消融

指标

- MSCOCO zero–shot retrieval
- ImageNet K Nearest–neighbor (KNN) tasks
- 这两个指标具有代表性，与其他指标具有很好的相关性

消融1：不同backbone

- image encoder: EfficientNet B1、B3、B5、B7
    - embedding dimension 640
    - EfficientNet B1–B7 添加线性激活的全连接层， B7 embedding dimensions是640
- text endoder: BERT–Mini 到 BERT–Large
- 越大性能越好
- 在图像任务，图像编码器越大越好
- 在检索任务，图像和文本编码器同样重要

消融2：embedding dimensions、number of random negatives in the batch、softmax temperature

- baseline model:
  - EfficientNet–B5 image encoder
  - BERT–Base text encoder
  - embedding dimension 640
  - all negatives in the batch
  - a learnable softmax temperature
- 结论
  - embedding dimensions 越大越好
  - number of random negatives in the batch 越高越好
  - 可学习的softmax temperature 使学习变得更容易

消融3：pretrain dataset

- 数据
  - full ALIGN training data
  - 10% randomly sampled ALIGN training data
  - Conceptual Captions (CC–3M, around 3M images)
    - 1/10 默认step
- 模型
  - EfficientNet–B7 + BERT–base
  - EfficientNet–B3 + BERT–mini
- 结论
  - 数据越多越好
  - 模型越小容易饱和