AI 特許紹介(36)
AI 特許を学ぶ！究める！
～対照学習(Contrastive Learning)～

2022 年 1 月 7 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
特許出願人　Google
出願日　2020 年 9 月 11 日
公開日　2020 年 8 月 11 日
公開番号 US20210319266
発明の名称　視覚的表現の対照学習のためのシステムと方法

266 特許は、教師データラベルを用いることなく精度を向上させる対照学習に関し、対照学習により学習された投影ヘッドニューラルネットワークと少ない教師データラベルを用いて、目的とする学習モデルをファインチューニングし、さらにファインチューニングした学習モデルを蒸留して学生モデルを生成する技術である。

2.特許内容の説明
教師データを用いることなく視覚表現を学習することは長年の課題である。本発明は下記の手法により当該課題を解決する。

図 2A は、対照学習のフレームワークを示す説明図である。

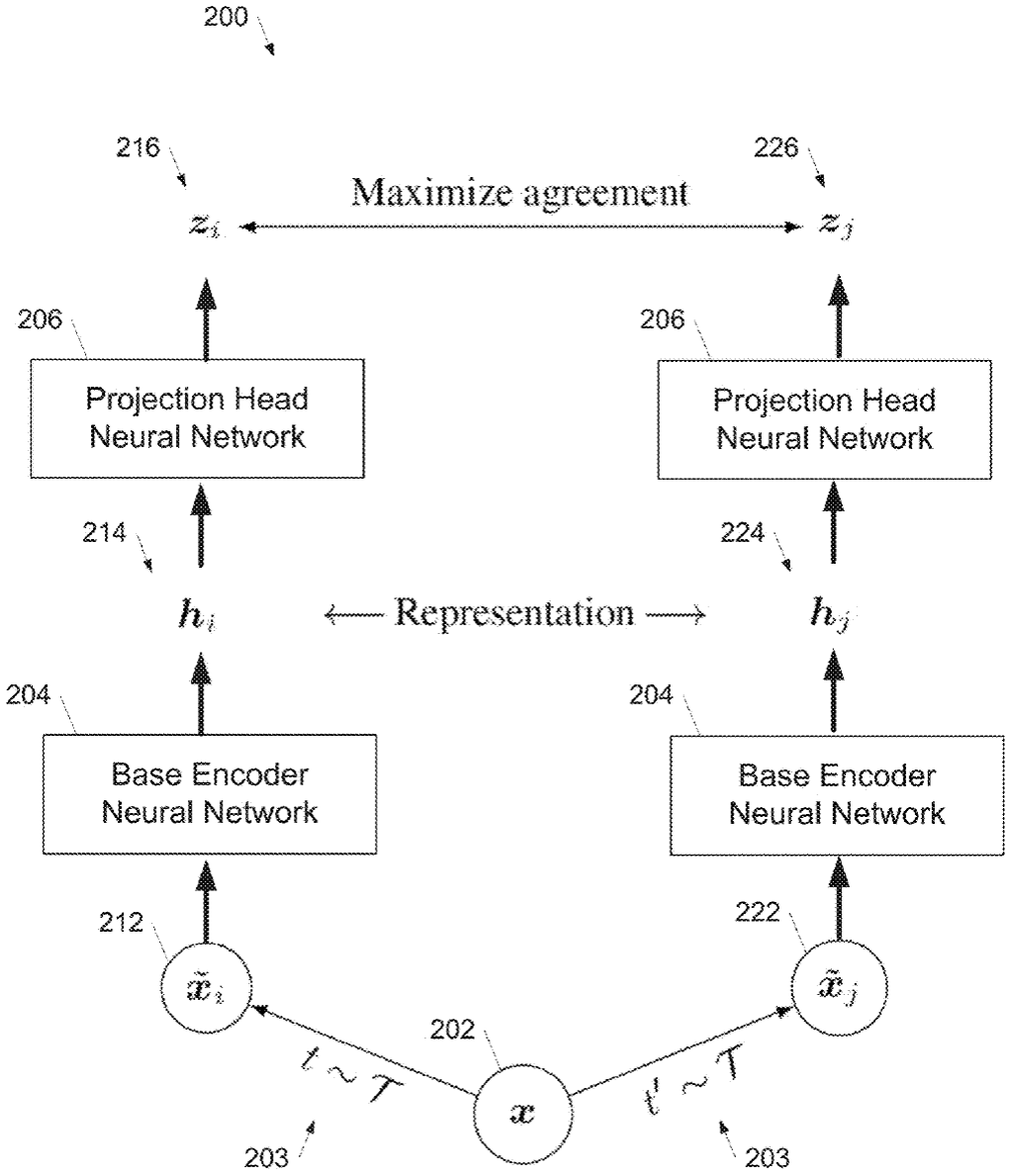


図 2A　対照学習のフレームワークを示す説明図

潜在空間における対照損失を介して、同じデータ例を異ならせて拡張したビュー間の一致を最大化することによって表現を学習する。図２Ａに示すように、フレームワーク 200 は、以下の４つの主要な構成要素を含む。

確率的データ拡張モジュール 203 は、任意のデータ例（たとえば、202 に示されている入力画像 x）をランダム変換して、同じ例の２つの相関ビューを生成する（x_i212,x_j213）。これらの拡張画像 212 および 222 は、正のペアと見なされる。

