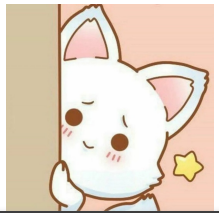
高效利用无标注数据：自监督学习简述

夕小瑶的卖萌屋 4月20日



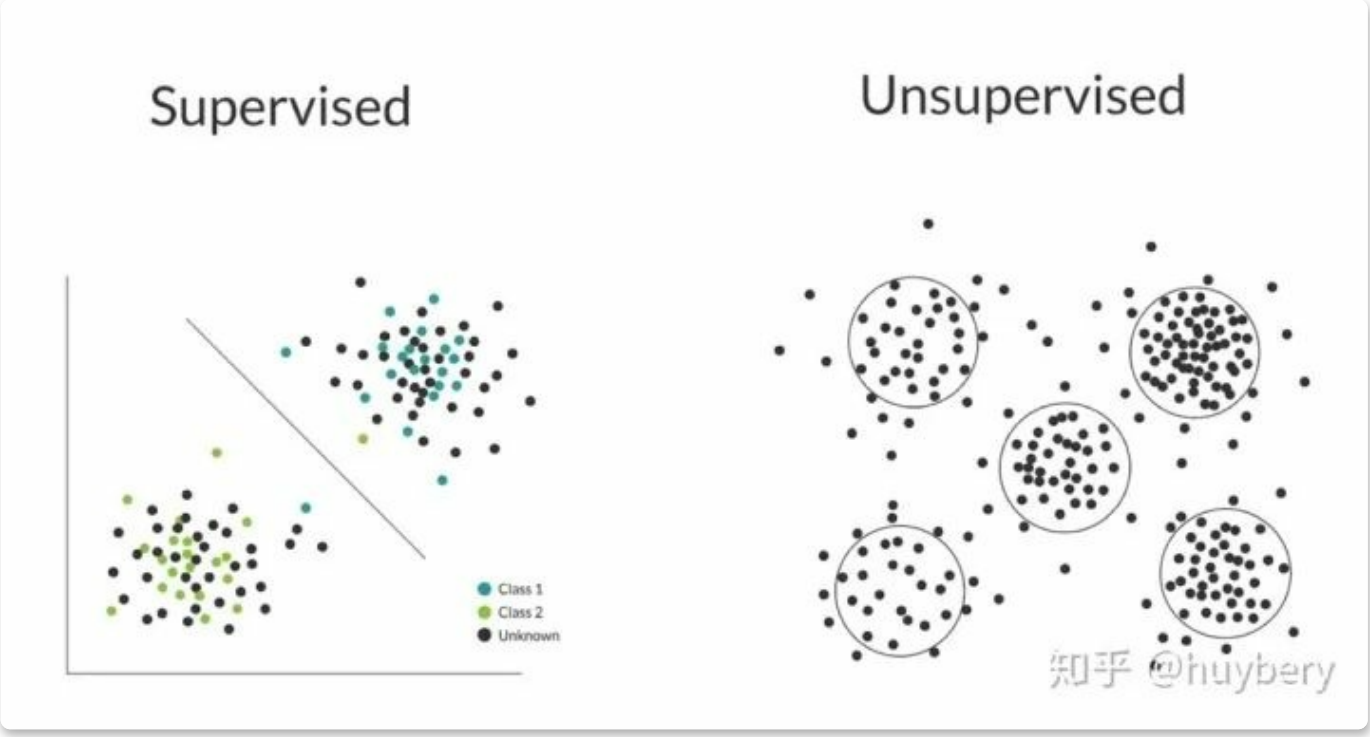
一只小狐狸带你解锁 **炼丹术&NLP** 秘籍

作者：huybery
来源：*https://zhuanlan.zhihu.com/p/108906502*

BERT的大热让自监督学习成为了大家讨论的热点，但其实word2vec和自编码器也都属于自监督学习范畴。本文通过整理自监督学习的一系列工作，把主流方法分成三大类，方便大家更全面的了解自监督学习的定义、方法、用途。

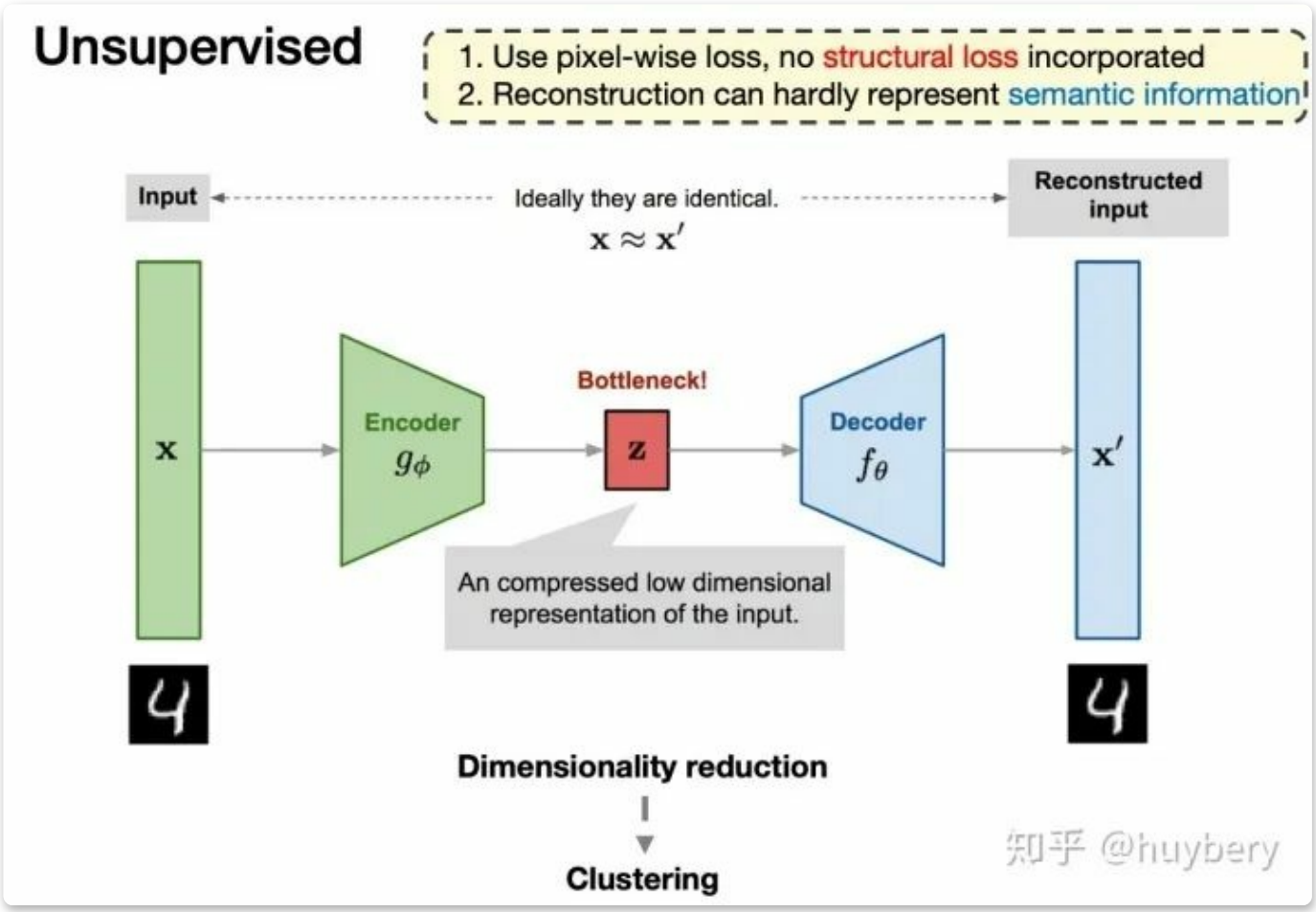
学习的范式

我们首先来回顾下机器学习中两种基本的学习范式，如图所示，**一种是监督学习，一种是无监督学习。**



监督学习利用大量的标注数据来训练模型，模型的预测和数据的真实标签产生损失后进行反向传播，通过不断的学习，最终可以获得识别新样本的能力。而无监督学习不依赖任何标签值，通过对数据内在特征的挖掘，找到样本间的关系，比如聚类相关的任务。有监督和无监督最主要的区别在于模型在训练时是否需要人工标注的标签信息。

无监督学习中被广泛采用的方式是自动编码器（autoencoder）：



编码器将输入的样本映射到隐层向量，解码器将这个隐层向量映射回样本空间。我们期待网络的输入和输出可以保持一致（理想情况，无损重构），同时隐层向量的维度大大小于输入样本的维度，以此达到了降维的目的，利用学习到的隐层向量再进行聚类等任务时将更加的简单高效。对于如何学习隐层向量的研究，可以称之为**表征学习（Representation Learning）**。

但这种简单的编码-解码结构仍然存在很多问题，基于像素的重构损失通常假设每个像素之间都是独立的，从而降低了它们对相关性或复杂结构进行建模的能力。尤其使用 L1 或 L2 损失来衡量输入和输出之间的差距其实是不存在语义信息的，而过分的关注像素级别的细节而忽略了更为重要的语义特征。对于自编码器，可能仅仅是做了维度的降低而已，我们希望学习的目的不仅仅是维度更低，还可以包含更多的**语义**特征，让模型懂的输入究竟是什么，从而帮助下游任务。而自监督学习最主要的目的就是学习到更丰富的语义表征。

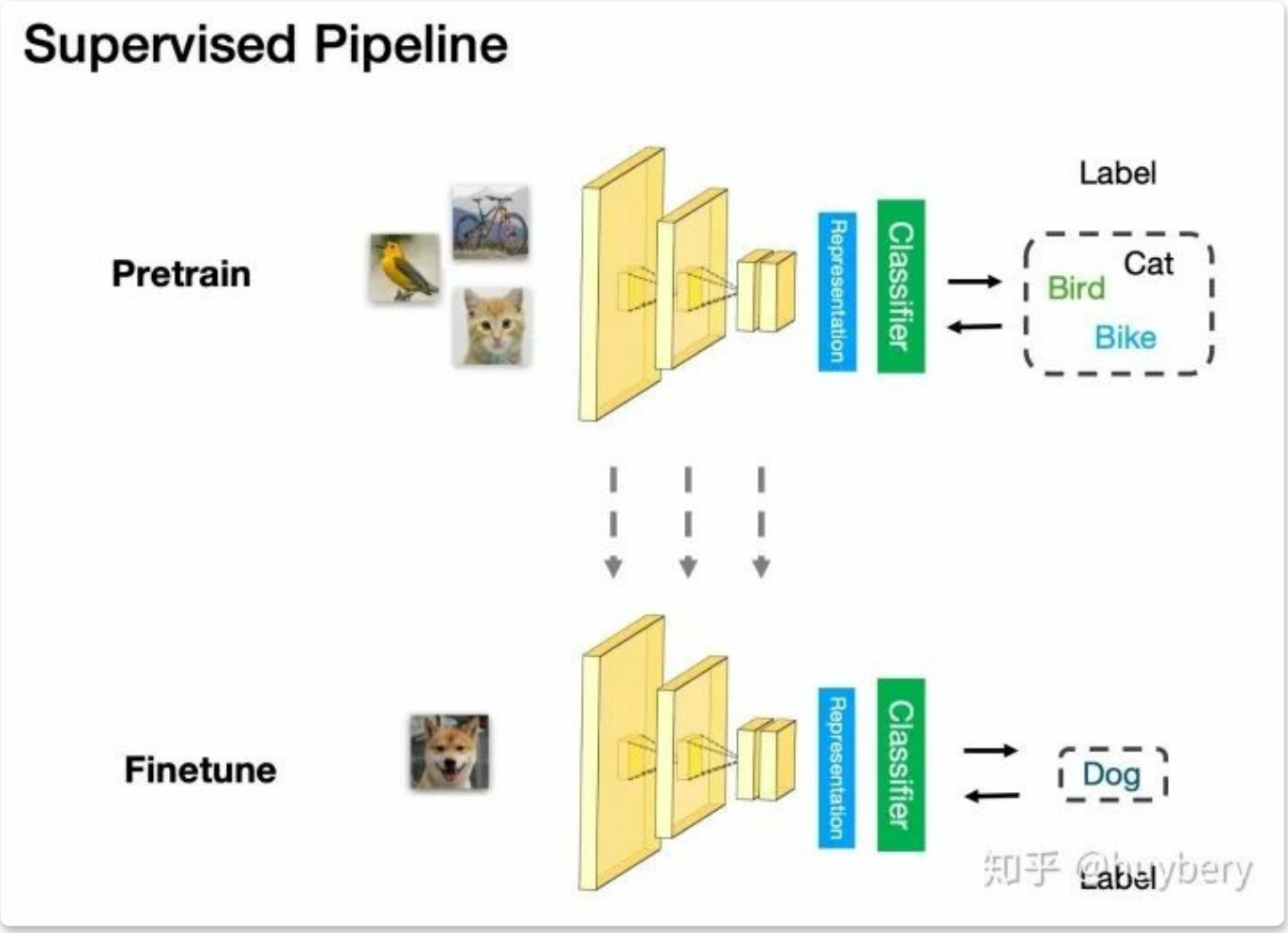
什么是自监督学习?

