每周研究进展阶段汇报

汇报人：杨凯冰
电　邮：tjuykb3022234232@163.com
时间段：2025 年 1 月 18 日 (周六) 至 2025 年 1 月 24 日 (周五)

一、本周工作：
1. 学习 pytorch 的基本用法，并根据 LoRA 代码进行实操理解；
2. 阅读冬暖学长论文列表中大模型综述中'Visual Understanding' 部分，并进行论文迭代，对该部分有了初步的了解
3. 和文轩学长讨论数据标注流程，并亲身进行数据集的标注。

二、思考总结：

Part 1.

将大一时懵懂安装的 anaconda、conda、cuDNN、pytorch 成功卸载并安装到合适版本，成功通过 pytorch 调用 GPU 并根据小土堆视频成功读取图片信息，实现基本的数据集建立和加载。

Part 2.

本周论文阅读主要以建立对多模态基本领域认识为主，以多模型综述中 Visual Understanding 为基本，通过迭代搜索了解最近多模态研究动态。主要包括 Label supervision, Language supervison 和 image-only self-supervision 的 general-purpose vision backbone.

Supervised Pre-training 作为一种特殊的 Label supervision，旨在将图像映射到与视觉相关联的离散标签，可用于图像分类等视觉问题。目前主流的 backbone 有传统的 AlexNet, ResNet [1] 和最近流行的 Vision Transformer 和 Swin Transformer。而标签映射和特征学习的好坏则会受到数据集的规模、类别和噪声影响，所以为了获得一个效果较好的监督预训练，需要考虑大规模的数据集（ImageNet-1K [2], ImageNet-21K [3] 等）、较好的数据质量（需要考虑噪声干扰）和模型的设计（不同的 loss function 如 margin loss [4] 等）。

Contrastive Language-Image Supervision(CLIP) [5] 则是在 Vison 基础上，加入语言并组合成 image-text pair, 将其作为正样本，不匹配的 pair 作为负样本，通过最大化正样本相似性，最小化负样本相似性来训练模型学习图像-文本对语义的方法，该方法能够有效的在同一语义空间中，将相关的文本和语义映射在尽可能近的位置，不相关的映射在尽可能远的位置。通过这种对比式学习，模型能够有优异的 zero-shot 能力，其能力主要由 batch size, data size 和 model size 决定。而因为 CLIP 的显著性能，有许多研究从上面三方面来进一步提高该模型。

- 在数据集方面，OpenAI、ALIGN 等分别通过在 400M、8B 文本图像对上训练，提高数据集规模的方式提升性能。但受限于数据集的不公开和计算资源，最近研究方向转化为从较小的数据集中提出更多有效性的方式，比如 interleaved image-text datasets(MMC4 [6] 等）；

- 在模型设计方面，则更为有意思。由于 CLIP 主要由 image encoder and language encoder 组成，故最近研究主要以优化两个编码器为主。在视觉编码上，FLIP [7] 提出了一个通过随机制造高遮蔽率的视觉补丁，只让能够看见的部分（未被遮蔽补丁遮挡的部分）进入视觉编码，这种方法能够在不降低 CLIP 性能的基础上提高其效率和鲁棒性。这方法十分的巧妙和神奇，展示了 AI 的魅力。它通过高遮蔽率（masking ratio），不仅能够提高 batch size 加速训练、降低显存，还能巧妙地起到了了正则化作用!
  在语言编码上，则尝试使用更多的外部数据如数据重写等；在可解释性上，STAIR [8] 则通过构建高维稀疏空间（large dictionary），利用子词 (subword) 为每个维度赋予对应权重，能够看到文本和图像在该空间中是如何关联起来的，使得文本图像在该语义空间具有更高的的可解释性，且该研究证明能够在提高 CLIP 的性能的基础上提高解释性；在更多模态融合上，ImageBind [9] 则将 pre-trained CLIP 冻结，并让其他模块以向其对其为目标进行学习，以此提高模型适应于更多模态的性能。

- 在目标函数上，也有许多新的尝试，如 FILIP [10] 没有使用传统的点积计算图像文本对的相似，而是使用 word-patch alignment 来进行更加精细对比；CoCa [11] 使用生成式损失，VirTex [12] 仅使用 Captioning loss, 尽管仅使用这种损失函数性能没有传统的 CLIP 好，但其 scaling 能力会更强；还有使用 Sigmoid [13] 函数进行学习，能够在零样本预测方面达到较好的性能。

CLIP 的诞生，提出了全新的预训练范式，其中对比学习的方法值得深度专研，同时 FILIP 的 mask 机制也十分巧妙，后续的论文阅读准备进一步深入了解。在对 CLIP 的初步了解的过程中，我也深刻感受到了一项优

秀技术的提出与创新，能够对该领域起到突破性的贡献，也能吸引众多研究者在此 baseline 进行进一步研究，产出更加优秀的模型，让我感受到了科研的魅力与神奇!

