吴侃

wkcn@live.cn | github.com/wkcn

求职方向：大模型训练（算法/框架，包含计算机视觉、自然语言处理）

教育经历

中山大学 计算机科学与技术专业（18 级直博生，与微软亚洲研究院联合培养） 2018.09 - 至今

计算机视觉方向。校内导师：朝红阳；微软导师：彭厚文、郭百宁

中山大学 计算机科学与技术专业（14 级本科） 2014.09 - 2018.06

研究成果

谷歌学术档案：https://scholar.google.com/citations?user=sK4JUL4AAAAJ

ICCV 2023 –TinyCLIP: CLIP Distillation via Affinity Mimicking and Weight Inheritance
我们提出了新的**跨模态模型蒸馏框架**，该框架包含**相似度模仿和权重继承**两个关键技术，并且将方法扩展到了**多阶段渐进蒸馏**。该蒸馏框架可以将 CLIP ViT-B/32 的**参数量减少一半**，并且维持相当的性能。和从头开始训练相比，我们的方法达到同等性能，训练速度提高了 1.4 到 7.8 倍。

ECCV 2022 - TinyViT: Fast Pretraining Distillation for Small Vision Transformers
我们研究了快速预训练蒸馏方法，将**大数据、大模型中的知识蒸馏**给小型视觉 Transformer 模型，我们的**21M 参数量**的模型在 ImageNet-1k 上达到了**84.8%**的验证集 Top-1 准确率。TinyViT-5M 用在 MobileSAM 工作上，与 ViT-L 相比模型缩小了 60 倍，速度提高 50 倍，得到了接近的性能。

CVPR 2022 - MiniViT: Compressing Vision Transformers with Weight Multiplexing
我们对视觉 Transformer 模型进行权重共享、权重复用和知识蒸馏，解决了共享权重模型在蒸馏时不稳定的问题，**减少**DeiT、Swin Transformer 的**近一半的参数量**，准确率有 1 个点以上的提高。

ICCV 2021 - Rethinking and Improving Relative Position Encoding for Vision Transformer
我们分析了相对位置编码中的关键因素，提出了四种**二维空间下的相对位置编码**和一种高效的实现方式，可以作为视觉 Transformer 的插件，提升图像分类、目标检测等任务的精度。

NeurIPS 2021 - Searching the Search Space of Vision Transformer
我们将视觉 Transformer 结构中各搜索变量解耦并分析，提出新的**自动设计神经网络的搜索空间**的方法。搜索出的模型精度超过了现有的 Transformer 模型，并在下游任务上有精度提升。

CVPR 2021 - LightTrack: Finding Lightweight Neural Networks for Object Tracking via One-Shot Architecture Search
使用神经网络结构搜索方法**搜索轻量的单目标跟踪模型结构**，在移动端的推理速度是先前人工设计的、最先进的模型 Ocean 的 12 倍，并且准确率更高。

大模型训练相关能力

大模型训练：具有**训练 GPT-6.7/13/175B、CLIP 和 MoE 模型**的经验。能够修改 Megatron-LM 和 Transformer Engine 中模型并行、分布式优化器（ZeRO 和混合精度训练方面）的代码以支持 FP8 混合精度训练，减少显存占用和通讯量，提高训练速度；能够实现读取**大规模数据的 Dataloader**；具有**MoE、ZeRO、模型并行、混合精度训练**底层代码实现经验；熟练使用并了解训练框架 (DeepSpeed, Megatron-LM, Metaseq)、并行策略 (DP/TP/PP/SP)、显存节省策略 (ZeRO, Gradient Checkpointing) 混合精度训练 (Apex, PyTorch AMP) 底层实现。

深度学习框架：熟悉 PyTorch、MXNet、Caffe 等深度学习框架及内部实现，了解分布式训练、Dataloader 的底层实现，能够阅读图优化、算子调度和内存池的实现。能够对底层代码（C++/CUDA）进行开发与调试。

实习经历

2021.07 - 至今 **微软亚洲研究院** - 多媒体搜索与挖掘组/视觉计算组全职实习生，导师：彭厚文、胡瀚

基于 Transformer 模型，进行**大模型训练（GPT 模型训练/微调、混合专家 MoE、多模态模型预训练）**以及模型**蒸馏压缩**的研究。**3 篇模型蒸馏、压缩相关论文（CVPR 2022, ECCV 2022, ICCV 2023）已被接收**。