自监督学习主要是利用辅助任务（pretext）从大规模的无监督数据中挖掘自身的监督信息，通过这种构造的监督信息对网络进行训练，从而可以学习到对下游任务有价值的表征。

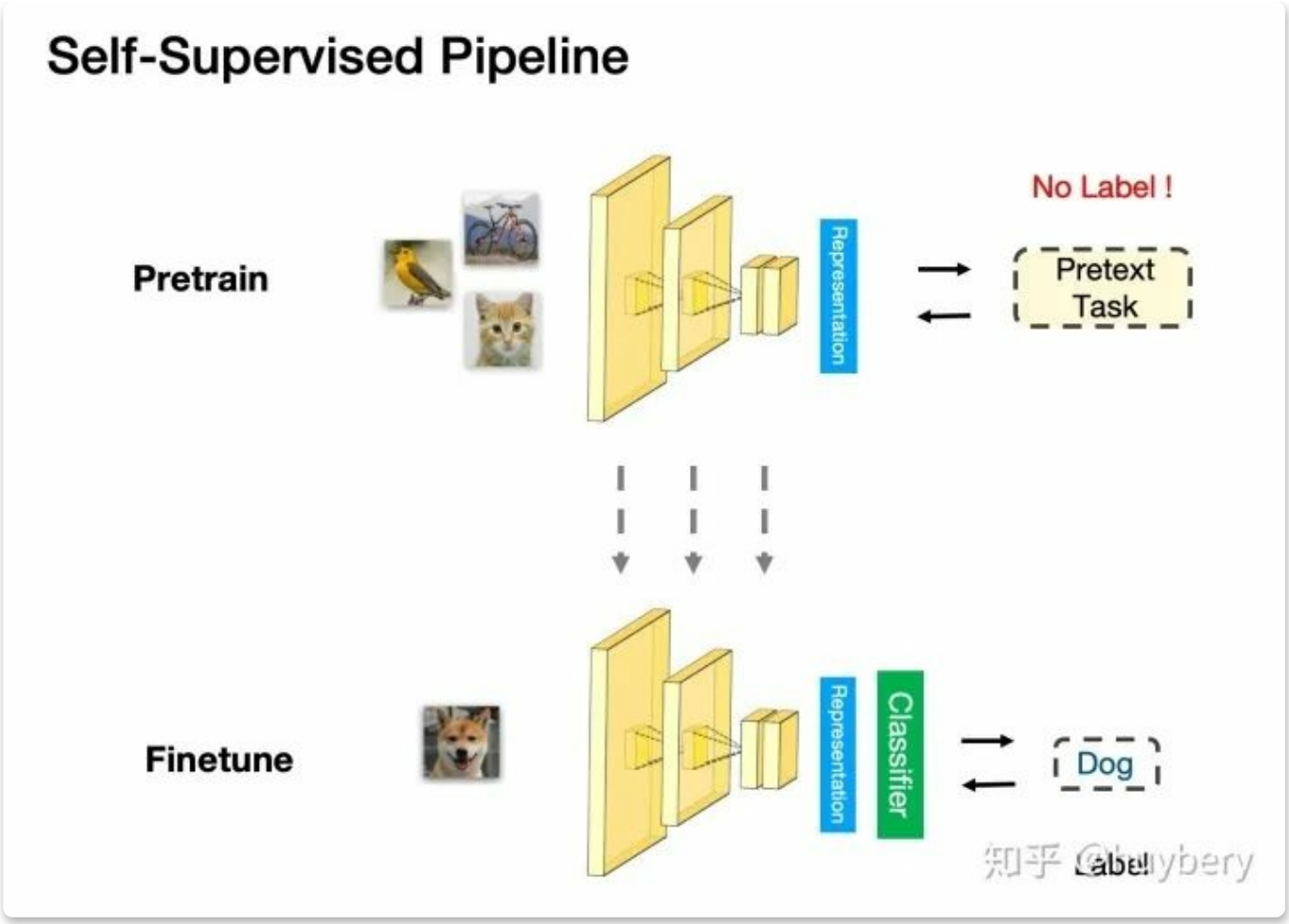
所以对于自监督学习来说，存在三个挑战：

- **对于大量的无标签数据，如何进行表征学习?**
- **从数据的本身出发，如何设计有效的辅助任务** pretext?
- **对于自监督学习到的表征，如何来评测它的有效性?**

对于第三点，评测自监督学习的能力，主要是通过 Pretrain-Fintune 的模式。我们首先回顾下监督学习中的 Pretrain - Finetune 流程：我们首先从大量的**有标签数据**上进行训练，得到预训练的模型，然后对于新的下游任务（Downstream task），我们将学习到的参数进行迁移，在新的有标签任务上进行「微调」，从而得到一个能适应新任务的网络。



而自监督的 Pretrain - Finetune 流程：首先从大量的 **无标签数据**中通过 pretext 来训练网络，得到预训练的模型，然后对于新的下游任务，和监督学习一样，迁移学习到的参数后微调即可。所以自监督学习的能力主要由下游任务的性能来体现。

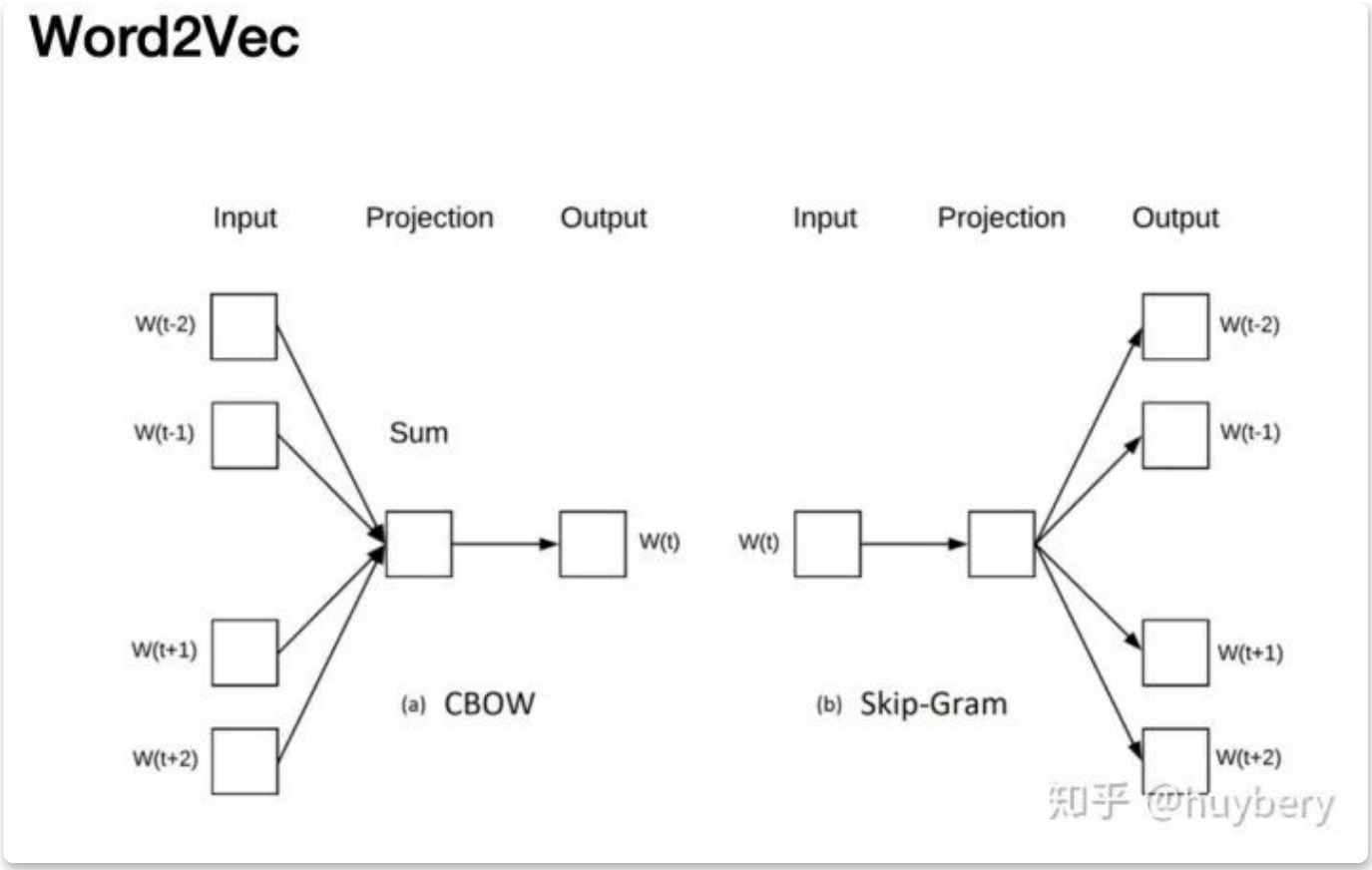


自监督学习的主要方法

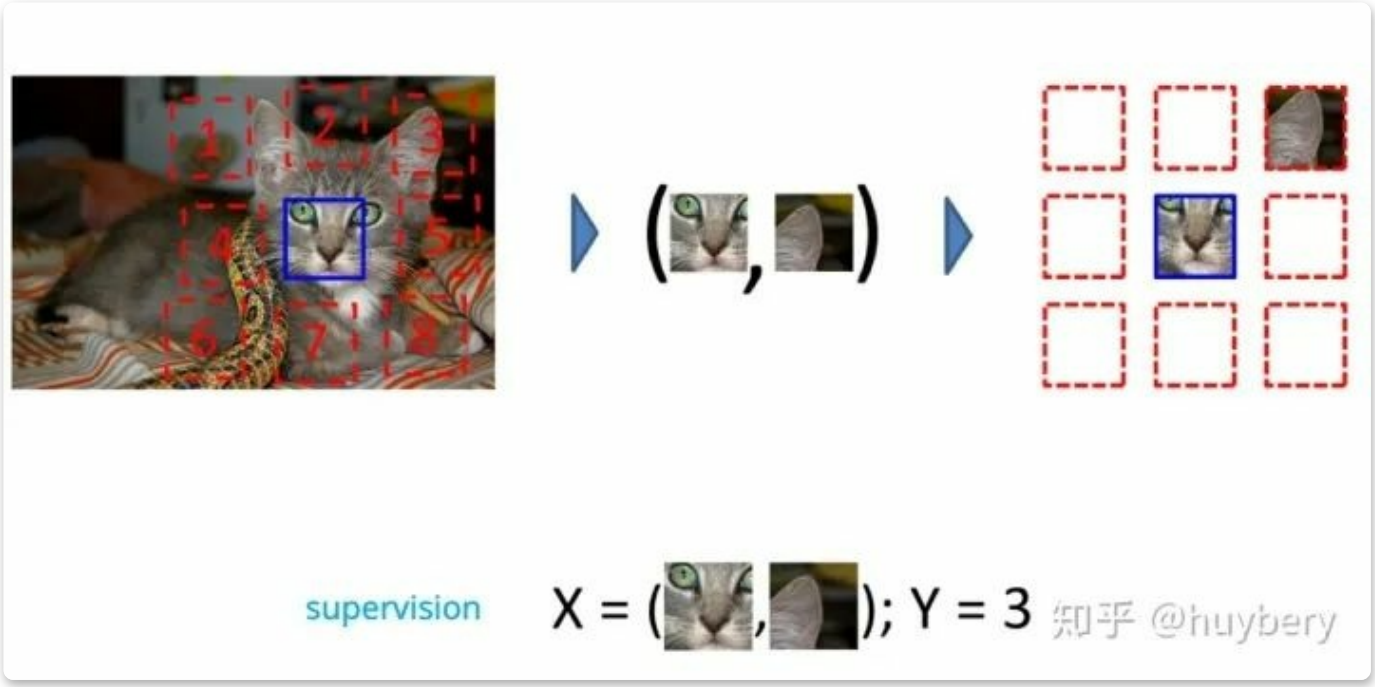
自监督学习的方法主要可以分为 3 类:基于上下文(Context based) 、基于时序(Temporal Based)以及基于对比（Contrastive Based）。

