AI 特許紹介(22)

AI 特許を学ぶ！究める！
～ニューラルネットワークを用いたビデオ分類～

2020 年 11 月 10 日

河野特許事務所
所長　弁理士　河野英仁

「AI 特許紹介」シリーズは、注目すべき AI 特許のポイントを紹介します。熾烈な競争となっている第４次産業革命下では AI 技術がキーとなり、この AI 技術・ソリューションを特許として適切に権利化しておくことが重要であることは言うまでもありません。

AI 技術は Google, Microsoft, Amazon を始めとした IT プラットフォーマ、研究機関及び大学から毎週のように新たな手法が提案されており、また AI 技術を活用した新たなソリューションも次々とリリースされています。

本稿では米国先進 IT 企業を中心に、これらの企業から出願された AI 特許に記載された AI テクノロジー・ソリューションのポイントをわかりやすく解説致します。

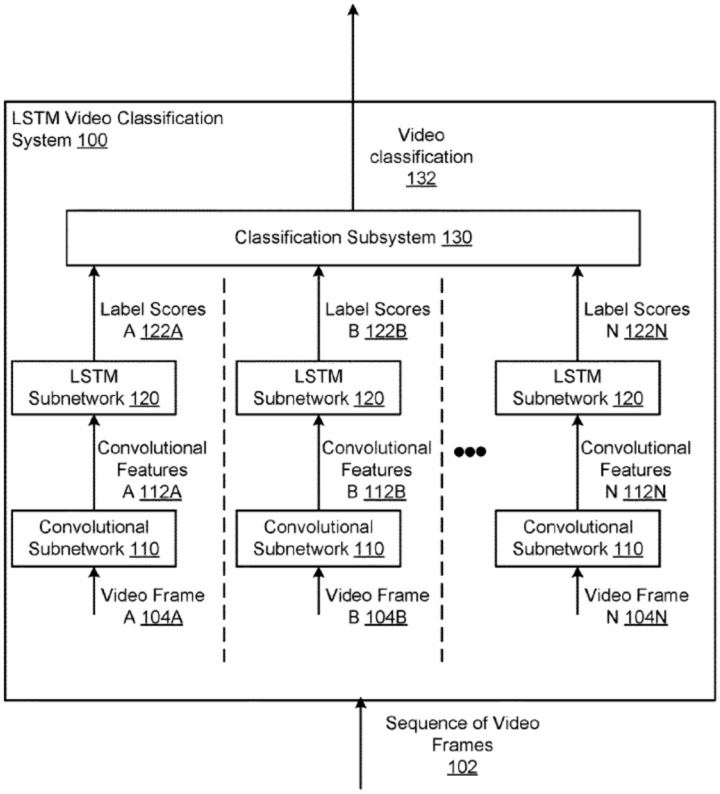
1.概要

特許権者　Google
出願日　2016 年 4 月 29 日
登録日　2019 年 5 月 14 日
登録番号 US10289912
発明の名称　ニューラルネットワークを使用したビデオの分類

912 特許は、ニューラルネットワークを用いてビデオのトピックを推定する技術に関する。

2.特許内容の説明

図 1 は、例示的な LSTM ビデオ分類システム 100 を示している。



ビデオ分類システム 100 は、入力ビデオから複数のビデオフレーム 104A～104N を含むシーケンス 102 を受信し、シーケンス内のビデオフレームを処理して、入力ビデオの分類 132 を生成する。分類 132 は、ビデオが関連するものとして分類された 1 つまたは複数のトピックを識別するデータである。ビデオ分類システム 100 は、畳み込みサブネットワーク 110、LSTM サブネットワーク 120、および分類サブシステム 130 を含む。

ビデオ分類システム 100 は、ビデオから一連のビデオフレーム 102 を受信する。シーケンス内のビデオフレーム 102 には、ビデオから抽出されたビデオフレームと、オプションで、ビデオ内の隣接するフレームから生成されたオプティカルフロー画像が含まれる。

畳み込みサブネットワーク 110 は、画像内のビデオフレームごとに、ビデオフレームを処理してビデオフレームの畳み込み特徴を生成するように構成される畳み込みニューラルネットワークである。例えば、畳み込みサブネットワーク 110 は、シーケンスからのビデオフレーム 104B を処理して、ビデオフレーム 104B の畳み込み特徴 112B を生成する。

LSTM サブネットワーク 120 は、1 つまたは複数の LSTM ニューラルネットワーク層および出力層を含み、畳み込み特徴 112Ａ～112Ｎのそれぞれを処理して、ビデオフレーム 104Ａ～104Ｎのそれぞれについてラベルスコア 122Ａ～122Ｎのそれぞれのセットを生成するように構成される。ラベルスコアの各セットには、所定のラベルセット内の各ラベルのそれぞれのスコアが含まれ、各ラベルはそれぞれのトピックを表す。LSTM ニューラルネットワーク層には、それぞれ 1 つ以上の LSTM メモリブロックが含まれている。