2019.06 - 2020.03　**微软亚洲工程院** - 小冰组全职实习生，导师：傅建龙、邱凯
设计多尺度的人群计数算法；参加 CVPR2019 细粒度分类 iNaturelist 比赛和医学图像分类 RSNA 比赛。

2017.07 - 2018.06　**微软亚洲研究院** - 多媒体搜索与挖掘组全职实习生，导师：梅涛、傅建龙
进行目标检测的研究，实现了支持单卡多图输入的二阶段目标检测器；进行图像分割和自动海报排版等实验项目。

项目经历

2022.07 - 至今　**MS-AMP** [https://github.com/Azure/MS-AMP]
MS-AMP 是一个**FP8 混合精度训练库**，它在模型权重、梯度、矩阵乘法、优化器和通讯上全方面实现了 FP8 精度，能够在维持模型性能的条件下**大幅减少显存消耗，提高训练速度**，可以高效地支持**GPT 等大模型**的训练，支持 DeepSpeed 和 Megatron-LM。我负责项目的**算法设计和原型开发**。

2018.03 - 至今　**Apache MXNet** [https://github.com/apache/mxnet]
我是开源深度学习框架 Apache MXNet 的**项目管理委员会（PMC）成员**，以及 DMLC（分布式（深度）机器学习社区）的成员。我提交的**30 多个 Pull Request**已经合并到 MXNet 的主分支中，包括 CPU/GPU 上的算子实现、Bug 修复、DLPack 接口、自定义算子并行。此外，我**审查了 260 个 PR**。

2021.07 - 至今　**Cream** [https://github.com/microsoft/cream]
开源并维护了研究成果：快速预训练蒸馏框架 TinyViT、视觉 Transformer 模型压缩 MiniViT 和二维图像相对位置编码 iRPE 等工作。

2018.03 - 至今　多个著名开源社区项目
我参加了 GitHub 上多个著名开源**深度学习训练/推理框架**项目，包括：Apache/MXNet (32 commits, 3276 lines), Tencent/ncnn (3 commits, 639 lines), OAID/Tengine (2 commits, 1744 lines), Oneflow-Inc/OneFlow (2 PR), MegEngine/MegEngine (2 PR).

个人开源项目：
2018.05 - 至今　**MobulaOP** [https://github.com/wkcn/MobulaOP]，43 stars
MobulaOP 是一个简单且灵活的**跨框架算子创建工具包**。不需要重新编译深度学习框架的源码，就可以创建自定义的 C++ 算子。而且只需要一份 C++ 代码实现和简单的定义，自定义算子就可以在 CPU 和 GPU 上运行。MobulaOP 部署了自动单元测试。我负责整个项目的编写、测试和部署。

2017.05 - 至今　**Mobula** [https://github.com/wkcn/mobula]，164 stars
Mobula 是一个**轻量级、灵活的深度学习框架**，并且是纯 Python + NumPy 实现中高效的深度学习框架。该项目拥有大部分常用算子，**单元测试覆盖率高达 96%**。我负责项目编写、测试和部署。

模型部署和编程经验

模型部署：具有使用 ONNXRuntime、DeepStream 和 TensorRT 部署多模态、多目标跟踪模型的项目经验。能够使用 NCNN、Tengine Lite 和 TNN 部署模型。
编程经验：十五年以上编程经验，在编程语言：C++/C，Python，CUDA 和汇编上有大量使用经验。能够使用 Linux 和 Windows 编程环境；能进行 OpenMP, OpenMPI, CUDA 编程，能实现分布式算法如 MapReduce；对数值计算工具 NumPy, 计算机视觉库 OpenCV, 图形程序接口 OpenGL 有大量使用经验；有能力在源码或汇编级别上进行调试；能够从头实现深度学习图像分类、目标检测、目标计数、图像分割、视频分类、关系检测等模型；能够在四旋翼飞行器上进行应用级别的编程。

获奖经历

2021-2022 学年中山大学腾讯奖学金特等奖
2018-2022 年中山大学计算机学院奖学金

KAN WU

wkcn@live.cn | github.com/wkcn

Job Objective: Large-scale Model Training (Algorithm/Framework, including CV and NLP).