1. 基于上下文（Context Based）

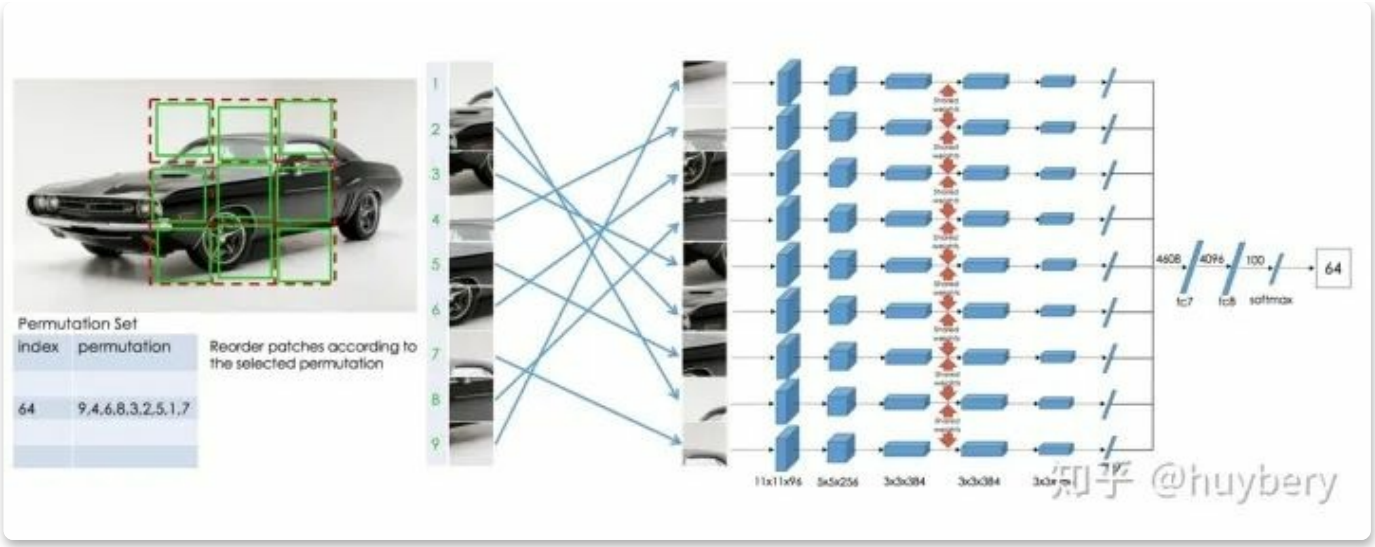
基于数据本身的上下文信息,我们其实可以构造很多任务,比如在 NLP 领域中最重要的算法 Word2vec 。Word2vec 主要是利用语句的顺序，例如 CBOW 通过前后的词来预测中间的词，而 Skip-Gram 通过中间的词来预测前后的词。



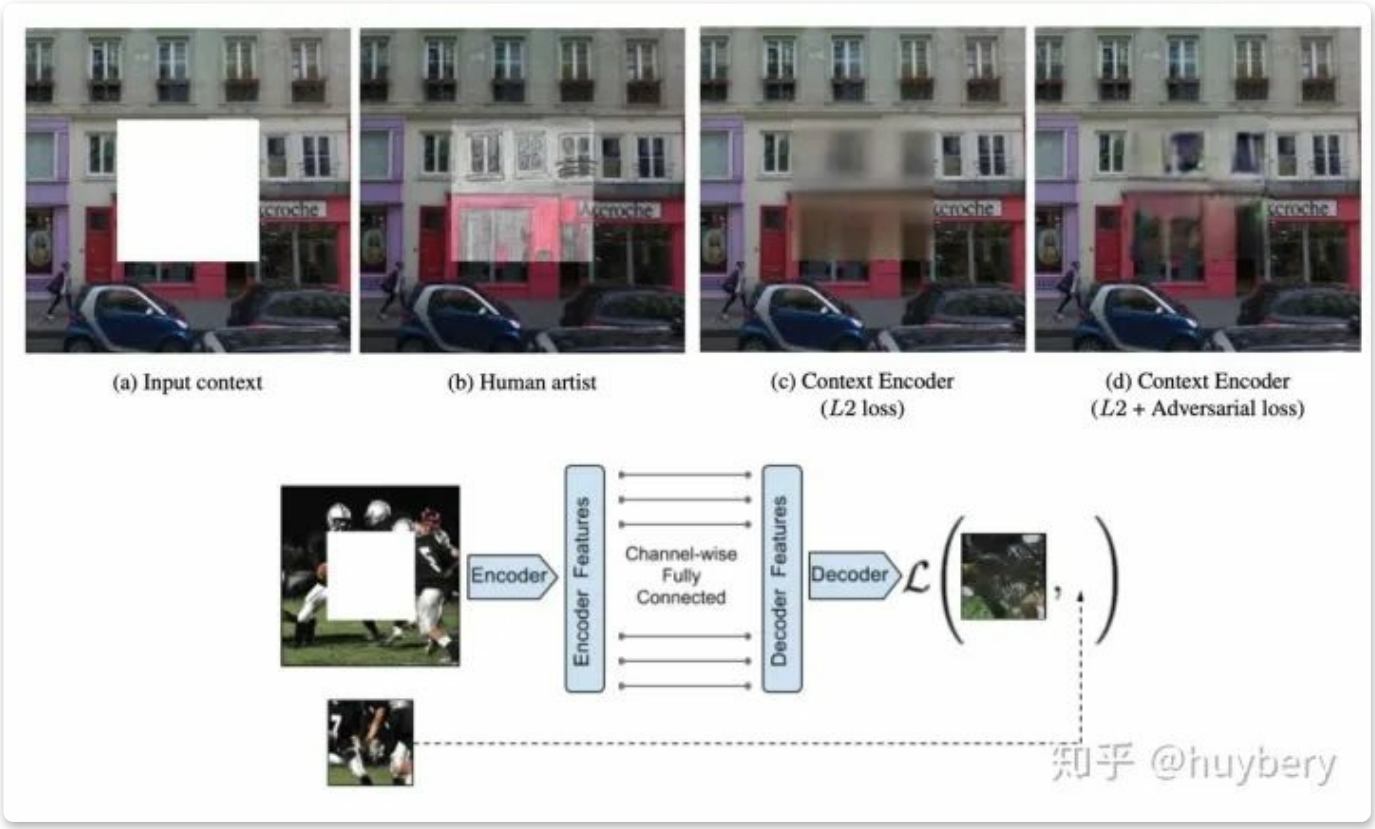
而在图像中，研究人员通过一种名为 Jigsaw（拼图）[7] 的方式来构造辅助任务。我们可以将一张图分成 9 个部分，然后通过预测这几个部分的相对位置来产生损失。比如我们输入这张图中的小猫的眼睛和右耳朵，期待让模型学习到猫的右耳朵是在脸部的右上方的，如果模型能很好的完成这个任务，那么我们就可以认为模型学习到的表征是具有语义信息的。



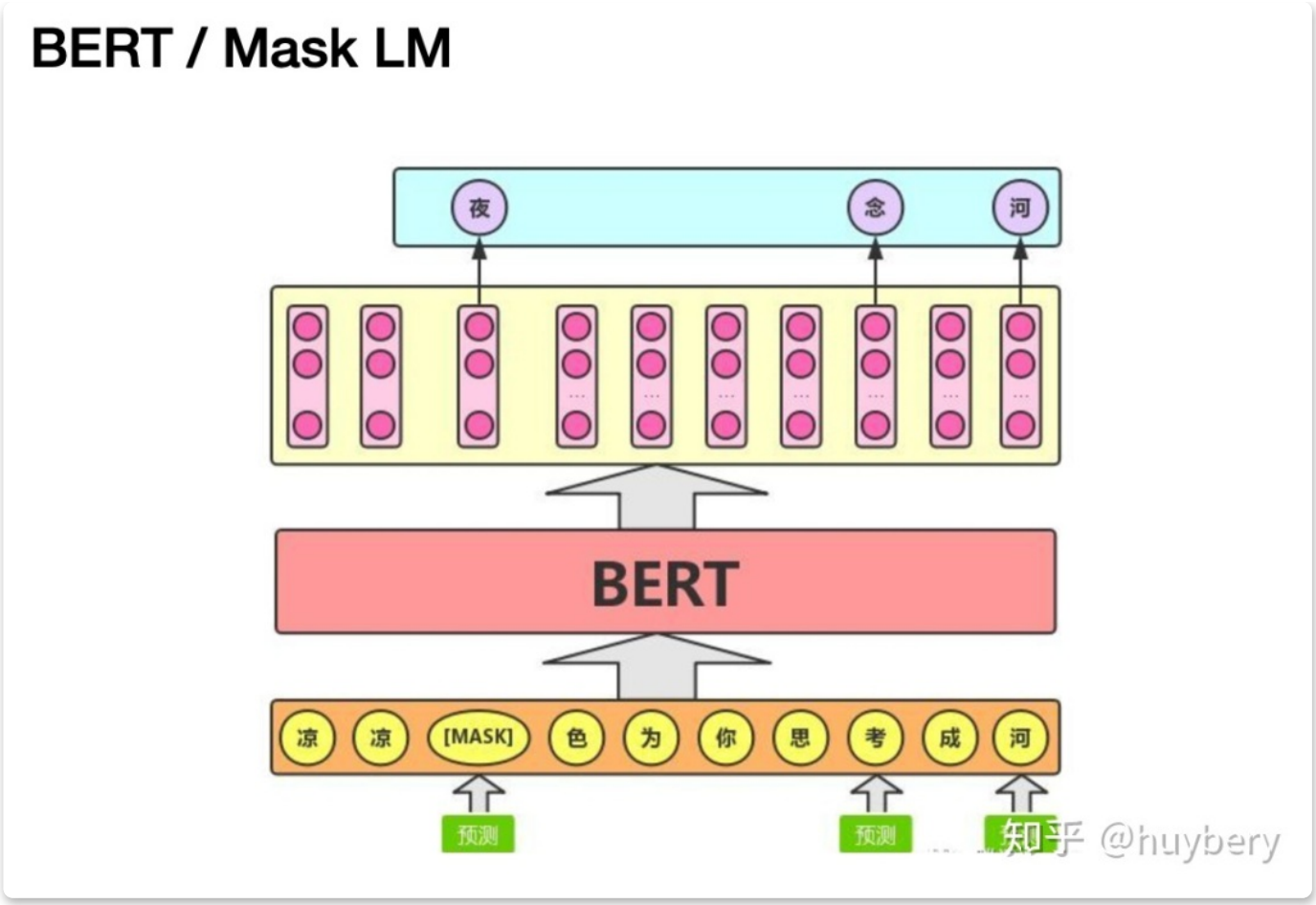
后续的工作[8]人们又拓展了这种拼图的方式，设计了更加复杂的，或者说更难的任务。首先我们依然将图片分为 9 块，我们预先定义好 64 种排序方式。模型输入任意一种被打乱的序列，期待能够学习到这种序列的顺序属于哪个类，和上个工作相比，这个模型需要学习到更多的相对位置信息。这个工作带来的启发就是使用更强的监督信息，或者说辅助任务越难，最后的性能越好。



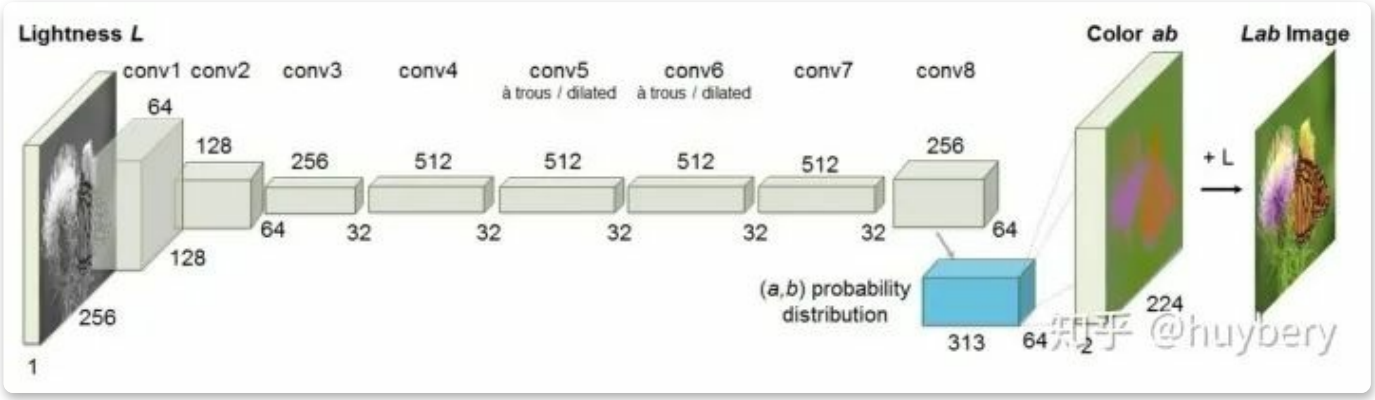
除了这种拼图的模式，还有一种是抠图[9]。想法其实也很简单粗暴，就是我们随机的将图片中的一部分删掉，然后利用剩余的部分来预测扣掉的部分，只有模型真正读懂了这张图所代表的含义，才能有效的进行补全。这个工作表明自监督学习任务不仅仅可以做表征学习，还能同时完成一些神奇的任务。



