多任务的应用
综述

王祥通
2018226215015

Abstract

首先介绍多任务的定义，必要性，以及遇到的挑战。然后，本文从理论上回顾现有工作的理论研究，包括对多任务以软硬共享方法进行的分类方法。最后，本文以视觉任务下多任务学习进行分类，主要包括场景感知的密集估计型和场景监测的非密集估计型两种。前者在多任务研究领域有较多的参考，后者由于任务的抽象程度更高，工作较少，也是本文的重点。

1. 多任务学习

多任务学习(Multi-Task Learning) 也称为联合学习(Joint Learning), 交叉学习（cross-task）是一种归纳传输方法，通过将相关任务的训练信号中包含的域信息用作归纳偏差来改进泛化。它通过在使用共享表示的同时并行学习任务来实现这一点;每项任务的学习内容可以帮助更好地学习其他任务[1]。

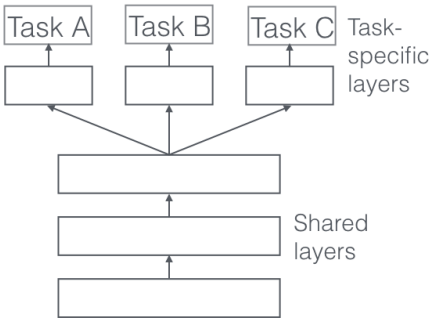
多任务学习通过的必要性有很多，例如： i) 推断时，计算资源有限设备中，通过共享参数能减少推断时间 ii) 训练时，多任务共同学习相比依次训练更节约时间。 iii) 评估时，通过多任务学习理论上能提升模型的评估精度，以及泛化能力。 iv) 对数据的利用更为高效。然而，多任务学习也面临着许多挑战，包括但不限于 i)多任务协同训练下，不同的任务需要有不同的学习率或者学习计划(learning schedule). ii)优化过程中某一个任务的变优也许会使得其他任务表现变差。iii)任务梯度可能会产生干扰，多个总和损失可能会使优化环境更加困难。

多任务关系（task relationship）让人自然而然的考虑到迁移关系（transfer relationship），但两者之间并没有很明显的相关性[2].

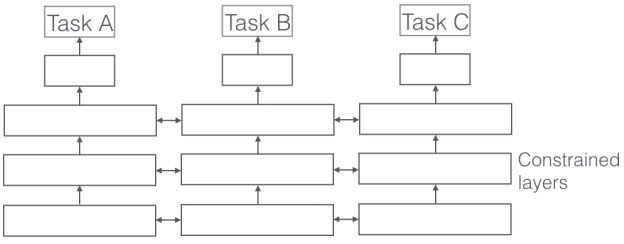
1.1. 基于理论的多任务学习分类

在深度学习中, 以参数的共享方法分类，多任务学习可以被分成两类, 硬参数共享(hard parameter sharing)与软参数共享(soft parameter sharing)[3], 硬参数共享的多任务会通过一个相同Encoder, 然后根据各自任务使用不同的Decoder; 软参数共享没用公共的层, 但是在不同任务的架构上会有跳跃连接来实现特征传输;

在解决视觉多任务领域中, 一个比较新的工作是[4]提出的十字绣网络, 该文旨在解决检测与分割任务. 文章重点在于讨论模型框架, 对两个任务并没过多介绍. 作者在两个相同结构的卷积网络的相同层之间添加十字绣单元(Cross-stitch Unit)来选择性的进行不同网络相同层之间的特征图传输.



(a) 深度网络中的硬参数共享



(b) 深度网络中的软参数共享

图 1. 硬参数共享的多任务会通过一个相同Encoder, 然后根据各自任务使用不同的Decoder; 软参数共享没用公共的层, 但是在不同任务的架构上会有跳跃连接来实现特征传输.

文章在消融试验上面花了较多篇幅讨论这种架构的模型面临的一些新问题, 比如十字绣单元的少量参数有着明确的约束, 需要单独学习, 学习率的设置对整体影响以及两个网络初始化问题. 此类问题多在理论层面分析和探讨多任务学习，而应用方面的，尤其是视觉方面的探讨比较少，此类问题又是工业界更为关注的，因此我们更倾向于以视觉任务为主干讨论多任务学习。

1.2. 视觉任务下多任务学习分类

多任务学习中的任务，在不同的情况下意义不同，

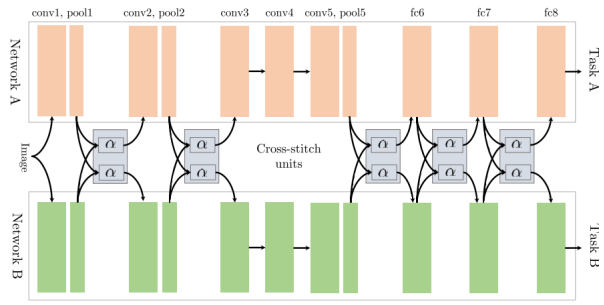


图 2. 使用 cross-stitch units 的两个AlexNet [5]. 在这种情况下，作者仅在pooling层和全连接的图层之后才用cross-stitch units，其可以将共享表示建模为输入激活图的线性组合。该网络尝试学习可以帮助完成任务A和B的表示。我们将从任务A直接监督的子网称为网络A(上), 另一个称为网络B(下)