例えば、203 では、ランダムなクロッピングとそれに続く元のサイズへのサイズ変更、ランダムな色の歪み、ランダムなガウスぼかし等である。なお、ランダムなクリッピングと色の歪みの組み合わせは、優れたパフォーマンスを発揮する。

ベースエンコーダニューラルネットワーク 204（f（·）として表記で表される）は、拡張データの例から中間表現ベクトルを抽出する。例えば、図 2A に示すように、ベースエンコーダニューラルネットワーク 204 は、拡張画像 212 および 222 からそれぞれ中間表現 210 および 224 を生成する。

サンプルフレームワーク 200 は、例えば ResNet アーキテクチャを採用して h_i を取得する。

$$h_i=f(\tilde{x}_i)=\mathrm{ResNet}(\tilde{x}_i)$$

ここで h_i は、平均プーリング層の後の出力である。

投影ヘッドニューラルネットワーク 206（g（·）として表記で表される）は、対照的な損失が適用される空間内の中間表現を最終表現にマッピングする。例えば、投影ヘッドニューラルネットワーク 206 は、中間表現 210 および 224 からそれぞれ最終表現 216 および 226 を生成する。

投影ヘッドニューラルネットワーク 206 は、z_i を得るために 1 つの隠れ層を有する多層パーセプトロンである。

$$z_i=g(h_i)=W^{(2)}\sigma(W^{(i)}h_i)$$

ここで、σ は、ReLU の非線形性である。

対照的な損失関数は、対照的な予測タスクに対して定義できる。一例として、例 x_i212 と x_j222 の正のペアを含む集合$\{x_k\}$が与えられた場合、対照予測タスクは、たとえば、それぞれの最終的な表現 216 と 226 の間に基づき、与えられた x_i の$\{x\}_{k\neq i}$ の x_j を識別することを目的としている。

フレームワーク内でトレーニングを実行するために、N 個の例のミニバッチをランダムにサンプリングし、ミニバッチから派生した拡張例のペアで対照予測タスクを定義し、2N 個のデータポイントを作成する。また負(negative)例は明示的にサンプリングされない。代わりに、正(positive)のペアが与えられると、ミニバッチ内の他の 2（N-1）拡

張例は負の例として扱うことができる。
$\mathrm{sim}(u,v)=u^{T}v/\|u\|\|v\|$
が 2 つのベクトル u と v の間の余弦の類似性を表すとする。

次に、例の正のペア(i,f)の 1 つの例の損失関数を下記式(1)として定義できる。

$$\ell_{i,j} = -\log \frac{\exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k\neq i]} \exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)}, \qquad (1)$$

ここで、$k\neq i\in\{0,1\}$は、$k\neq i$ で τ が温度パラメータを表す場合に 1 と評価される指示関数である。

最終的な損失は、ミニバッチで(i,j) 及び(j,i)の両方のすべての正のペアにわたって計算される。この損失は、NT-Xent(正規化された温度スケーリングされたクロスエントロピー損失）と呼ばれる。

図 2B はフレームワーク訓練後のベースエンコーダニューラルネットワークの使用例を示す説明図である。

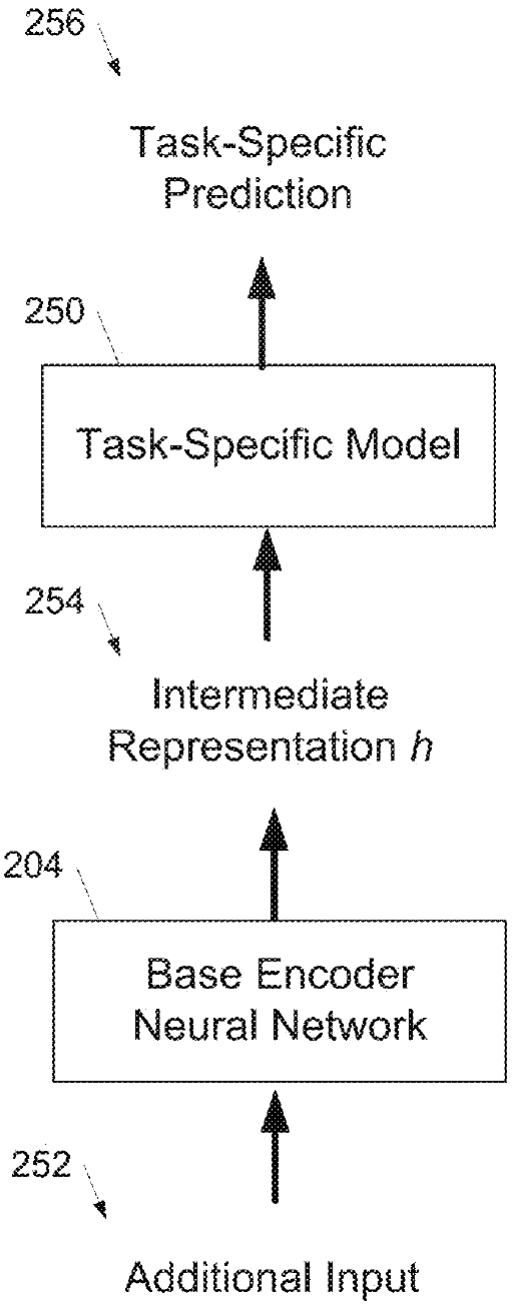


図 2B　フレームワーク訓練後の使用例

ベースエンコーダニューラルネットワーク 204 が抽出され、追加のタスク固有モデル 250 がベースエンコーダニューラルネットワーク 204 に追加される。例えば、タスク固有のモデル 250 は、線形モデルまたはニューラルネットワークなどの非線形モデルを含む任意の種類のモデルである。