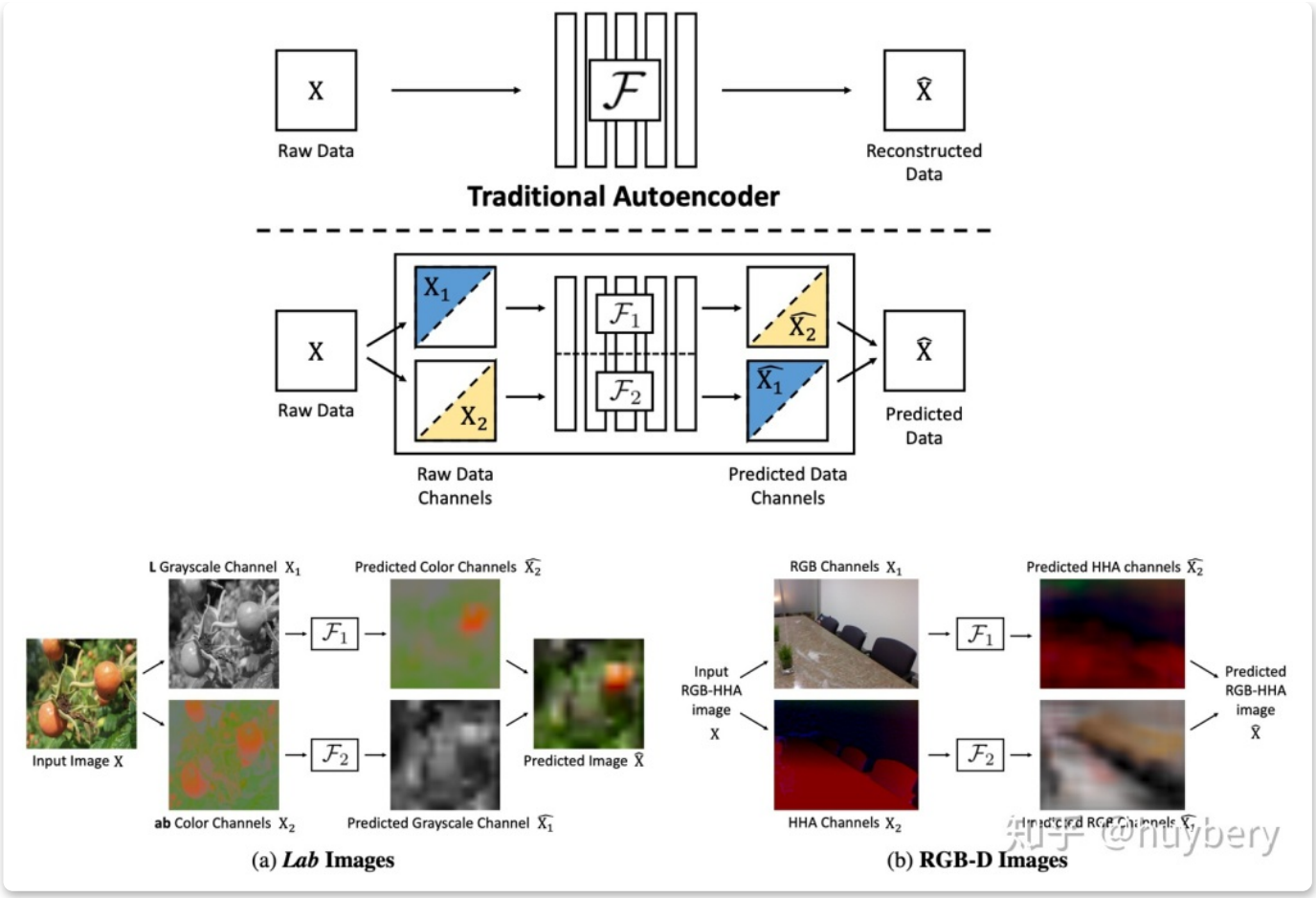
而对于这种抠图的方式，其实和 nlp 中的 BERT [10] 的 MASK LM 训练方式有异曲同工之妙，BERT 在训练时也可以是看做随机扣掉一些词，然后来预测扣掉的词，从而让模型读懂句子。



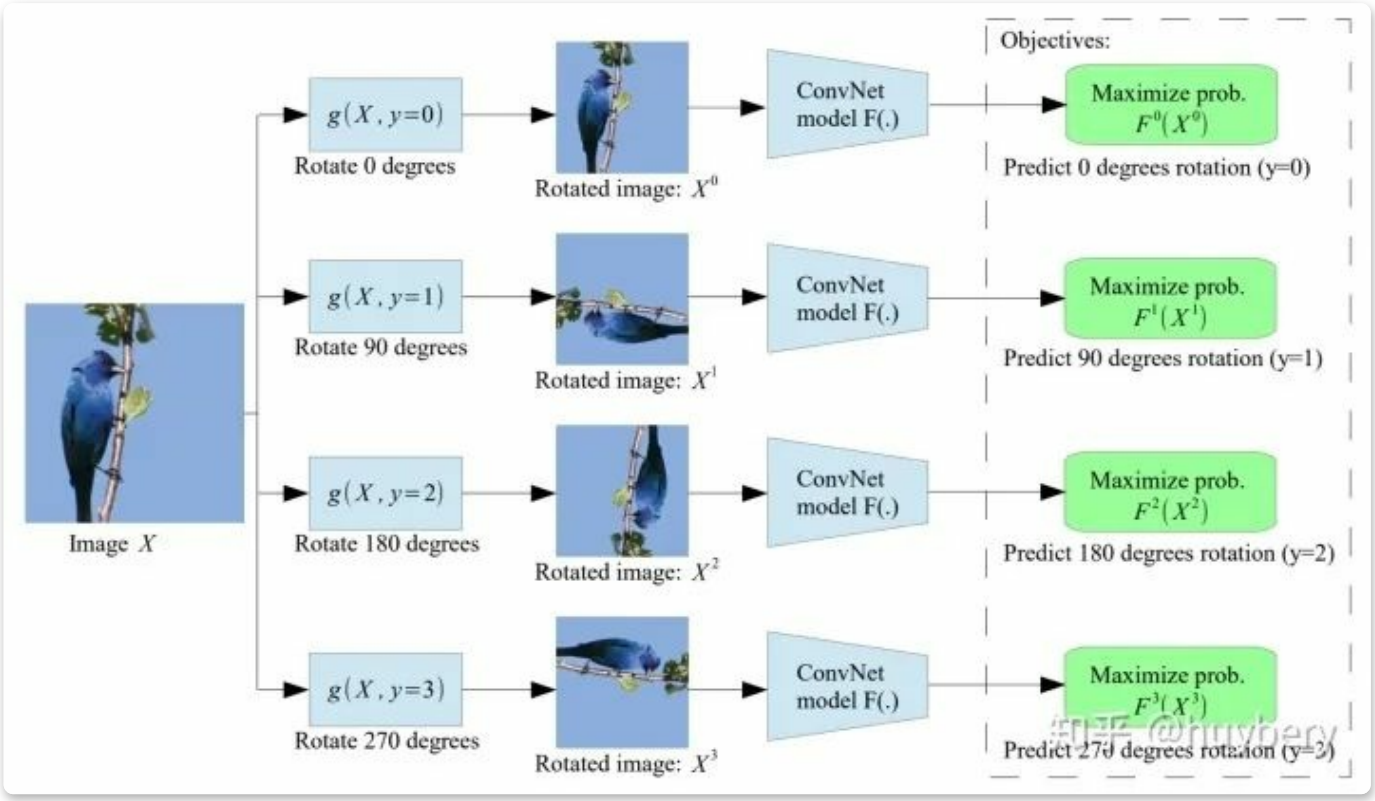
还有一种思路是通过图片的颜色信息[11]，比如给模型输入图像的灰度图，来预测图片的色彩。只有模型可以理解图片中的语义信息才能得知哪些部分应该上怎样的颜色，比如天空是蓝色的，草地是绿色的，只有模型从海量的数据中学习到了这些语义概念，才能得知物体的具体颜色信息。同时这个模型在训练结束后就可以做这种图片上色的任务。



这种基于预测颜色的生成模型带给了人们新的启发，其实这种灰度图和 ab 域的信息我们可以当做是一张图片的解耦表达，所以只要是解耦的特征，我们都可以通过这种方式互相监督的学习表征，著名的 Split-Brain Autoencoders [12] 就在做这样一件事情。对于原始数据，首先分成两部分，然后通过一部分的信息来预测另一部分，最后再合成完成的数据。和传统编码器不同的是，这种预测的方式可以促使模型真正读懂数据的语义信息才能够实现，所以相当于间接地约束编码器不单单靠 pixel-wise 层面来训练，而要同时考虑更多的语义信息。

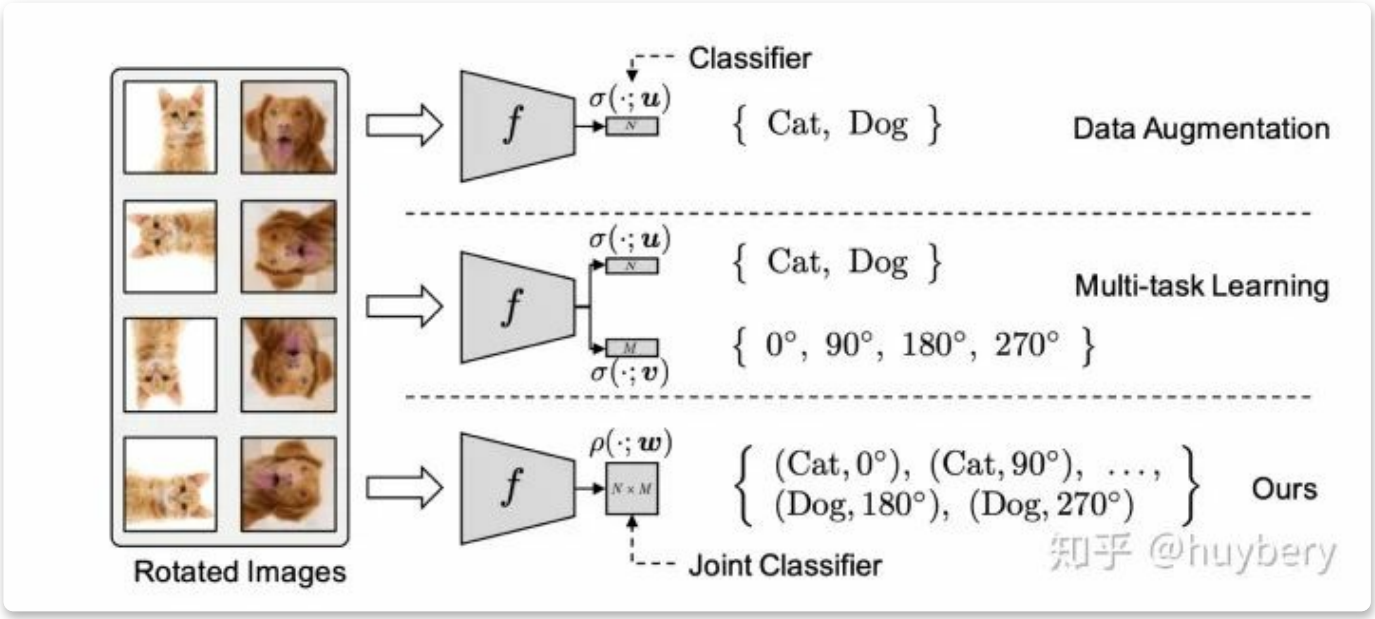


最后我们要介绍的是根据类似数据增广的方式来寻找自监督上下文。ICLR 2018 [13]的工作是给定一张输入的图片，我们对其进行不同角度的旋转，模型的目的是预测该图片的旋转角度。这种朴素的想法最后带来的增益竟然是非常巨大的，所以数据增强对于自监督学习也是非常有益处的，我个人的想法是数据增强不仅带来了更多的数据，还增加了预训练模型的鲁棒性。

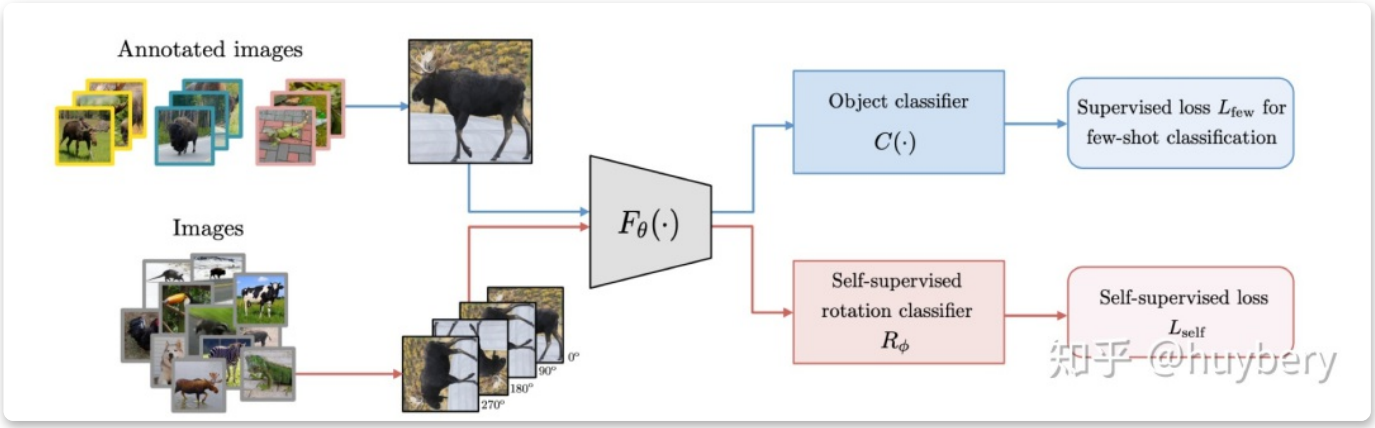


自监督学习在预训练模型中的成功让研究人员觉得非常兴奋，同时也激发了更多的灵感。我们之前介绍的模型都是在专注如何寻找自监督信息，而自监督学习一定要脱离下游的具体任务吗？答案是否定的，越来越多的工作开始思考自监督学习和具体任务紧