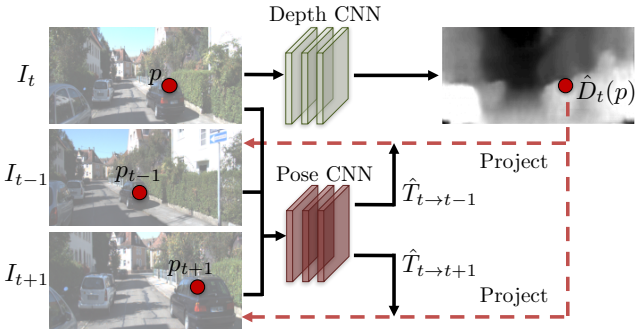


图 3. 基于深度估计任务与相机自运动位姿变换两个任务下的框图[17]

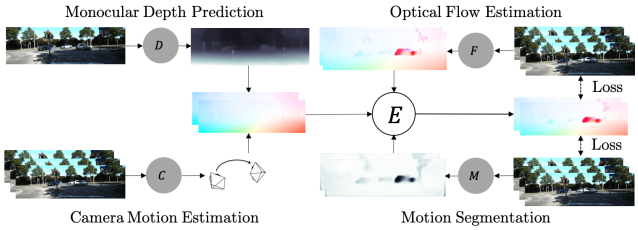


图 4. 包含了深度估计，位姿变换，分割以及光流估计四个任务的框图[8]

有时是代表不同尺度或者视觉层次的相同任务[6]，有时是代表一个视觉任务的不同阶段[7]，有时也代表不同种类的视觉任务[8]。本文根据应用，将多任务分成两个分支，分别是以场景感知为主，模型多数为密集估计（Dense Prediction）的多任务集合，以及以场景监测为主，模型多数为稀疏估计的多任务集合。

1.2.1 场景感知类任务集

在场景感知类任务集中，重点处理对象为场景的背景，即全局性感知，涉及任务包括深度估计[9, 10]，场景表面法向估计[11]，边缘检测[12]，光流估计[13, 14, 15]，语义分割[8, 16]，视觉里程计[17],自运动速度检估计[18]任务, 图.3 和图.4分别是两个任务以及四个任务下的多任务学习框架，可以看到每个子任务通过几何关系偶合在一起，并企图提升各自任务的效果。

此类任务的一个特点就是数据难以标注，人工处理困难极大等。例如, 光流任务是要估计出图像中每个像素的移动, 并通过色彩映射标注出来, 以此估计相对相机视角下每个点的运动状况,但此种任务无法直接标注. 此类任务的另一个特点就是耦合关系较强,各个任务之间的层次性较弱. 文献[19, 20]主要介绍了此类任务集的耦合关系, 并量化了两两任务之间的关联度, 并且还在[21]中基于前面的工作提出了如何鲁棒学习的策略，使得模型的泛化能力更好。

以上任务多数都是基于视觉导航类的任务需求，重点在探测场景感知环境上，相机一般是运动状态，算法一般搭载在机器人上，计算资源有限，与场景监测类任务有着较大的不同。

1.2.2 场景监测类任务集

在场景监测类任务集中, 重点处理对象是场景中的物体，即局部性感知，涉及任务包括物体检测,跟踪,运动轨迹预测,实例分割，物体位姿估计等. 该任务集中的子任务抽象程度较高, 没有明显的关系来关联, 因此相关研究较为空白. 此类任务的特点之一就是相比于场景感知类任务的数据，标注较为容易。第二个特点是此类任务的关键难点，也是本文研究的重点，即任务之间的耦合性并非通过几何耦合这类数学关系体现，而是通过逻辑关系，各个任务之间的层次性较强。因此，本文根据不同的视觉任务以及层次关系，我们将该任务集绘制如图.5