Image-Only Self-Supervised Learning 仅使用图像本身进行自监督学习，要求对所给定的图像进行多种变化（数据增强）如分割旋转等，学习其中的相似性和差异性，从而理解不同图像在不同特征不同表现之下的表征含义，捕捉图像的图意、语义、视觉模式等。该综述主要从 contrastive learning、non-contrastive learning 和 masked image learning 三方面进行介绍。

- Contrasitve learning 是上文 CLIP 的主要思维，通过学习不同图片（一般是对既有图片进行变化）之间的相似性和差异性，能够更好的捕捉图片和语义的联系，增强模型鲁棒性。核心思想是使用正例子对和重用负例子对。最新的研究主要是采用' 对一张照片使用两种数据增强手段-通过在编码器后使用投影头 (projector header) 执行对比学习来判断是否为同一张图片-在下游执行任务' 的模型。这要求较多的负样本对。
- Non-contrastive learning 是将图片进行变化重建，让模型恢复原状的学习方法，使用非对称结构代替负样本对。比较出众的结构有 SimSiam [14],DINO [15] 等结构。
- Masked Image Modeling(MIM) 将图片的部分 patch 掩盖，让模型通过学习部分能够看见的 patch 来预测被 masked 的 patch，从而挖掘图像文本之间的语义关系，主要有两部分构成：tokenizer 用于将图片转化为离散的视觉 tokens 用于训练，tokenizer 也被用于自回归图像生成领域；使用被遮盖后的数据进行预测。MIM 因其优秀性能而被广泛应用，主要可以分为两部分：Low-level pixels/features as targets, High-level pixels/features as targets，需要注意的则是其 scaling properties 和可能无法学到全局图像特征。

同时，CLIP 还能和众多方法结合，完成更加复杂的任务，做到更优的效果提高：

- CLIP with label supervision 将 noisy labels 和 text supervision 结合用于视觉预训练。
- CLIP with image-only (non-)contrastive learning
- CLIP with MIM

在仔细阅读 Visual Understanding 后，对于计算机如何进行视觉理解感动十分新奇，通过一系列十分有趣、巧妙的方法，将数据进行处理，使得机器能够理解并进行处理，达到 general-purpose understanding 的效果，也深深感受到了自己相关知识的匮乏。

Part 3.

和文轩师兄讨论数据集标注流程并实际标注八个数据集。感受到了一个高质量的数据集构建的辛苦，数据集标注流程确定的规范性、人工标注的艰难性。同时感受到了机器初标的效能，通过 gemini、Dino 的标注，已经能够达到较为高效、准确的结果。

三、下周规划：
下周为春节周，在继续学习基础知识的基础上，跟随文轩师兄项目学习。
1. 继续学习 pytorch 的基本用法，如有时间完成李宏毅老师 ML 课程的 pytorch 左右；
2. 阅读冬暖学长论文列表中大模型综述中'Visual Generation' 部分，并进行论文迭代。
3. 和文轩师兄讨论数据集标准流程，配合相关工作；
4. 协助文轩师兄长尾实验。

References

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015. [Online]. Available: https://arxiv.org/abs/1512.03385

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., vol. 25. Curran Associates, Inc., 2012. [Online]. Available: https://proceedings.neurips.cc/paper$_f$*iles*/*paper*/2012/*file*/*c*399862*d*3*b*9*d*6*b*76*c*8436*e*924*a*68*c*45*b* − *Paper.pdf*

[3] T. Ridnik, E. Ben-Baruch, A. Noy, and L. Zelnik-Manor, "Imagenet-21k pretraining for the masses," 2021. [Online]. Available: https://arxiv.org/abs/2104.10972

[4] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," 2018. [Online]. Available: https://arxiv.org/abs/1801.09414

[5] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021. [Online]. Available: https://arxiv.org/abs/2103.00020

[6] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models," *arXiv preprint arXiv:2304.10592*, 2023.

[7] Y. Li, H. Fan, R. Hu, C. Feichtenhofer, and K. He, "Scaling language-image pre-training via masking," 2023. [Online]. Available: https://arxiv.org/abs/2212.00794

[8] C. Chen, B. Zhang, L. Cao, J. Shen, T. Gunter, A. M. Jose, A. Toshev, J. Shlens, R. Pang, and Y. Yang, "Stair: Learning sparse text and image representation in grounded tokens," 2023. [Online]. Available: https://arxiv.org/abs/2301.13081

[9] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra, "Imagebind: One embedding space to bind them all," 2023. [Online]. Available: https://arxiv.org/abs/2305.05665

[10] L. Yao, R. Huang, L. Hou, G. Lu, M. Niu, H. Xu, X. Liang, Z. Li, X. Jiang, and C. Xu, "Filip: Fine-grained interactive language-image pre-training," 2021. [Online]. Available: https://arxiv.org/abs/2111.07783

[11] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu, "Coca: Contrastive captioners are image-text foundation models," 2022. [Online]. Available: https://arxiv.org/abs/2205.01917

[12] K. Desai and J. Johnson, "Virtex: Learning visual representations from textual annotations," 2021. [Online]. Available: https://arxiv.org/abs/2006.06666

[13] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," 2023. [Online]. Available: https://arxiv.org/abs/2303.15343

[14] X. Chen and K. He, "Exploring simple siamese representation learning," 2020. [Online]. Available: https://arxiv.org/abs/2011.10566

[15] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," 2021. [Online]. Available: https://arxiv.org/abs/2104.14294