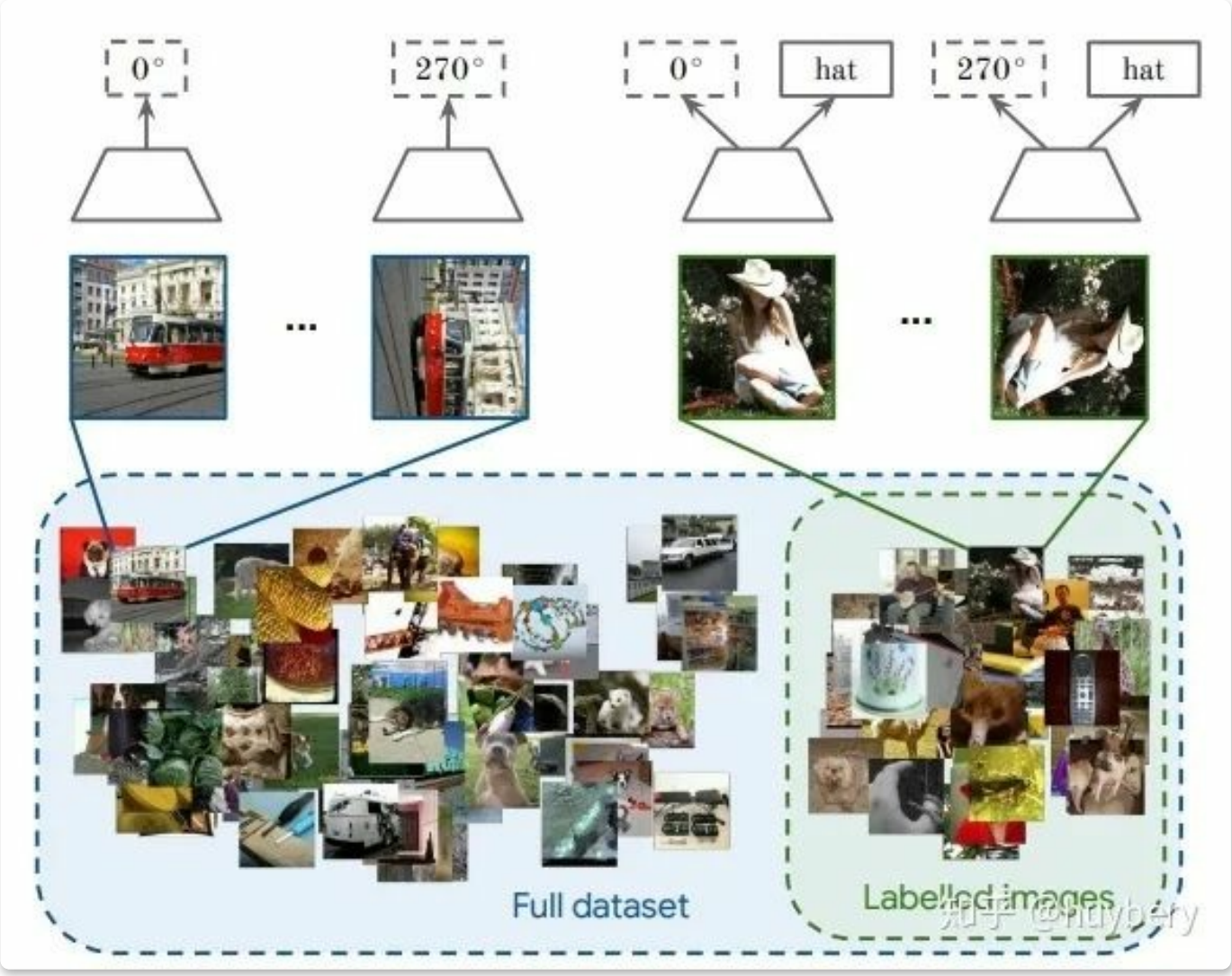
密结合的方法(**Task Related Self-Supervised Learning**)。



Lee, Hankook et al [14]探索了在多任务学习中增加自监督学习的可能，他们将普通的分类任务中嵌入了旋转预测任务。除了简单的多任务学习，也可以设计联合学习策略，直接预测两种监督信息。同样的想法也被用到了小样本学习[15]中，一个分支进行传统的小样本分类，另一个分支来进行自监督旋转预测，虽然这篇文章的想法和设计不是很亮眼，但提升还是比较明显的。



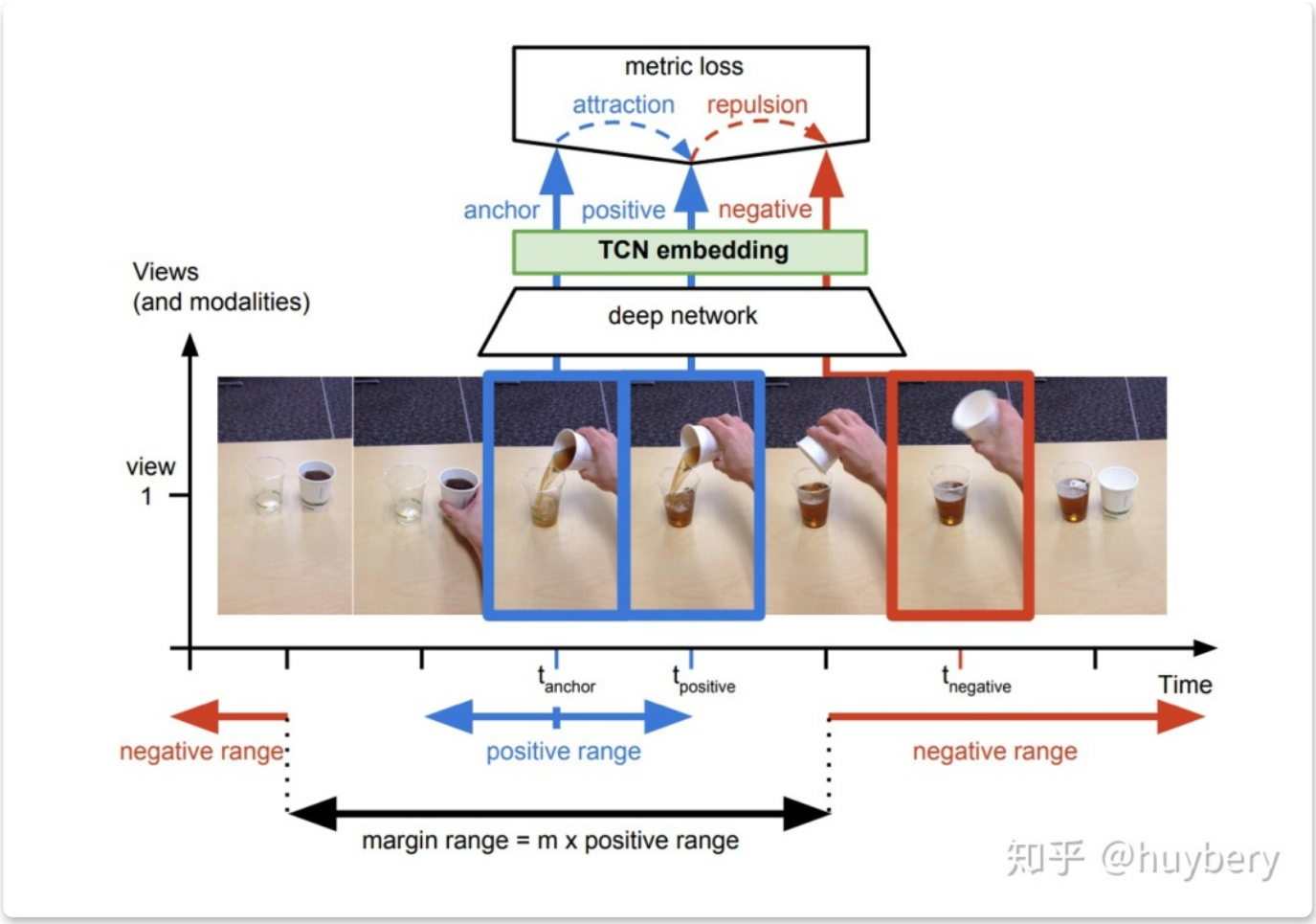
而自监督和半监督学习[16]也可以进行结合，对于无标记的数据进行自监督学习（旋转预测），和对于有标记数据，在进行自监督学习的同时利用联合训练的想法进行有监督学习。通过对 imagenet 的半监督划分，利用 10% 或者 1% 的数据进行实验，最后分析了一些超参数对于最终性能的影响。



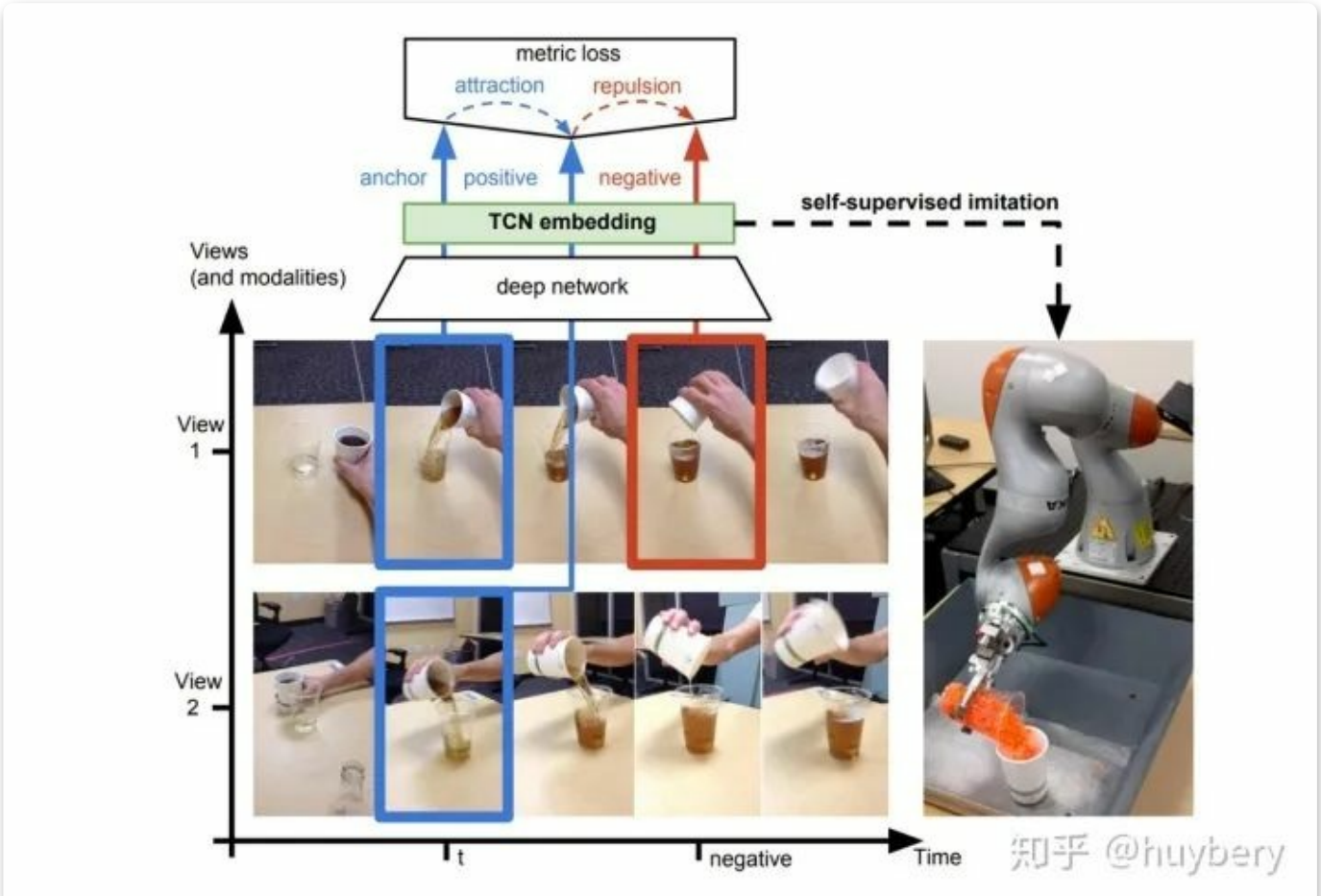
这两篇文章最后都中了 ICCV 2019，说明目前来说审稿人对于这类任务相关的自监督模型都是比较感兴趣的。