EDUCATION

Sun Yat-sen University, Computer Science (Ph.D. candidate) Sep 2018 - Present
SYSU-MSRA Joint Ph.D. program, Computer Vision and Deep Learning.
Academic Advisor: Hongyang Chao; Microsoft Mentor: Houwen Peng, Advisor: Baining Guo

Sun Yat-sen University, Computer Science (B.S.) Sep 2014 - Jun 2018

RESEARCH ACHIEVEMENTS

Google Scholar Profile: https://scholar.google.com/citations?user=sK4JUL4AAAAJ

ICCV 2023 –TinyCLIP: CLIP Distillation via Affinity Mimicking and Weight Inheritance
We propose a new **cross-modal model distillation framework**, involving two key techniques: **similarity mimicking and weight inheritance**. They are extended to **multi-stage progressive distillation**. It reduces the size of *CLIP ViT-B/32* by **50%** while maintaining performance. Compared to training from scratch, our method achieves comparable performance with 1.4 to 7.8 times faster training speed.

ECCV 2022 - TinyViT: Fast Pretraining Distillation for Small Vision Transformers
We study fast pretraining distillation methods which transfer knowledge from **large data and large models** to small vision transformer models. Our model with **21M parameters** achieves **84.8%** top-1 accuracy on ImageNet-1k validation set. TinyViT-5M is applied to MobileSAM, which reduces the model size by 60 times and improves speed by 50 times, compared to ViT-L with comparable performance.

CVPR 2022 - MiniViT: Compressing Vision Transformers with Weight Multiplexing
We compress vision transformer models by weight sharing, weight transformation, and knowledge distillation, addressing the instability training issue of shared-weight models during distillation. The proposed method reduces the parameter size of DeiT and Swin Transformer by **nearly 50%** with an accuracy increase of over 1 point.

ICCV 2021 - Rethinking and Improving Relative Position Encoding for Vision Transformer
We analyze the key factors in relative position encoding and propose four **2D spatial relative position encoding** methods and an efficient implementation. The proposed encoding can serve as a plugin for vision transformers, improving the accuracy of vision tasks such as image classification and object detection.

NeurIPS 2021 - Searching the Search Space of Vision Transformer
We decouple and analyze various search dimensions in vision transformer, then we propose a novel method for **searching the search space**. The discovered models surpass existing models in performance in image classification and downstream tasks.

CVPR 2021 - LightTrack: Finding Lightweight Neural Networks for Object Tracking via One-Shot Architecture Search
We design an object tracking search space, and search **a light single-object tracking model**, whose inference speed is **12× faster** than the previous state-of-the-art light model Ocean.

ABILITY FOR LARGE MODEL TRAINING

Large Model Training: I have the experience in **training GPT-6.7/13/175B, CLIP, and MoE models**. I can modify Megatron-LM and Transformer Engine to support **FP8 mixed precision training** on model parallelism and distributed optimizer (ZeRO and mixed precision training), reducing memory footprint and communication cost, and improving training efficiency. I can design the **dataloader for large-scale data**, and I have the experience in implementing **MoE, ZeRO, model parallelism, and mixed precision training**.

I understand training frameworks (DeepSpeed, Megatron-LM, Metaseq), parallel strategies (DP/TP/PP/SP), memory-saving strategies (ZeRO, Gradient Checkpointing), and mixed precision training (Apex, PyTorch AMP).

Deep Learning Frameworks: I am familiar with deep learning frameworks like PyTorch, MXNet, Caffe, etc., and their low-level implementations. I understand distributed training and dataloader's implementations, and I am able to read the implementations of graph optimization, operator scheduling, and memory pool. I can develop and debug the libraries in C++/CUDA.