タスク固有のモデル 250 およびベースエンコーダニューラルネットワーク 204 は、追加のトレーニングデータ（例えば、タスク固有のデータ）について追加的にトレーニング（例えば、「ファインチューニング」）することができる。追加のトレーニングは、たとえば、教師あり学習トレーニングである。

ファインチューニング後、追加の入力 252 を、中間表現 254 を生成するベースエンコーダニューラルネットワーク 204 に提供する。タスク固有のモデル 250 は、中間表現 254 を受信および処理して、タスク固有の予測 256 を生成する。例として、タスク固有の予測 256 は分類予測、検出予測、認識予測、セグメンテーション予測、または他の予測タスクである。

図 3 は、対照表現学習のためのデータ拡張手法の例を示す説明図である。

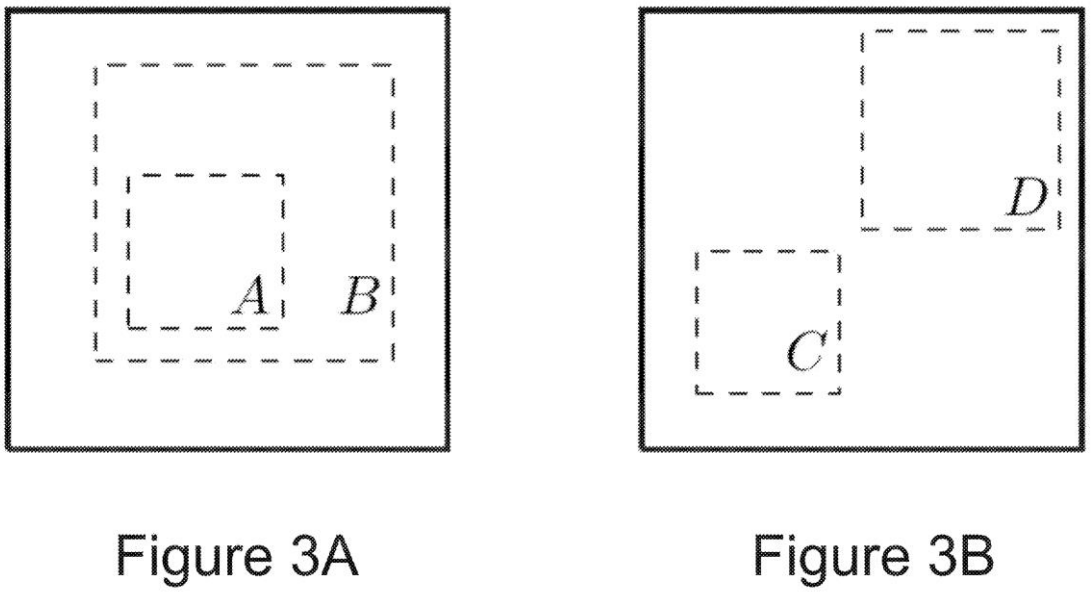


図 3　対照表現学習のためのデータ拡張手法の例を示す説明図

データ拡張は、予測タスクを定義する。多くの既存のアプローチは、アーキテクチャを変更することによって対照的な予測タスクを定義する。Hjelm と Bachman は、ネットワークアーキテクチャの受容野を制約することにより、グローバルからローカルへのビュー予測を実現る。一方、Oord と Hénaff は、固定画像分割手順とコンテキスト集約ネットワークを介して隣接ビューの予測を実現する。ただし、これらのカスタムアーキテクチャは複雑さを増し、結果モデルの柔軟性または適用性を低下させる。

ここで説明する手法は、ターゲット画像の単純なランダムクロッピング（サイズ変更を伴う）を実行することでこの複雑さを回避する。これにより、上記の既存のアプローチを含む一連の予測タスクが作成される。

図３Ａおよび３Ｂは、この原理を示している。 図 3A はグローバルビューとローカルビューを示し、図 3B は隣接するビューを示している。具体的には、実線の長方形は画像であり、破線の長方形はランダムな切り抜きである。画像をランダムにクロッピングすることにより、提案されたシステムは、グローバルからローカルビュー（B→A）または隣接ビュー（D→C）の予測を含む対照的な予測タスクをサンプリングできる。

拡張のタイプには、データの空間的/幾何学的変換（水平方向の反転を使用）、回転、切り抜きなどが含まれる。別のタイプの拡張には、色の歪み（色のドロップ、明るさ、コントラスト、彩度、色相を含む）、ガウスぼかし、ソーベルフィルタリングなどの外観変換が含まれる。図４に拡張例を示す。