2. 基于时序（Temporal Based）

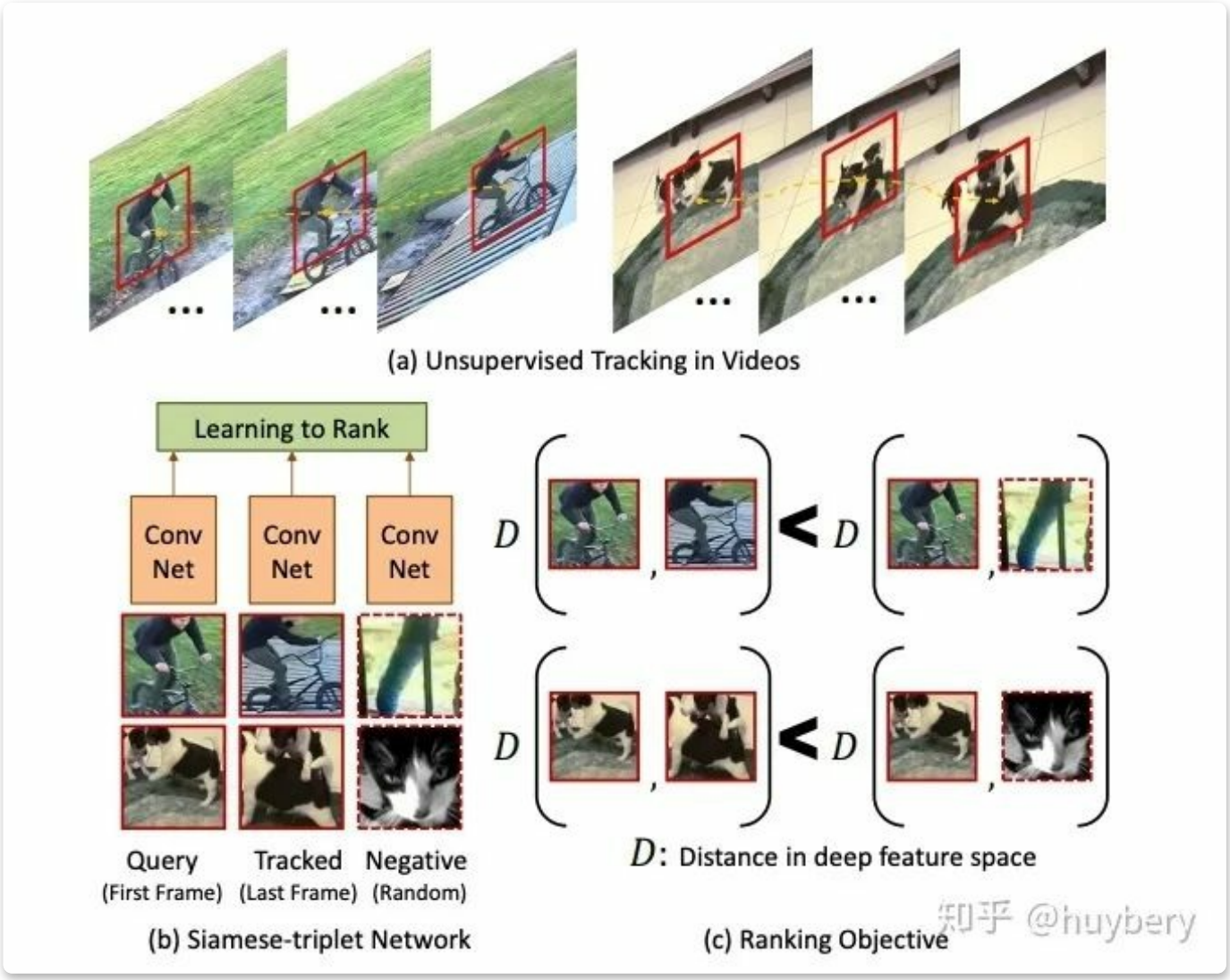
之前介绍的方法大多是基于样本自身的信息，比如旋转、色彩、裁剪等。而样本间其实也是具有很多约束关系的，这里我们来介绍利用时序约束来进行自监督学习的方法。最能体现时序的数据类型就是视频了（video）。



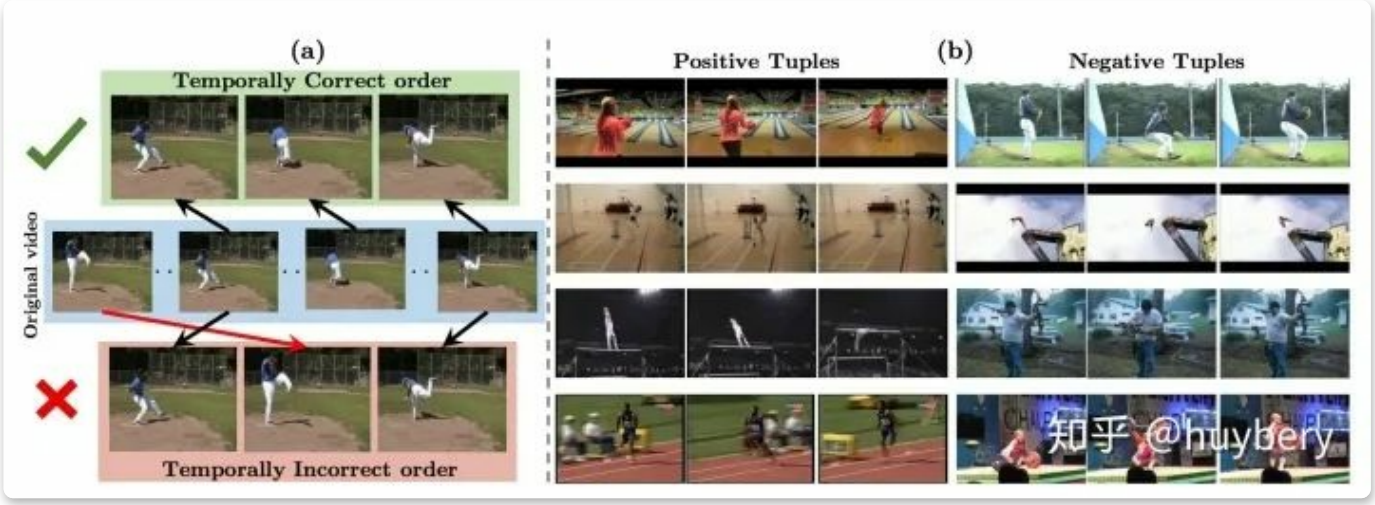
第一种思想是基于帧的相似性[17]，对于视频中的每一帧，其实存在着特征相似的概念，简单来说我们可以认为视频中的相邻帧特征是相似的，而相隔较远的视频帧是不相似的，通过构建这种相似（position）和不相似（negative）的样本来进行自监督约束。



另外，对于同一个物体的拍摄是可能存在多个视角（multi-view），对于多个视角中的同一帧，可以认为特征是相似的，对于不同帧可以认为是不相似的。

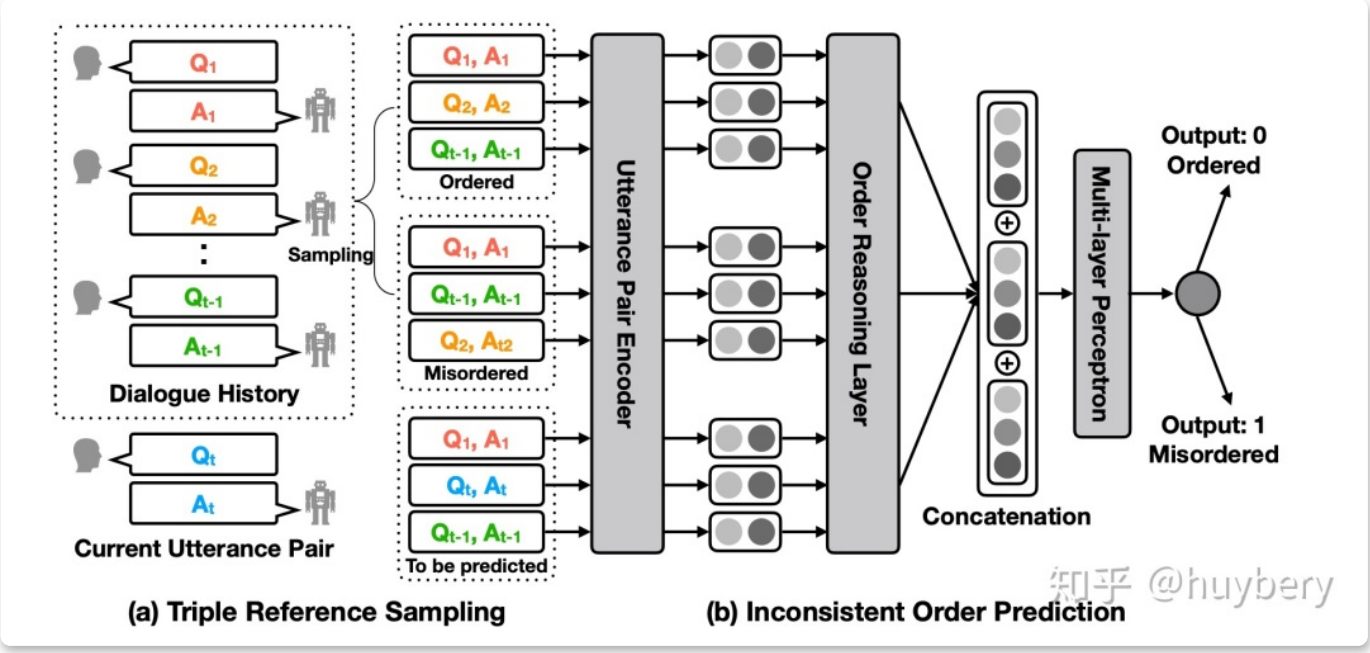


还有一种想法是来自 @Xiaolong Wang 大佬 ICCV 2015 [18]的基于无监督追踪方法，首先在大量的无标签视频中进行无监督追踪，获取大量的物体追踪框。那么对于一个物体追踪框在不同帧的特征应该是相似的（positive），而对于不同物体的追踪框中的特征应该是不相似的（negative）。

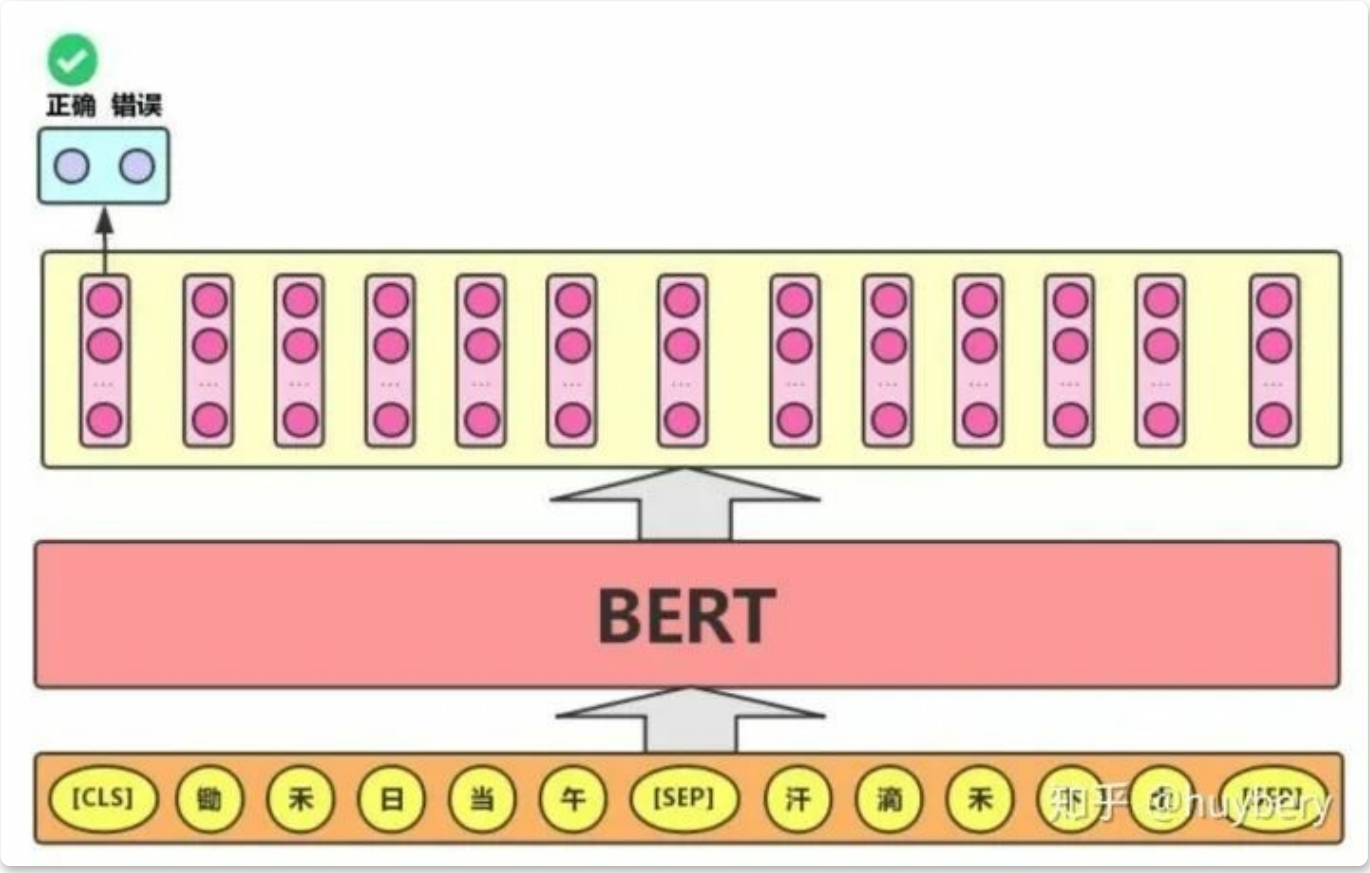


除了基于特征相似性外，视频的先后顺序也是一种自监督信息。比如ECCV 2016, Misra, I. [19] 等人提出基于顺序约束的方法，可

以从视频中采样出正确的视频序列和不正确的视频序列，构造成正负样本对然后进行训练。简而言之，就是设计一个模型，来判断当前的视频序列是否是正确的顺序。



基于顺序的约束还被应用了到了对话系统中，ACL 2019 [20] 提出的自监督对话学习就是基于这种思想。这篇文章主要是想解决对话系统中生成的话术连贯性的问题，期待机器生成的回复和人类交谈一样是符合之前说话的风格、习惯等等。从大量的历史预料中挖掘出顺序的序列（positive）和乱序的序列（negative），通过模型来预测是否符合正确的顺序来进行训练。训练完成后就拥有了一个可以判断连贯性的模型，从而可以嵌入到对话系统中，最后利用对抗训练的方式生成更加连贯的话术。