INTERNSHIP EXPERIENCE

Jul 2021 - Present **Microsoft Research Asia** - Full-time Intern

in Multimedia Search and Mining Group / Visual Computing Group, Mentors: Houwen Peng, Han Hu

Research on **large model training (GPT model training/fine-tuning, mixed of expert (MoE), multimodal model pretraining)** as well as **model distillation and compression**. **3 papers about model distillation and compression (CVPR 2022, ECCV 2022, ICCV 2023) have been accepted.**

Jun 2019 - Mar 2020 **Microsoft Search Technology Center Asia** - Full-time Intern

in XiaoIce Group, Mentors: Jianlong Fu, Kai Qiu

Designed a multi-scale crowd counting algorithm; participated in the competitions of CVPR2019 iNaturelist Fine-Grained Classification and RSNA medical image classification.

Jul 2017 - Jun 2018 **Microsoft Research Asia** - Full-time Intern

in Multimedia Search and Mining Group, Mentors: Tao Mei, Jianlong Fu

Research on object detection and implemented a single-GPU multi-image two-stage object detector; participated in projects about image segmentation and automatic design for poster layout.

PROJECT EXPERIENCE

Jul 2022 - Present **MS-AMP** [https://github.com/Azure/MS-AMP]

MS-AMP is a **FP8 mixed precision training library**, which fully applies FP8 format on model weights, gradients, matrix multiplication, optimizer states, and communication. It significantly reduces memory footprint and boosts training speed while maintaining accuracy. It supports **training of large models like GPT** as well as DeepSpeed and Megatron-LM. I am responsible for **algorithm design and prototype development**.

Mar 2018 - Present **Apache MXNet** [https://github.com/apache/mxnet]

I am a member of the **Project Management Committee (PMC)** for the deep learning framework Apache MXNet, as well as a member of the Distributed (Deep) Machine Learning Community (DMLC). I have submitted **over 30 Pull Requests** that have been merged into MXNet, including CPU/GPU operator implementations, bug fixes, DLPack interface, and custom operator parallelization. Besides, I have **reviewed 260 PRs**.

Jul 2021 - Present **Cream** [https://github.com/microsoft/cream]

Maintaining the open-sourced codes of our research works: fast pretraining distillation framework (TinyViT), model compression for vision transformer (MiniViT), and 2D image relative position encoding (iRPE).

Mar 2018 - Present Multiple well-known open-source projects

I have participated in well-known open-source **deep learning training/inference framework** projects on GitHub, including: Apache/MXNet (32 commits, 3276 lines), Tencent/ncnn (3 commits, 639 lines), OAID/Tengine (2 commits, 1744 lines), Oneflow-Inc/OneFlow (2 PR), MegEngine/MegEngine (2 PR).

Personal Open-source Projects:

May 2018 - Present **MobulaOP** [https://github.com/wkcn/MobulaOP], 43 stars

MobulaOP is a simple and flexible **cross-framework custom operator toolkit**. It allows creating custom C++ operators without recompiling the source code of deep learning framework and can run custom operators on both

CPU and GPU with just one C++ implementation and simple definitions. Unit testing has been deployed. I am responsible for the **development, testing, and deployment**.

May 2017 - Present **Mobula** [https://github.com/wkcn/mobula], 164 stars

Mobula is a **lightweight and flexible deep learning framework** with high efficiency in Python + NumPy implementation. The project includes most commonly used operators, with **unit test coverage of up to 96%**. I am responsible for the development, testing, and deployment.

EXPERIENCE OF MODEL DEPLOYMENT AND PROGRAMMING

Model Deployment: Experienced in deploying multi-modal and multi-object tracking models using ONNXRuntime, DeepStream, and TensorRT. Capable of deploying models using NCNN, Tengine Lite, and TNN.

Programming Experience: Over 15 years of programming experience in C++/C, Python, CUDA, and assembly languages. Proficient in programming on Linux and Windows; capable of OpenMP, OpenMPI, CUDA programming; able to implement distributed algorithms like MapReduce; familiar with and have extensive experience using numerical computing libraries such as NumPy, computer vision library OpenCV, graphics library OpenGL; capable of debugging at source code or assembly level; able to implement deep learning models for image classification, object detection, object counting, image segmentation, video classification, relationship detection, etc., from scratch; capable of application-level programming on quadcopter.

AWARDS

Academic Year 2021-2022 Tencent Scholarship (Special Prize) at Sun Yat-sen University

Academic Year 2018-2022 Scholarships at School of Computer Science and Engineering, Sun Yat-sen University