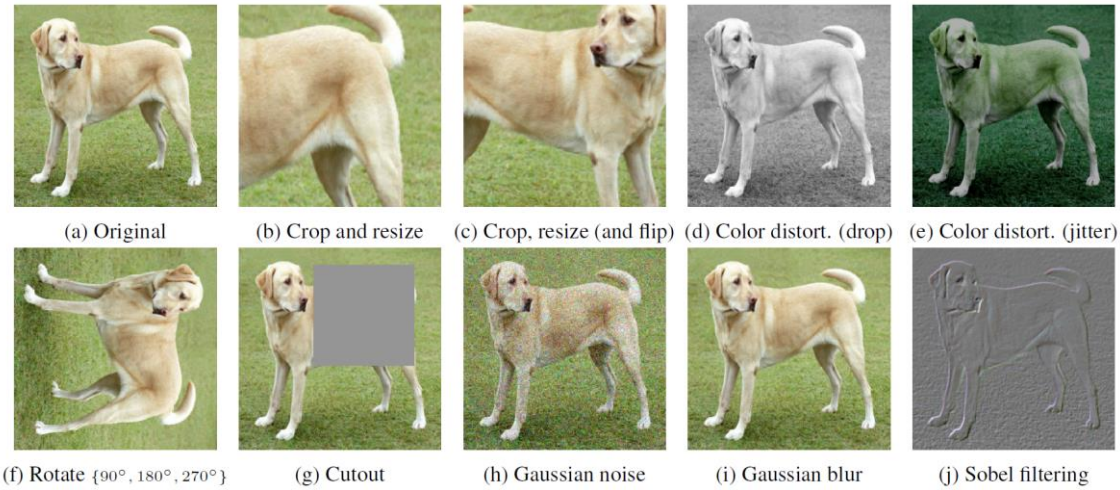


図４　拡張例

クロップ、サイズ変更（および反転）、色の歪み（ドロップ）、色の歪み（ジッター）、回転、切り取り、;ガウスノイズ、ガウスぼかし、および Sobel フィルタリングが含まれる。

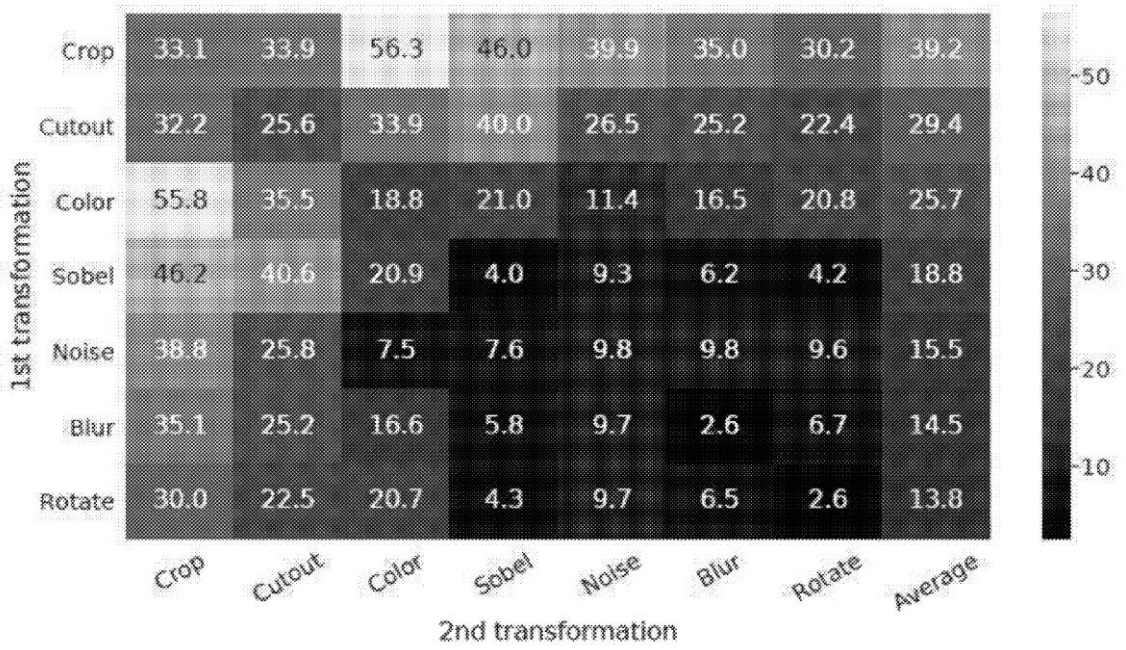


図 5 は、変換の個々の構成の下での線形評価結果を示している。すべての列で、対角線のエントリは単一の変換に対応し、非対角線は 2 つの変換の合成に対応する。最後の列は、行全体の平均を反映している。

図 5 から、モデルが対照タスクの正のペアをほぼ完全に識別できる場合でも、優れた表現を学習するのに単一の変換では不十分であることがわかる。合成拡張の場合、対照的な予測タスクは難しくなるが、表現の品質は劇的に向上する。合成拡張の 1 つ「ランダムクロッピングとランダムな色の歪み」の構成が際立っている。

