革新的コンピューティング技術の開拓
2019 年度採択研究者

2020 年度
年次報告書

孫鶴鳴

早稲田大学理工総合研究所
次席研究員（研究院講師）

リアルタイム低電力深層学習適用による革新的な動画像圧縮システム

§1. 研究成果の概要

本年度は学習型静止画像圧縮の固定小数点フレームワークを開発し、圧縮率の有効性とクロスプラットフォームコーディングの実行性を確認した。

固定小数点フレームワークの開発には主に三つの部分がある。(1)まずは重みの量子化である。重みのチャネルごと及びレイヤーごとのラプラス分布を観察し、大きい値の発生確率が低いということを確認した上で、重みの範囲を縮めた。そうすることで、同じビットでより高い量子化精度を実現した。そして、クリッピング損失を補うために、ネットワークを再訓練した。(2)重みに加えて、出力アクティベーションも量子化した。各チャネルには物理的な表現があるので、一部の DC チャネルの出力が大きくなり、AC チャネルの出力が小さくなる。この現象に基づいて、異なる出力の範囲を十分に利用するために、いくつのハードウェア向きの非線形量子化カーネルを提案した。結果として、重みと出力を両方 8 ビットに量子化する場合、32 ビットの浮動小数点と比べると、圧縮率の損失は小さくなった。

(3)そして、様々なプラットフォームで浮動小数点計算の丸め誤差があるので、エンコードとデコードのプラットフォームが異なる場合、事前確率(prior distribution)の結果も異なる。そのため、ある一つの違うプライアサンプルが現れたら、その後の全てのデコード画像に誤差が伝搬してしまう。この問題を解決するために、ルックアップテーブル(LUT)を確率モデルとして使用する。具体的に、ガウスプライア分布の場合、logフィールドでガウスの分散を線形量子化し、それぞれの量子化された分散に対して、LUT を用意する。結果として、確率モデルの精度を維持しつつ、クロスプラットフォームコーディングの実行性も確認した。

本研究結果の一部は IEEE ICIP, IEEE PCS で発表し、また IEEE Journal にも投稿した。