学 位 論 文 内 容 の 要 旨

博士の専攻分野の名称　博士（情報科学）　氏名　寺尾　颯人

学 位 論 文 題 名

Reducing Annotation and Computation Costs for Efficient Compressed Video Action Recognition
（効率的な圧縮動画分類に向けたアノテーションコストと計算コストの削減手法）

ソーシャルメディアや動画共有サイトの普及に伴い, オンライン上で利用可能な動画データの数は増加しており, それらの動画を活用するための動画解析技術の必要性は高まり続けている. 近年では深層学習が急激に進歩していることから, 深層学習を用いた動画解析技術が数多く提案されている. 特に, 入力として与えられた動画データを事前に決められた動画ラベル群に含まれる最も適したラベルに割り当てる動画分類問題に対する手法は広く研究されている.

本論文では, 動画分類の一種である圧縮動画分類に焦点を当てる. 一般的な動画分類手法では RGB フレーム列をディープネットワークの入力として扱うのに対し, 圧縮動画分類は MPEG4 などの動画圧縮アルゴリズムによって圧縮された状態の動画ファイルに保存されている,I-frame, motion vector, residual と呼ばれる複数の情報を直接ディープネットワークに入力して動画を分類する手法である. これによって RGB フレームを取得するために必要であったデコード処理を介さずにディープネットワークへの入力を得ることができ, より効率的な動画分類が可能となる. 圧縮動画分類はその効率性を活かして, エッジデバイスやモバイルデバイスのような安価で省電力なデバイス上での運用が期待されている.

本論文は圧縮動画分類の効率性を更に向上させることを目的とし, 特にアノテーションコストと計算コストを削減するための手法を提案している. 本研究ではこれらのコスト削減をそれぞれ独立した問題と捉え, それぞれ異なるアプローチで取り組んだ. 本論文は全 5 章で構成される. 第 1 章では研究背景と研究目的, および本論文の構成について, 続く第 2 章では動画分類や圧縮動画分類に関する先行研究についてそれぞれまとめる.

第 3 章では, 圧縮動画分類モデルの学習に必要なアノテーションコストを削減するための半教師あり学習手法について述べる. 従来の圧縮動画分類手法のほとんどは学習に大量の教師付きデータが必要な教師あり学習であり, 学習に用いるデータの作成にはアノテーションコストがかかってしまうという課題があった. 半教師あり学習は教師付きデータセットと教師なしデータセットの両方を用いる学習方法で, アノテーションコストの削減が可能である. まず, 研究背景として動画分類の半教師あり学習手法に加えて, コンピュータビジョン分野で近年大きく発展した画像分類の半教師あり学習手法を紹介する. 次に, 半教師あり学習手法の一つである擬似ラベル法を圧縮動画分類に適用した実験を紹介する. ここで擬似ラベル法とは, 教師なしデータを学習中のディープネットワークに 1 度入力し, 得られた予測を基に生成した擬似ラベルを通常の教師ラベルの代わりとして用いる学習法である. この実験では,I-frame, motion vector, residual それぞれに対応する 3 つのディープネットワークを別々に学習し, それらの予測を平均して評価をおこなった. 結果として, 圧縮動画の入力を用いる方が RGB フレームを入力として用いるよりも擬似ラベル法で学習した後のモデルの精度が高くなることが示される. また,RGB フレームを入力とするほかの半教師あり学習手法と比較した場合でも, 擬似ラベルと圧縮動画の入力を組み合わせた手法のほうが高い精度を達成できる

ことを示す. さらに, 半教師あり学習だけでなく教師付き学習においても, 教師データが限られた状況では圧縮動画の入力を用いる方が高い分類精度を達成できることが示される. これらの結果から, 圧縮動画分類は通常の動画分類よりも半教師あり学習が有効に働くと考えられる. この結果を受けて, 学習段階から I-frame, motion vector, residual それぞれに対応するディープネットワークの予測をアンサンブルすることでより信頼のおける擬似ラベルを生成する Compressed Video Ensemble based Pseudo Labeling (CoVEnPL) を提案する. 学習時にアンサンブルを導入することで擬似ラベルの精度を高め, 結果としてよりよいモデルを学習することができる. 更にこのアプローチと擬似ラベル法を改良した最先端の半教師あり学習手法である Fixmatch を組み合わせることで,RGB フレームのみを入力とするほとんどの手法と比較して高い精度を達成できることを示す. また, データの読み込み速度についても比較をおこない, 提案手法が複数の入力を用いているにも関わらず RGB フレームよりも高速にデータ読み込みが可能であることを示し, 提案手法の効率性についても議論する. 最後に, これらの結果を受けた考察および将来への展望について述べる.

第 4 章では, 圧縮動画分類の計算コストを削減するために提案した手法について述べる. 計算コストはモバイルデバイスやエッジデバイス上のような非力なデバイス上での運用において重要な問題である. これまで提案されてきた研究では,I-frame, motion vector, residual をそれぞれ軽量なディープネットワークで処理することで計算コストを削減する手法が多く提案されている. 一方, 本論文では圧縮動画分類の入力を処理するネットワークの数自体を減らすことで計算コストを削減するアプローチを提案し, そのようなモデルとして単一のネットワークで I-frame, motion vector, residual を同時に処理する multi-stream single network (MussNet) を提案する. 本章ではまず事前知識として, 圧縮動画分類のような複数の入力を扱うモデルについて, それぞれの入力から得られる情報をどこで混ぜるかという観点から分別した Early fusion と Late fusion を紹介する. ここで,Early fusion はディープネットワークの入力時点で情報を混ぜるモデルで,Late fusion はそれぞれの入力を別々の独立したディープネットワークで処理したのち, それらの出力を混ぜるモデルである.Early fusion は単一のネットワークですべての入力を処理するため,Late fusion よりも効率的な圧縮動画分類を実現できる一方, 学習後の分類精度が低くなってしまうという欠点がある. そこで,MussNet は単一ネットワークに対して, 内部で Late fusion がおこなわれるような学習をおこなう. 結果として,MussNet は Early fusion のような単一ネットワークによる効率性を維持しつつ,Late fusion と同等の精度を達成できることを示す. 加えて,MussNet は先行研究と比較しても高い効率性を達成しつつ, 先行研究と同等の分類精度を達成できることが示される. 最後に,MussNet の実験で得た結果をもとに考察と今後の展望について述べる.

第 5 章においては, 本論文の結論を述べる. 特に, 本論文が取り組んだ問題と提案手法について簡単にまとめ, その後本論文で得られた結果全体を受けた今後の展望について述べる.