図 11 は、視覚表現の半教師あり対照学習を実行するためのフローチャートである。

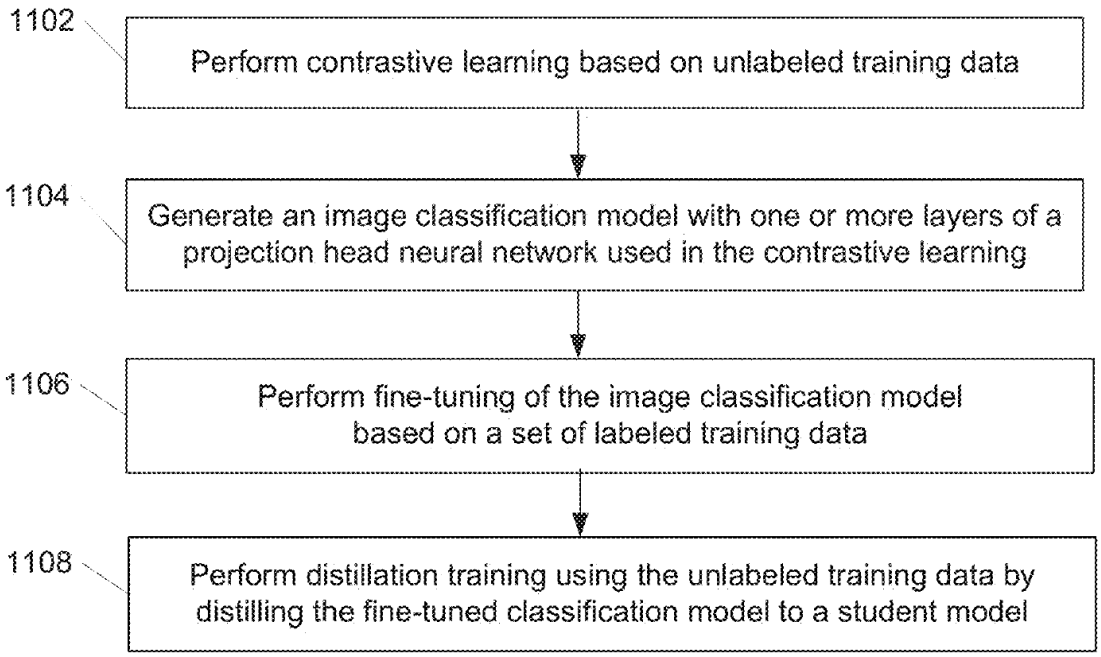


図 11 視覚表現の半教師あり対照学習を実行するためのフローチャート

ブロック 1102 においてコンピュータシステムは、ラベル付けされていないトレーニングデータのセットに基づく対照学習を使用して、モデルの教師なし事前トレーニングを実行する。たとえば、コンピュータシステムは、多数のラベルなしトレーニングデータを使用して、タスクにとらわれない大規模な一般畳み込みネットワークを事前トレーニングする。一例では、コンピュータシステムは、修正バージョンの SimCLR を使用して、大規模モデルの教師なし事前トレーニングを実行する。

ブロック 1104 で、コンピューティングシステムは、対照学習で使用される投影ヘッドニューラルネットワークの 1 つまたは複数の層を備えた画像分類モデルを生成する。

図 12 は、ファインチューニングを実行するためのグラフである。

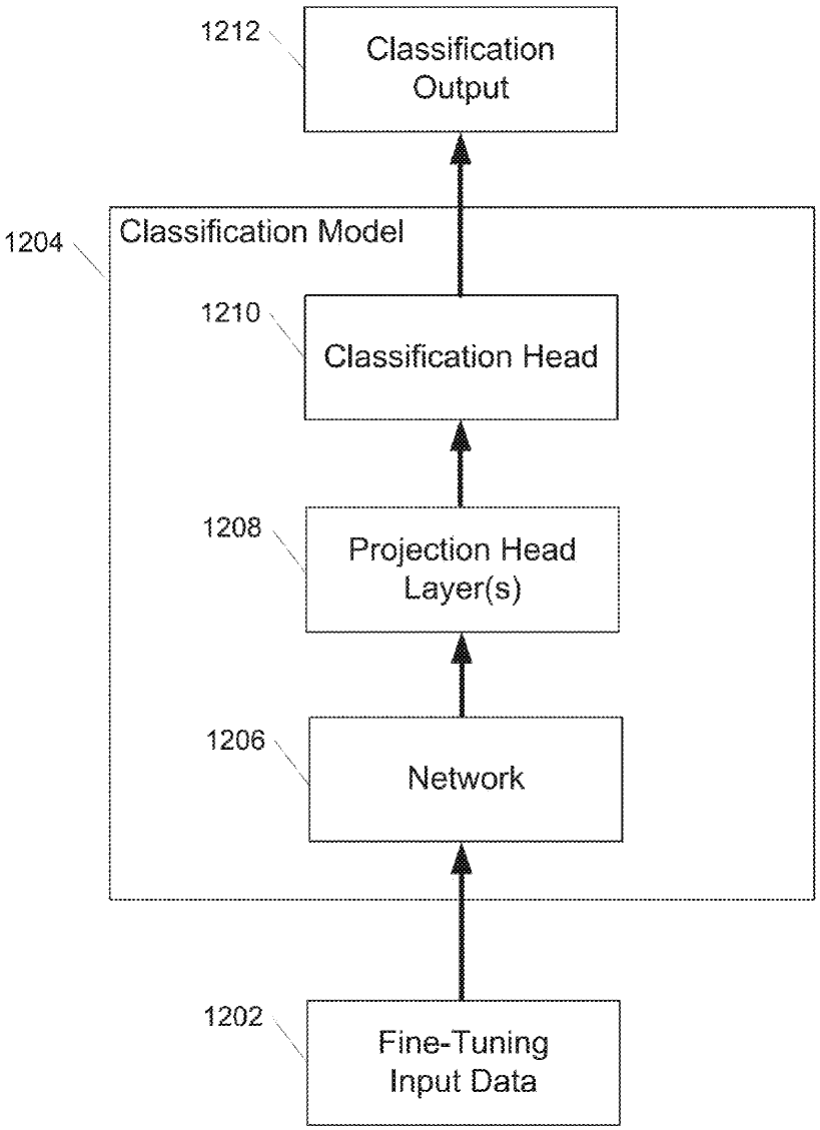


図 12　ファインチューニングを実行するためのグラフ

グラフ 1200 は、ファインチューニング入力データ 1202、分類モデル 1204、および分類出力 1212 を含む。分類モデル 1204 は、ネットワー1206、投影ヘッド層 1208、および分類ヘッド 1210 をさらに含む。