而 BERT 的另一种训练方式，Next Sentence Prediction 也可以看作是基于顺序的约束，通过构造大量的上下文样本，目的是让模型理解两个句子之间的联系。这一任务的训练语料可以从语料库中抽取句子对包括两个句子A和B来进行生成，其中50%的概率B

是A的下一个句子，50%的概率B是语料中的一个随机句子。该任务预测B是否是A的下一句。

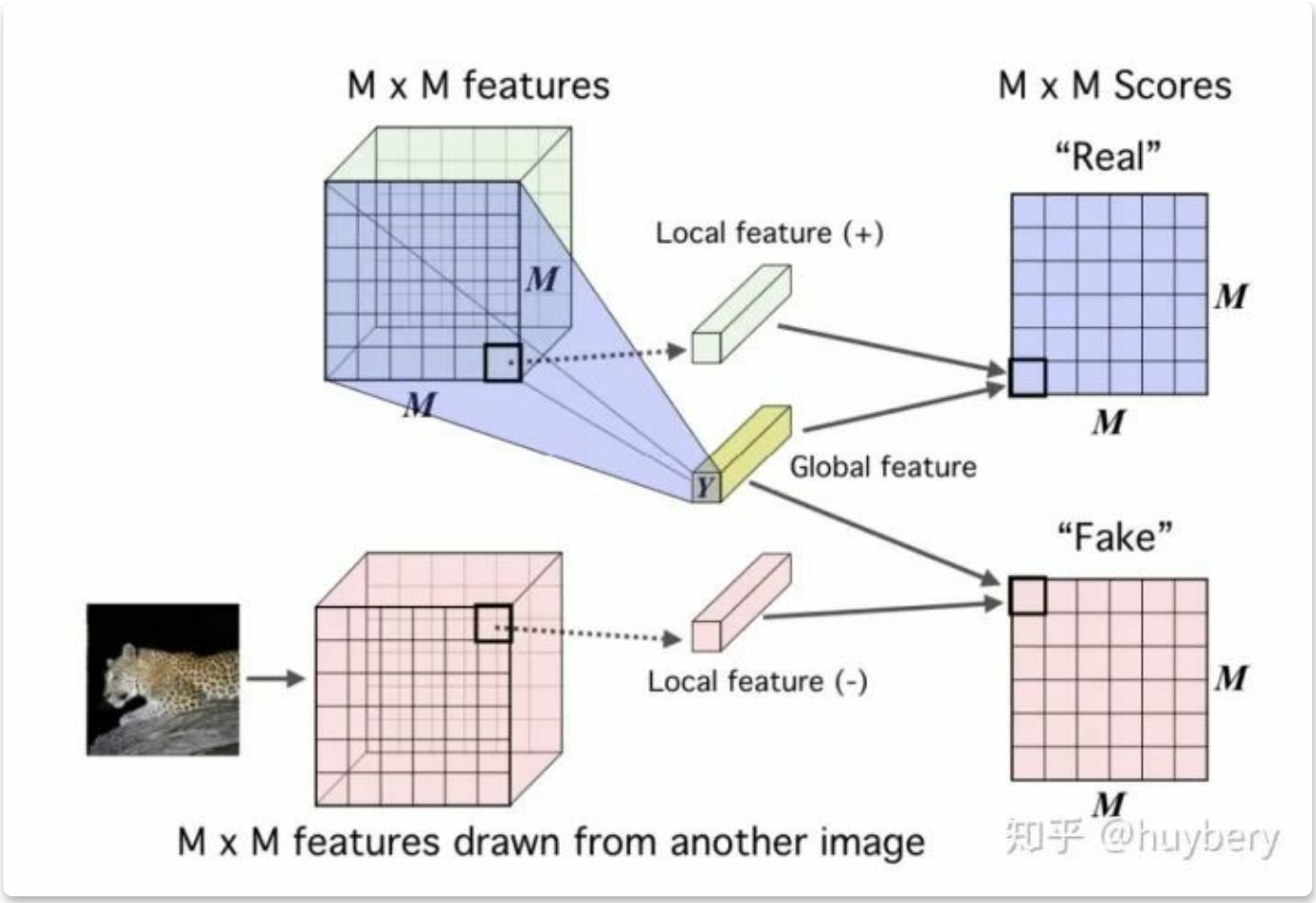
3. 基于对比（Contrastive Based）

第三类自监督学习的方法是基于对比约束，它通过学习对两个事物的相似或不相似进行编码来构建表征，这类方法的性能目前来说是非常强的，从最近的热度就可以看出，很多大牛的精力都放在这个方向上面。关于这个方向的方法，[22] 总结的比较好。这里我们再简单的阐述一下，加上一些我个人的看法。

其实我们第二部分所介绍的基于时序的方法已经涉及到了这种基于对比的约束，通过构建正样本（positive）和负样本（negative），然后度量正负样本的距离来实现自监督学习。核心思想样本和正样本之间的距离远远大于样本和负样本之间的距离：

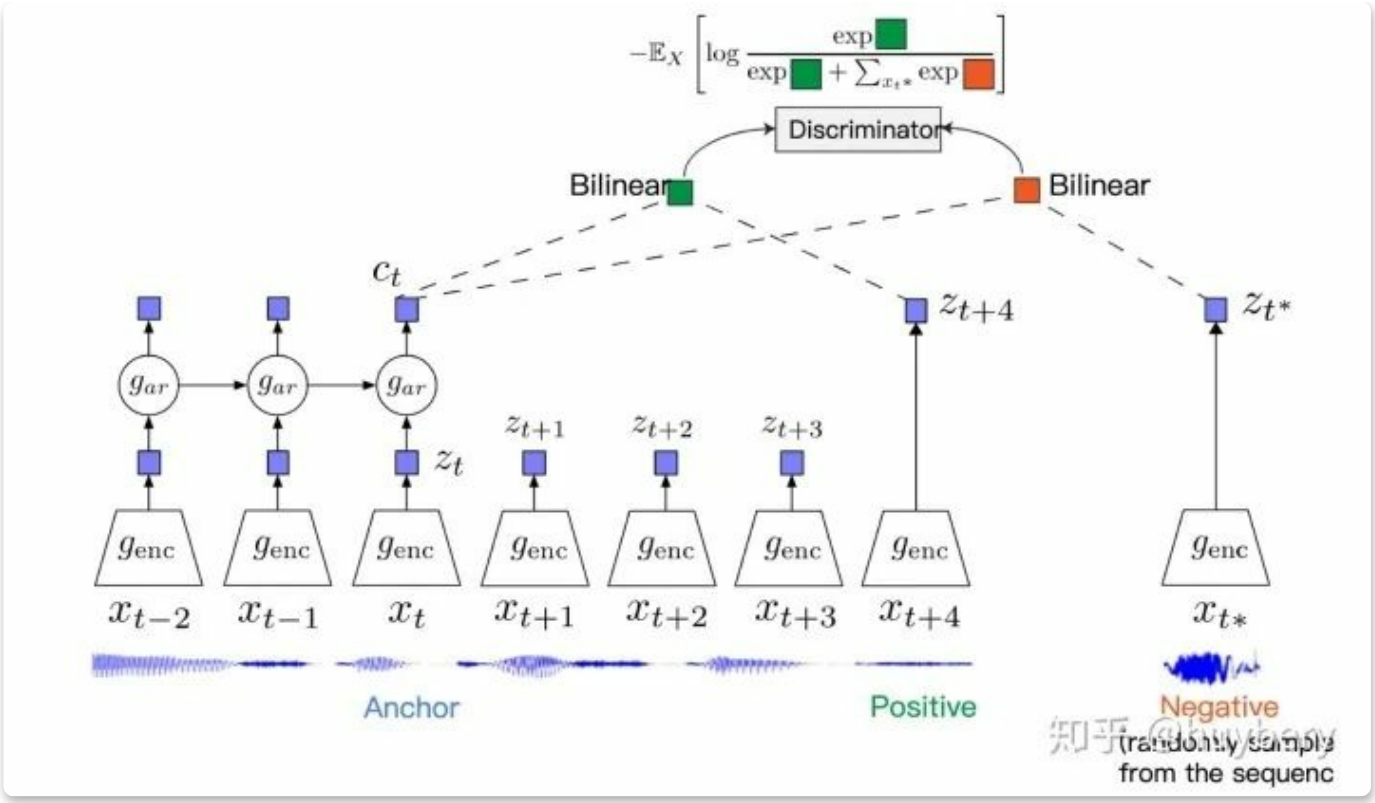
这里的 x 通常也称为「anchor」数据，为了优化 anchor 数据和其正负样本的关系，我们可以使用点积的方式构造距离函数，然后构造一个 softmax 分类器，以正确分类正样本和负样本。这应该鼓励相似性度量函数（点积）将较大的值分配给正例，将较小的值分配给负例：

通常这个损失也被称为 InfoNCE （多么炫酷的名字啊），后面的所有工作也基本是围绕这个损失进行的。



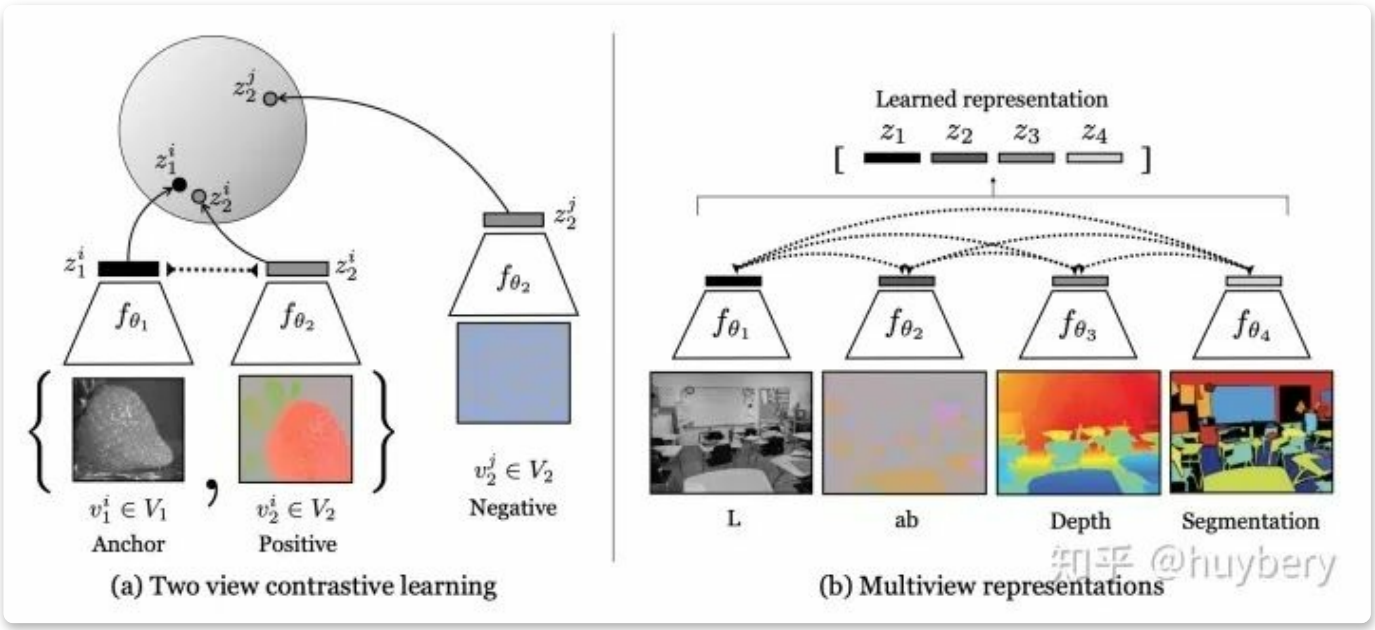
DIM

我们首先介绍 ICLR 2019 的 DIM [23]，DIM 的具体思想是对于隐层的表达，我们可以拥有全局的特征（编码器最终的输出）和局部特征（编码器中间层的特征），模型需要分类全局特征和局部特征是否来自同一图像。所以这里 x 是来自一幅图像的全局特征，正样本是该图像的局部特征，而负样本是其他图像的局部特征。这个工作的开创性很强，已经被应用到了其他领域，比如 graph [24]。