各 LSTM メモリブロックは、それぞれが入力ゲート、忘却ゲート、および出力ゲートを含む 1 つまたは複数のセルを含み、これにより、セルは、例えば、現在の活性化を生成する際に使用するための、または LSTM サブネットワーク 120 の他のコンポーネントに提供されるための隠れ状態として、セルによって生成された以前の活性化を記憶することができる。

シーケンス内のフレームごとに、1 つ以上の LSTM ニューラルネットワーク層がフレームの畳み込み特徴を集合的に処理して、LSTM 出力を生成する。LSTM ニューラルネットワーク層は隠れ状態を維持するため、特定のフレームの LSTM 出力は通常、フレームだけでなく、前のシーケンスのフレームに先行するフレームにも依存する。

出力層は、例えば、ソフトマックス層であり、フレームごとに、フレームの LSTM 出力を処理して、フレームのラベルスコアのセットを生成するように構成される。

分類サブシステム 130 は、ラベルスコア 122Ａ～122Ｎのセットを受信し、ラベルスコアのセットを使用してビデオを分類する。分類サブシステム 130 は、様々な方法のいずれかでスコアのセットを使用してビデオを分類することができる。

例えば、分類サブシステム 130 は、シーケンス内のビデオフレームのラベルスコアに従って、1 つまたは複数の最高スコアのラベルによって表されるトピックに関連するものとしてビデオを分類することができる。

別の例として、分類サブシステム 130 は、時系列の後半にあるフレームに割り当てられた重みが、時系列の前にあるフレームに割り当てられた重みよりも高くなるように、各フレームに重みを割り当てることによって、所定のラベルの結合されたラベルスコアを生成することができる。

次に、分類サブシステム 130 は、フレームごとに、フレームのラベルのラベルスコア

にフレームの重みを乗算することによってラベルの加重ラベルスコアを生成し、ラベルの加重ラベルスコアを合計することにより、ラベルの結合ラベルスコアを生成する。

トレーニング中および特定のトレーニングシーケンスに対して、システムはトレーニングシーケンスに関連付けられたラベルとタイムステップのラベルスコアを使用して各タイムステップの勾配を決定し、パラメータの現在値を更新するために LSTM ニューラルネットワーク及び畳み込みニューラルネットワークを介してタイムステップの勾配を逆伝播する。

システムは各タイムステップに重みを割り当て、トレーニングシーケンスの後半のタイムステップに割り当てられた重みが、トレーニングシーケンスの前のタイムステップに割り当てられた重みよりも高くなり、タイムステップの勾配を逆伝播する前に、タイムステップの重みを使用して勾配を調整する。たとえば、勾配にタイムステップの重みを掛ける。

さらに、一部の実装では、システムは、トレーニングビデオから派生したオプティカルフロー画像で畳み込みニューラルネットワークと LSTM ニューラルネットワークをトレーニングし、後期融合(Late Fusion)を実行して、一連のトレーニングシーケンスと、オプティカルフロー画像での畳み込みニューラルネットワークと LSTM ニューラルネットワークのトレーニングに関し、畳み込みニューラルネットワークと LSTM ニューラルネットワークのトレーニングの結果を結合する。

3.クレーム

912 特許のクレーム 1 は以下の通りである。

1. 複数のタイムステップのそれぞれにおける特定のビデオからのそれぞれのビデオフレームを含むビデオフレームの時間シーケンスを取得し、

複数のタイムステップの各タイムステップについて：

ビデオフレームの特徴を生成するために、畳み込みニューラルネットワークを使用してタイムステップでビデオフレームを処理し、

タイムステップのラベルスコアのセットを生成すべく、長期短期記憶（LSTM）ニューラルネットワークを使用してビデオフレームの特徴を処理し、ラベルスコアのセットは、所定のラベルセットの各ラベルのそれぞれのラベルスコアを含み、所定のラベルのセットのセット内の各ラベルは、それぞれのトピックを表し、

複数のタイムステップのそれぞれのラベルスコアから、ラベルのセット内のラベルによって表される 1 つ以上のトピックに関連するビデオを分類し、ここで、ビデオをトピ

ックの 1 つ以上に関連するものとして分類することは、以下を含み、

ビデオフレームの時間シーケンスの各タイムステップにそれぞれの重みを割り当て、ここで、時間シーケンスの後のタイムステップに割り当てられた重みは、時間シーケンスの前のタイムステップに割り当てられた重みよりも高く、

各ラベルのそれぞれの加重結合ラベルスコアを生成し、各ラベルについて、

各タイムステップについて、重み付けされたラベルスコアを生成するために、(i) タイムステップのラベルの LSTM ニューラルネットワークによって生成されたラベルスコアに（ii）タイムステップに割り当てられた重みを乗じ、