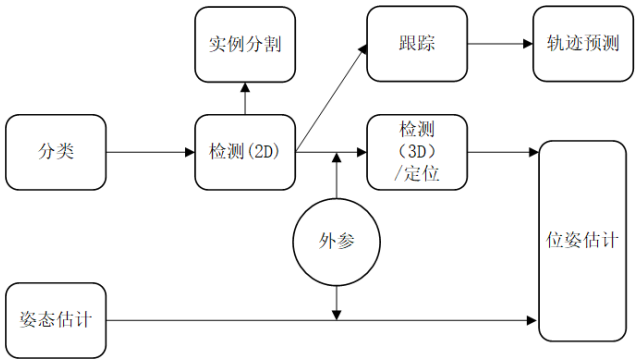


图 5. 非密集估计类任务层次关系

在场景检测类视觉任务衍生出的应用级任务中，人脸检测以精度要求高，难度较大，应用前景广阔收到更大的关注。由于需要在大量人脸中准确分别，所以一些人脸检测应用中引入了更多辅助任务以增强性能。 Zhang等人[7]提出通过堆叠式多重网络来进行该任务，对人脸分类、锚箱回归、面部特征点定位作为三个任务，并分别通过网络各自进行任务。但此种方

法的任务实质上是一个应用级视觉任务的不同阶段，子任务的结果无法直接拿来应用，与之前提及的任务有本质不同。同样是人脸识别任务，Ranjan等人[22]则零任务集中包含了人脸检测、关键点定位、人脸姿态估计以及性别识别四个子任务，效果如图.6

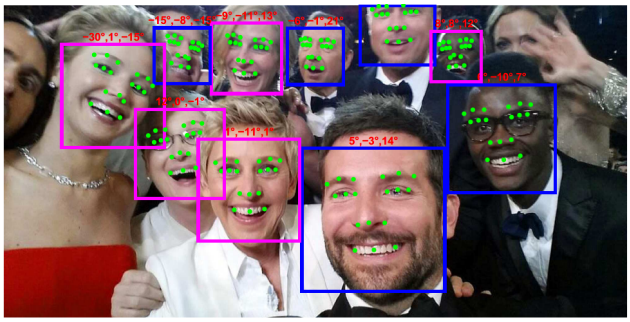


图 6. HyperFace效果展示，可见程序的结果包含人脸检测（框出）、脸部关键点定位（绿点）、性别识别（红蓝色框）以及姿态估计（俯仰偏转翻滚角度）。

参考文献

[1] Ramtin Mehdizadeh Seraj. Multi-task Learning. *Machine Learning*, 28:41–75, 1997. 1

[2] Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 2021. 1

[3] Sebastian Ruder. An Overview of Multi-Task Learning in Deep Neural Networks. (May), 2017. 1

[4] Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. Cross-Stitch Networks for Multi-task Learning. In *CVPR*, volume 2016-Decem, pages 3994–4003, 2016. 1

[5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *NIPS*, 2012. 2

[6] Iasonas Kokkinos. Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6129–6138, 2017. 2

[7] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016. 2

[8] Optical Flow and Motion Segmentation. Competitive Collaboration: Joint Unsupervised Learning of Depth, Camera Motion, Optical Flow and Motion Segmentation. In *CVPR*, 2019. 2

[9] Eigen. Depth Map Prediction from a Single Image using a Multi-Scale Deep Network. In *NIPS*, pages 1–9, 2014. 2

[10] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *ICCV*, 2015. 2

[11] Wei Yin, Yifan Liu, Chunhua Shen, and Youliang Yan. Enforcing geometric constraints of virtual normal for depth prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5684–5693, 2019. 2

[12] Zhenheng Yang, Peng Wang, Yang Wang, Wei Xu, and Ram Nevatia. Lego: Learning edge with geometry all at once by watching videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 225–234, 2018. 2

[13] Philipp Fischer, Eddy Ilg, H Philip, Caner Hazırbas, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. FlowNet: Learning Optical Flow with Convolutional Networks. In *ICCV*, 2015. 2

[14] Zhichao Yin and Jianping Shi. GeoNet: Unsupervised Learning of Dense Depth, Optical Flow and Camera Pose. In *CVPR*, 2018. 2

[15] Yuliang Zou, Zelun Luo, and Jia Bin Huang. DF-Net: Unsupervised Joint Learning of Depth and Flow Using Cross-Task Consistency. In *ECCV*, 2018. 2

[16] Marvin Klingner, Jan-Aike Termohlen, Jonas Mikolajczyk, and Tim Fingscheidt. Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance. In *European Conference on Computer Vision*, pages 582–600. Springer, 2020. 2

[17] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 2017. 2

[18] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[19] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3712–3722, 2018. 2

[20] Trevor Standley, Amir Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Which tasks should be learned together in multi-task learning. In *International Conference on Machine Learning*, pages 9120–9132, 2020. 2

[21] Amir R Zamir, Alexander Sax, Nikhil Cheerla, Rohan Suri, Zhangjie Cao, Jitendra Malik, and Leonidas J Guibas. Robust learning through cross-task consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11197–11206, 2020. 2

[22] Rajeev Ranjan, Vishal M Patel, and Rama Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE transactions on pattern analysis and machine intelligence*, 41(1):121–135, 2017. 3