CPC

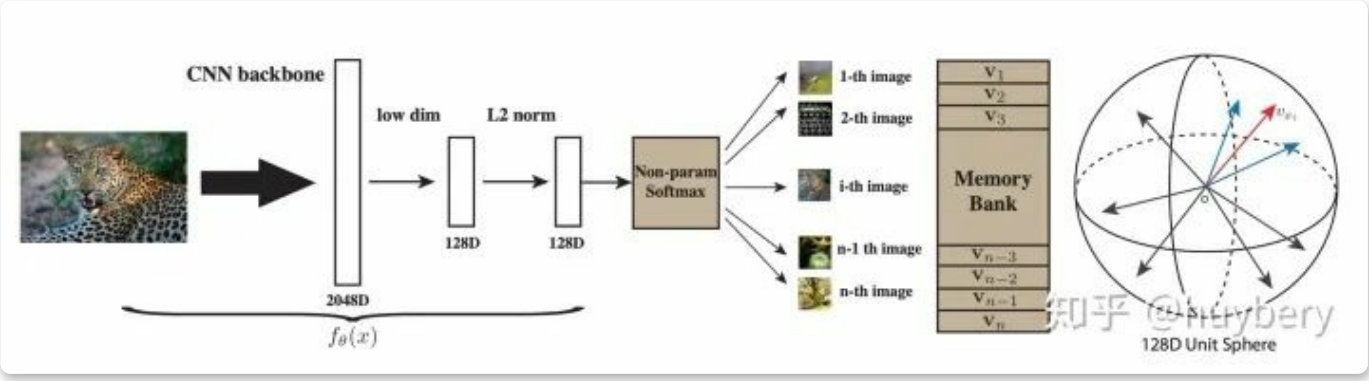
CPC 同样是一个基于对比约束的自监督框架，主要是可以应用于能够以有序序列表示的任何形式的数据:文本、语音、视频、甚至图像（图像可以被视为像素或块的序列，后面作者也给出了具体的想法）。CPC 主要是利用自回归的想法，对相隔多个时间步长的数据点之间共享的信息进行编码来学习表示，这个表示 c_t 可以代表融合了过去的信息，而正样本就是这段序列 t 时刻后的输入，负样本是从其他序列中随机采样出的样本。CPC的主要思想就是基于过去的信息预测的未来数据，通过采样的方式进行训练。



CMC

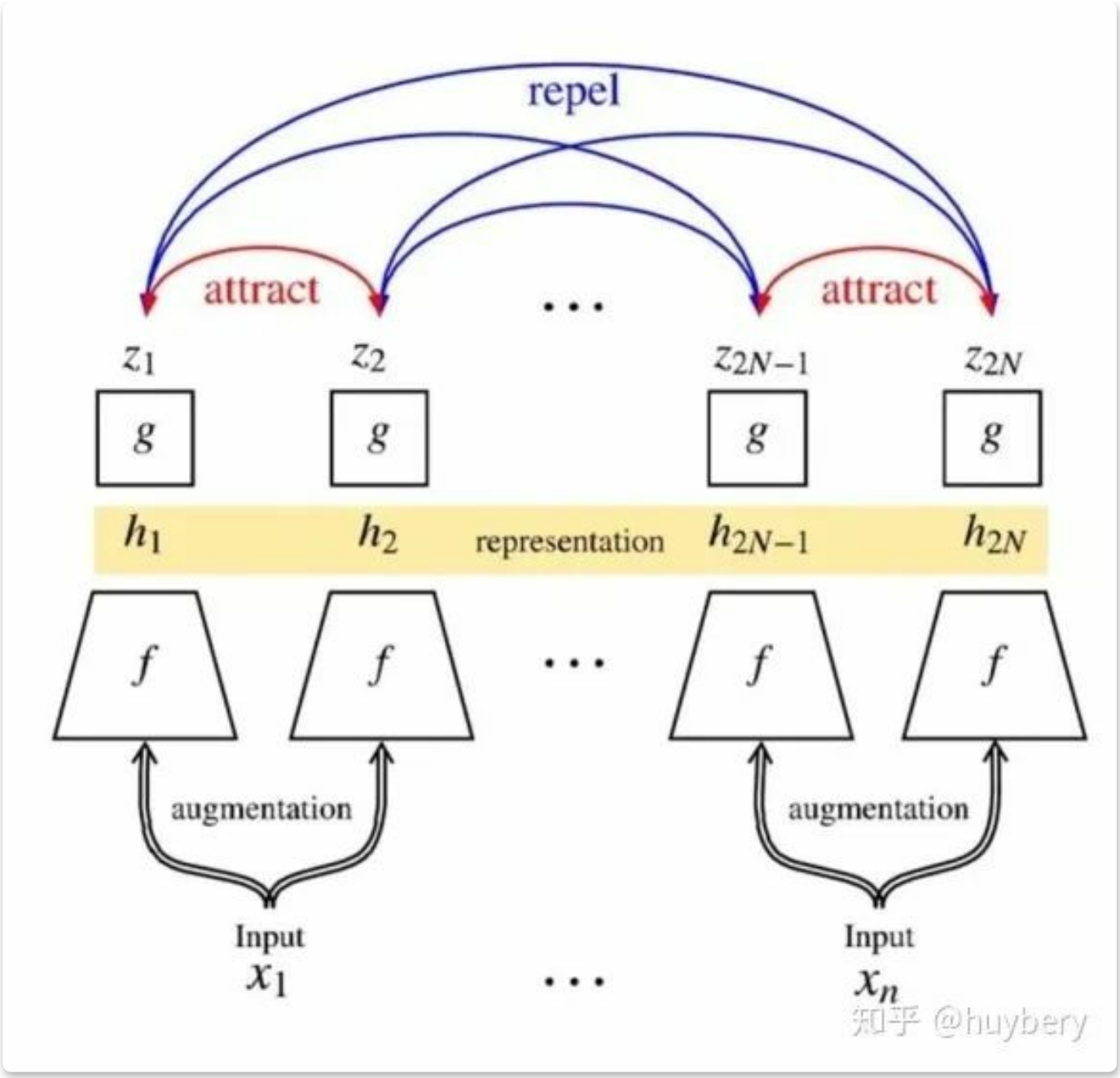
所以基于对比约束的自监督方法主要围绕如何选取正负样本， @慕容腹黑 大佬提出了利用多模态（多视角）的信息来构造样本[26] ，一个样本的多个模态为正样本，其他样本的模态为负样本。我认为这个工作还是很有启发性的，很遗憾 ICCV2019 没有

中，真心希望这篇文章能够有一个好的归宿。



Memory Bank

对于具体的实现上，因为存在大量的样本，如何存取和高效的计算损失是急需解决的。研究人员提出了memory bank [27]的概念，也就是说我们把之前模型产生样本特征全部存起来，当前计算损失的时候直接拿来用就可以了，每次模型更新完后将当前的特征重新更新到 memory bank 中，以便下一次使用。这个工作的缺点就在于每次需要将所有样本的特征全部存起来。后续 kaiming 大神提出的 Moco[28]，主要的贡献是 Momentum Update、shuffleBN 等技术点来优化这个过程。关于 Moco 知乎上已经有了很多的解释了，推荐大家阅读 [2]，这里我们就不展开介绍了。



SImCLR

最近 hinton 组又放出了 SimCLR[29]，这个工作主要是对于一个输入的样本，进行不同的数据增广方式，对于同一个样本的不同增广是正样本，对于不同样本的增广是负样本。整个过程比之前kaiming提出的动量对比（MoCo）更加的简单，同时省去了数据存储队列。这个工作的创新主要有两个：

1. **在表征层和最后的损失层增加了一个非线性映射可以增加性能 （这个地方我比较好奇，希望能有大佬给出更直观的解释）。**
2. **数据增广对于自监督学习是有益的，不同数据增广方式的结合比单一增广更好。**

   同时作者公布了非常多的实验经验，比如自监督学习需要更大的 batch 和更长的训练时间。

Discussion

通过阅读这些经典工作，我自己的思考主要如下：

- 找到合适的辅助任务（pretext）对于自监督学习是最需要解决的问题。
- 数据和资源越多，自监督预训练的效果会更好（Bert, MoCo, SimCLR）。
- 自监督直接和具体任务的结合（Task Related Self-Supervised Learning）是个可探索的方向，已经在很多任务中初露头角，也比较符合审稿人的口味。

可能喜欢

- 斯坦福大学最甜网剧：知识图谱CS520面向大众开放啦！
- Google|突破瓶颈，打造更强大的Transformer
- ACL2020|对话数据集Mutual：论对话逻辑，BERT还差的很远
- ACL2020|FastBERT：放飞BERT的推理速度
- LayerNorm是Transformer的最优解吗?

参考文献

[1] https://lawtomated.com/supervised-vs-unsupervised-learning-which-is-better/

[2] https://zhuanlan.zhihu.com/p/102573476

[3] https://zhuanlan.zhihu.com/p/107126866

[4] https://zhuanlan.zhihu.com/p/30265894

[5] https://zhuanlan.zhihu.com/p/108625273

[6] https://lilianweng.github.io/lil-log/2018/08/12/from-autoencoder-to-beta-vae.html

[7] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised Visual Representation Learning by Context Prediction. In ICCV 2015

[8] Noroozi, M., & Favaro, P. Unsupervised learning of visual representations by solving jigsaw puzzles. In ECCV 2016.

[9] Deepak Pathak et al. Context Encoders: Feature Learning by Inpainting. In *CVPR 2016*.

[10] Devlin, Jacob et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” NAACL-HLT (2019).

[11] Zhang, R., Isola, P., & Efros, A. A. Colorful image colorization. In ECCV 2016.

[12] Zhang, R., Isola, P., & Efros, A. A. Split-Brain Autoencoders: Unsupervised Learning by Cross-Channel Prediction. In CVPR 2017

[13] Gidaris, Spyros et al. “Unsupervised Representation Learning by Predicting Image Rotations.” In ICLR 2018

[14] Lee, Hankook et al. “Rethinking Data Augmentation: Self-Supervision and Self-Distillation.” ArXiv abs/1910.05872 (2019): n. pag.

[15] Gidaris, Spyros et al. “Boosting Few-Shot Visual Learning with Self-Supervision.” ICCV 2019

[16] Zhai, Xiaohua et al. “SL: Self-Supervised Semi-Supervised Learning.” ” ICCV 2019

[17] Sermanet, Pierre et al. “Time-Contrastive Networks: Self-Supervised Learning from Video.” 2018 IEEE International Conference on Robotics and Automation (ICRA) (2017): 1134-1141.

[18] Wang, Xiaolong and Abhinav Gupta. “Unsupervised Learning of Visual Representations Using Videos.” *2015 IEEE International Conference on Computer Vision (ICCV)* (2015): 2794-2802.

[19] Misra, I., Zitnick, C. L., & Hebert, M. Shuffle and learn: unsupervised learning using temporal order verification. In ECCV 2016.

[20] Wu, Jiawei et al. “Self-Supervised Dialogue Learning.” ACL (2019).

[21] https://cloud.tencent.com/developer/article/1389555

[22] https://ankeshanand.com/blog/2020/01/26/contrative-self-supervised-learning.html

[23] Hjelm, R. Devon et al. “Learning deep representations by mutual information estimation and maximization.” . ICLR 2019

[24] Velickovic, Petar et al. “Deep Graph Infomax.” *ArXiv* abs/1809.10341 (2018): n. pag.

[25] Oord, Aäron van den et al. “Representation Learning with Contrastive Predictive Coding.” ArXiv abs/1807.03748 (2018): n. pag.

[26] Tian, Yonglong et al. “Contrastive Multiview Coding.” ArXiv abs/1906.05849 (2019): n. pag.

[27] Wu, Zhirong et al. “Unsupervised Feature Learning via Non-parametric Instance Discrimination.” CVPR 2018

[28] He, Kaiming et al. “Momentum Contrast for Unsupervised Visual Representation Learning.” ArXiv abs/1911.05722 (2019): n. pag.

[29] Chen, Ting et al. “A Simple Framework for Contrastive Learning of Visual Representations.” ArXiv abs/2002.05709 (2020): n. pag.

文章已于修改