ラベルの加重結合ラベルスコアを生成するために、タイムステップの加重ラベルスコアを組み合わせて、

加重結合ラベルスコアに従って、1 つ以上の最高スコアのラベルで表されるトピックを選択する。

4. ニューラルネットワークを用いたビデオ分類モデルに関する論文

本特許に関連するビデオ分類に関する論文[1]が Joe Yue-Hei Ng 氏らにより発表されている。

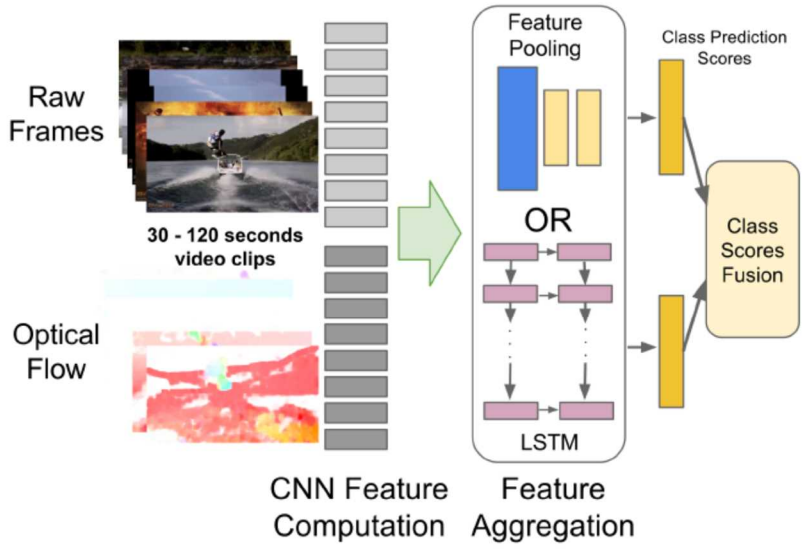


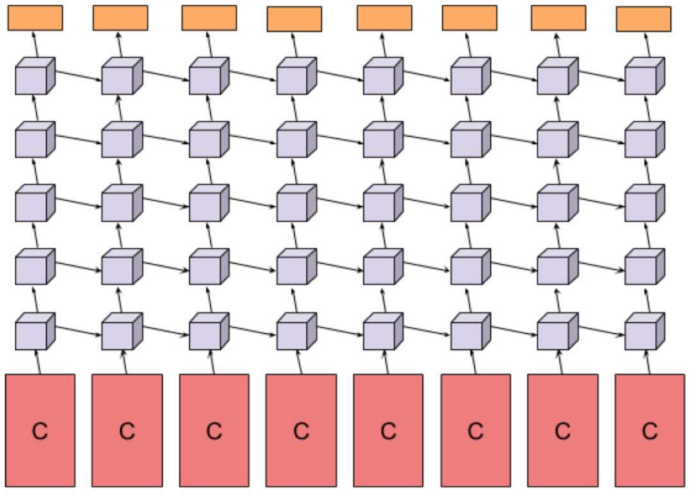
Figure 1: Overview of our approach.

図 1 は、分類モデルのネットワーク構成図である。分類システムには、ビデオのフレーム画像が入力される。また精度を向上させるためにオプティカルフロー画像も入力さ

[1] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, George Toderici “Beyond Short Snippets: Deep Networks for Video Classification” arXiv:1503.08909v2

れる。

フレーム画像が LSTM 層に入力され、並行してオプティカルフロー画像も LSTM 層に入力される。また LSTM 層に代えて特徴プーリング層をも利用することができる。



上記図は C で示す畳み込み層と、その後段の LSTM 層を示す。Deep Video LSTM は、連続する各ビデオフレームで最終 CNN レイヤーからの出力を受け取る。CNN 出力は、時間の経過とともに順方向に処理され、スタックされた LSTM の 5 つの層を介して上方に処理される。ソフトマックス層は、各タイムステップでクラスを予測する。畳み込みネットワーク（ピンク）とソフトマックス分類器（オレンジ）のパラメータは、タイムステップ間で共有される。

Category	Method	Frames	Clip Hit@1	Hit@1	Hit@5
Prior Results [14]	Single Frame	1	41.1	59.3	77.7
	Slow Fusion	15	41.9	60.9	80.2
Conv Pooling	Image and Optical Flow	120	**70.8**	72.4	**90.8**
LSTM	Image and Optical Flow	30	N/A	**73.1**	90.5

上記テーブルは、従来例と、本論文における特徴プーリング手法(Conv-Pooling)及び LSTM 手法との性能比較を示す。グローバルビデオレベル記述子を利用し、Sports-1M データセットを用いて評価を行った。LSTM と Conv-Pooling は、従来例と比較して Hit @ 1 で 20%の増加を達成し、また Hit @ 5 においても 10%の増加を達成している。

以上

著者紹介

河野英仁

河野特許事務所、所長弁理士。立命館大学情報システム学博士前期課程修了、米国フランクリンピアースローセンター知的財産権法修士修了、中国清華大学法学院知的財産夏季セミナー修了、MIT(マサチューセッツ工科大学)コンピュータ科学・AI 研究所 AI コース修了。

AI 特許コンサルティング、医療 AI 特許コンサルティングの他、米国・中国特許の権利化・侵害訴訟を専門としている。著書に「世界のソフトウエア特許(共著)」、「FinTech 特許入門」、「AI/IoT 特許入門 2.0」がある。