分類モデル 1204 は、ラベル付けされていないトレーニングデータを使用して対照学習で事前トレーニングされたタスクに依存しないネットワーク（例えば、事前トレーニングされたベースエンコーダニューラルネットワーク、規模の大きい畳み込みニューラルネットワークなど）などのネットワーク 1206 を含む。

分類モデル 1204 はまた、ラベル付けされていないトレーニングデータ（すなわち、投影ヘッド層 1208）を使用して対照学習で事前訓練された投影ヘッドニューラルネットワークの複数の層の一部を再利用する。

分類ヘッド 1210 は、一般に、１つまたは複数の表現を受信および処理して、分類予測、検出予測、認識予測、セグメンテーション予測または他のタイプの予測および予測タスクなどの分類出力 1212 を生成する。

ブロック 1106 で、コンピューティングシステムは、ラベル付けされたトレーニングデータのセットに基づいて画像分類モデルのファインチューニングを実行する。例えば、コンピュータシステムは、ラベルのない事前トレーニングサンプルの数と比較して、比較的少数または割合のラベル付けされたトレーニングデータ（例えば、１％、５％、１０％）を含む一組のファインチューニング入力データ 1202 に基づいて、事前訓練された分類モデル 1204 のファインチューニングを実行する。

分類モデル 1204 は、ラベル付けされたファインチューニング入力データ 1202 のセットを取得する。ラベル付けされたファインチューニング入力データ 1202 は、ネットワーク 1206 および投影ヘッド層 1208 によって処理される。ネットワーク 1206 は、ラベルのないトレーニングデータを使用した対照的な学習を使用して事前トレーニングされた、タスクに依存しない事前トレーニングされたネットワークである。

すべてではないがいくつかの事前訓練された投影ヘッド層 1208 は、事前訓練されたベースエンコーダニューラルネットワーク 204 である事前訓練されたネットワークの上部に１つまたは複数のそれぞれの線形変換層として追加される。

したがって、分類モデル 1204 のファインチューニングは、例えば、監視あり交差エントロピー損失または他のタイプの損失関数を使用して、ラベル付けされたファインチューニング入力データ 1202 に基づいて様々なパラメータを調整することによって実行され、分類を可能にする。分類モデル 1204 を使用して、1 つ以上の特定のタスクの内部表現をわずかにファインチューニングする。

ブロック 1108 で、コンピューティングシステムは、対照学習からのラベルのないトレーニングデータを使用して蒸留トレーニングを実行し、ファインチューニングされた分類モデルが比較的小さい学生モデルに蒸留される。

たとえば、コンピューティングシステムは、1 つまたは複数の対象タスクに特化した

軽量の学生ネットワークのトレーニングの一部として蒸留を実行するときに、ラベルのない事前トレーニングデータを直接再利用できる。

そのため、ラベルのないトレーニングデータは、最初にタスクに依存しない方法で事前トレーニングに使用され、次にファインチューニングを実行した後に蒸留で再び使用され、1 つ以上の特殊な対象タスクの学生ネットワークをトレーニングする。

