华中科技大学

本科毕业设计（论文）任务书

题　　目　面向推理系统的编码计算动态优化方案研究

（任务起止日期：2023 年 6 月 1 日～2023 年 6 月 1 日）

院　　系　计算机科学与技术学院

专业班级　计算机 2103 班

姓　　名　余正浩

学　　号　U202115404

指导教师　胡燏翀

教研室（系、所）负责人 _____2023 年 6 月 1 日审查

院（系）负责人 _____2023 年 6 月 1 日批准

课题内容：

设计自适应算法来实现编码器和解码器复杂度的动态调整。算法应能够实时监测任务特点和数据分布的变化，并根据预先设定的规则自动调整编码器和解码器的结构和参数。例如，可以采用基于机器学习的算法，通过对大量任务和数据的学习，不断优化调整策略。 根据任务特点和数据分布的分析结果，制定编码器和解码器复杂度动态调整的依据和规则。例如，如果任务复杂度较高且数据分布不均匀，可能需要增加编码器和解码器的复杂度；如果任务对实时性要求较高且数据分布相对均匀，则可以适当降低编码器和解码器的复杂度。 通过大量实验验证动态调整策略的有效性。在不同类型的任务和数据分布下进行实验，对比使用动态调整策略前后推理系统的性能指标。根据实验结果，对动态调整策略进行优化调整，进一步提高推理系统的性能。 建立一套全面的性能评估指标体系，用于衡量动态调整编码器和解码器复杂度后推理系统的性能。指标可以包括准确性（如准确率、召回率等）、延迟（如平均推理时间、响应时间等）、资源利用率（如内存占用、计算资源消耗等）等方面。

课题任务要求：

实现根据不同任务特点和数据分布，自动且精准地动态调整编码器和解码器的复杂度，以达到在各种情况下推理系统性能的最优提升。具体而言，在保证准确性的前提下，最大程度地降低延迟，并有效利用计算资源。

主要参考文献（由指导教师选定）：

Asymmetric Coded Distributed Computation for Resilient Prediction Serving Systems （HPDC 24） Parity models: erasure-coded resilience for prediction serving systems. In: Proc. of ACM SOSP (2019) 12. Kosaian, J., Rashmi, K., Venkataraman, S.: Learning-based c ApproxIFER: A model-agnostic approach to resilient and robust prediction serving systems. In: Proc. of AAAI (2022)

同组设计者：

指导教师签名：

年　　月　　日