図 13 は、学生モデルの蒸留訓練を実行するためのグラフ図 1300 を示す。

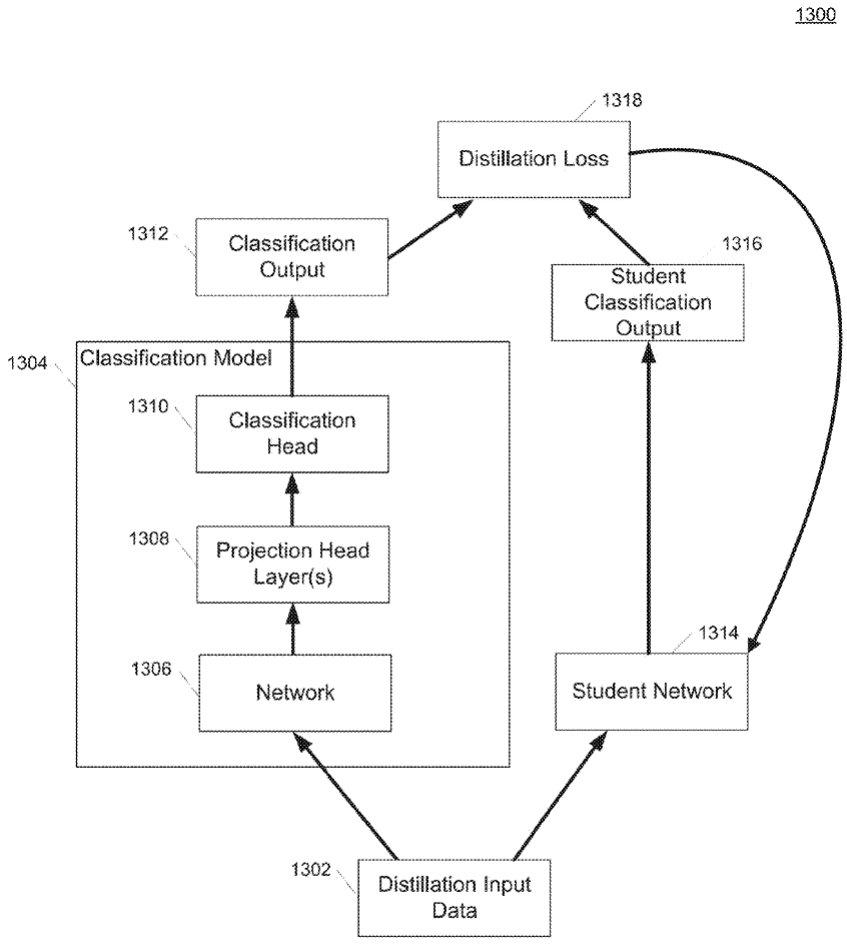


Figure 13

図 13 学生モデルの蒸留訓練を実行するためのグラフ図

グラフ図 1300 は、蒸留入力データ 1302、分類モデル 1304、ネットワーク 1306、投影ヘッド層 1308、分類ヘッド 1310、分類出力 1312、学生ネットワーク 1314、学生分類出力 1316、および蒸留損失 1318 を含む。

コンピューティングシステムは、蒸留入力データ 1302 を取得する。蒸留入力データ 1302 は、一般に、モデルの事前訓練で使用されるラベルのないデータの一部またはすべてを含む。

ラベルなし蒸留入力データ 1302 は、タスクに依存しない方法でモデルを事前トレーニングする際に最初に使用され、一つまたは複数のタスクに特化した学生モデルを蒸留するために、モデルのファインチューニングを実行した後に再利用される。

ラベルなし蒸留入力データ 1302 が、処理のために分類モデル 1304 および学生ネットワーク 1314 に提供される。分類モデル 1304 は、画像分類モデルまたは任意の他のタイプの分類モデルである。分類モデル 1304 は、事前に訓練され、ファインチューニングされた分類モデルである。

分類モデル 1304 は、ラベルなしトレーニングデータを使用して対照学習で最初に事前トレーニングされたネットワーク（例えば、微調整されたベースエンコーダニューラルネットワーク、大規模な畳み込みニューラルネットワークなど）などのファインチューニングされたネットワーク 1306 を含み、その後、ラベル付けされたトレーニングデータの比較的少数のセットに基づいてファインチューニングされる。

分類モデル 1304、ネットワーク 1306、および投影ヘッド層 1308 のファインチューニングは、一般に、ブロック 1106 で実行される。分類ヘッド 1310 は、1 つまたは複数の表現を受信および処理して、分類予測、検出予測、認識予測、セグメンテーション予測、または他のタイプの予測および予測タスクなどの分類出力 1312 を生成する。

分類モデル 1304 は、標的タスクにより特化された学生ネットワーク 1314 を訓練するために使用される。例えば、ファインチューニングされた分類モデル 1304 は、蒸留トレーニングを実行して、画像分類モデルと比較して比較的少数のパラメータを含む学生ネットワーク 1314 にモデルを蒸留する際に使用される。

そのため、学生ネットワーク 1314 は一般に軽量であり、ローカルコンピューティングリソースが限られているクライアントコンピューティングデバイスに展開するのに適している。たとえば、学生ネットワーク 1314 は、モバイルデバイス、モノのインターネット（IoT）エッジデバイス、またはリモート処理のために送信されるのではなく、ローカルでデータが処理されるその他のクライアントデバイスを含む様々なタイプのクライアントコンピューティングデバイスで使用するために展開できる。

学生ネットワーク 1314 は、ラベルのない蒸留入力データ 1302 を取得し、学生分類出力 1316 を生成する。対照学習における事前訓練段階のラベル付けされていないデータは、ターゲットタスクのために学生ネットワーク 1314 を訓練するために再利用される。

ファインチューニングされた分類モデル 1304 は、学生ネットワーク 1314 を訓練するためのラベルを提供し、蒸留損失 1318 が最小化される。

3.クレーム

266 特許のクレーム 1 は以下の通りである。

1. 視覚的表現の半教師あり対照学習を実行するためのコンピュータ実装方法において、

1 つまたは複数のラベルのないトレーニング画像のセットでトレーニング画像を取得し、

トレーニング画像に対して、第 1 の拡張画像を取得するために、複数の第 1 の拡張オペレーションを実行し、

複数の第 1 の拡張オペレーションを実行することとは別に、トレーニング画像に対して、第 2 の拡張画像を取得するために、複数の第 2 の拡張オペレーションを実行し、

ベースエンコーダニューラルネットワークを用いて、第 1 の拡張画像の第 1 の中間表現および第 2 の拡張画像の第 2 の中間表現をそれぞれ生成するために、第 1 の拡張画像および第 2 の拡張画像を処理し、

複数の層を含む投影ヘッドニューラルネットワークを用いて、それぞれ第 1 の拡張画像の第 1 の投影表現および第 2 の拡張画像の第 2 の投影表現を生成するために、第 1 の中間表現および第 2 の中間表現を処理し、

第 1 の投影表現と第 2 の投影表現の相違を評価する損失関数を評価し、

損失関数に少なくとも部分的に基づいて、ベースエンコーダニューラルネットワークおよび投影ヘッドニューラルネットワークの一方または両方の 1 つまたは複数のパラメータの 1 つまたは複数の値を変更し、

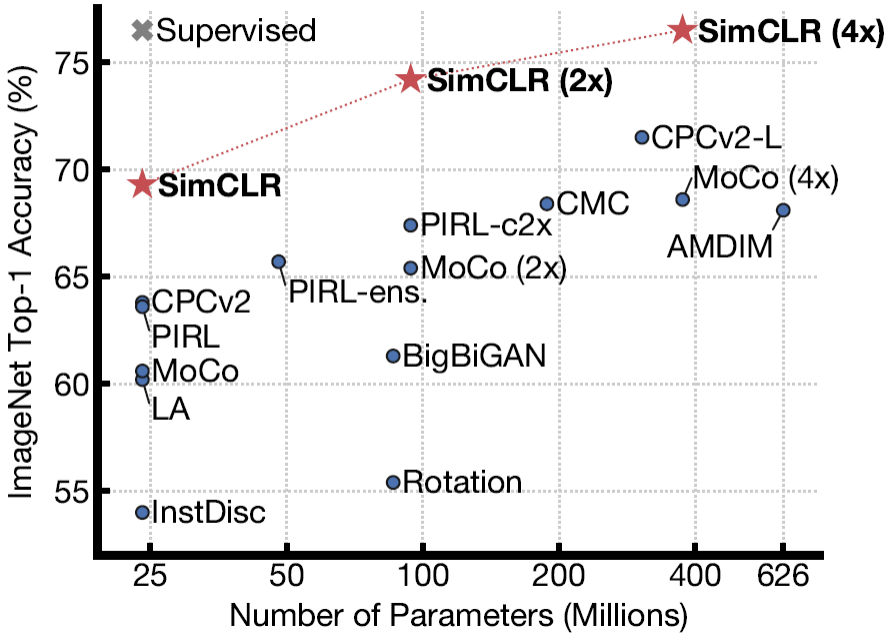
前記変更後、ベースエンコーダニューラルネットワークおよび投影ヘッドニューラルネットワークから画像分類モデルを生成し、該画像分類モデルは、投影ヘッドニューラルネットワークの複数の層の全てではないがいくつかを含み、

ラベル付けされた画像のセットに基づいて画像分類モデルのファインチューニングを実行する。

4. 本特許に関する論文

本特許に関する論文“A Simple Framework for Contrastive Learning of Visual Representations”が、Ting Chen らにより公表されている[1]。

下記図は、さまざまな自己教師あり方式で学習された表現でトレーニングされた線形分類器の ImageNetTop-1 精度（ImageNet で事前トレーニング済み）を示す。



灰色の×は、教師あり ResNet-50 を示す。SimCLR は太字で示されている。教師無しの対照学習である SimCLR はパラメータ数の増加に伴い、教師あり ResNet-50 と同等の精度を示していることがわかる。

提案されたシステムは、ImageNet での自己教師あり、及び、半教師あり学習の以前の方法を大幅に上回ることが経験的に示されている。提案されたシステムと方法の実装例によって学習された自己教師あり表現でトレーニングされた線形分類器は、76.5%のトップ 1 精度を達成する。これは、教師あり ResNet-50 のパフォーマンスに匹敵する、以前の最先端技術に比べて 7%の相対的な改善である。さらに、ラベルの 1%のみをファインチューニングすると、提案された手法の実装例は 85.8%のトップ 5 精度を達成し、100 少ないラベルで AlexNet を上回る。

[1] Ting Chen, Simon Kornblith, Mohammad Norouzi, Geoffrey Hinton “A Simple Framework for Contrastive Learning of Visual Representations” arXiv:2002.05709v3 [cs.LG] 1 Jul 2020

下記テーブル 8 は、ImageNet で事前トレーニングされた ResNet-50（4X）モデルについて、12 の自然画像分類データセットにわたる教師ありベースラインと自己教師ありアプローチの転移学習パフォーマンスとを比較したものである。

A Simple Framework for Contrastive Learning of Visual Representations

	Food	CIFAR10	CIFAR100	Birdsnap	SUN397	Cars	Aircraft	VOC2007	DTD	Pets	Caltech-101	Flowers
Linear evaluation:												
SimCLR (ours)	**76.9**	**95.3**	80.2	48.4	**65.9**	60.0	61.2	**84.2**	**78.9**	89.2	**93.9**	**95.0**
Supervised	75.2	**95.7**	**81.2**	**56.4**	64.9	**68.8**	**63.8**	83.8	**78.7**	**92.3**	**94.1**	94.2
Fine-tuned:												
SimCLR (ours)	**89.4**	**98.6**	**89.0**	**78.2**	**68.1**	**92.1**	**87.0**	**86.6**	**77.8**	92.1	**94.1**	97.6
Supervised	88.7	98.3	**88.7**	**77.8**	67.0	91.4	**88.0**	86.5	**78.8**	**93.2**	**94.2**	**98.0**
Random init	88.3	96.0	81.9	**77.0**	53.7	91.3	84.8	69.4	64.1	82.7	72.5	92.5

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」、「ブロックチェーン 3.0(共著)」がある。