2C1-OS-06a-3

Deep Learningの中間層学習表現を利用した動画像の意味解析

Semantic analysis of video using an intermediate layer representation of deep neural network

松本泰幸 *1　篠崎隆志 *2　上原邦昭 *3
Yasuyuki Matsumoto　Takashi Shinozaki　Kuniaki Uehara

*1神戸大学工学部情報知能工学科
Department of Computer Science and Systems Engineering, Kobe University

*2国立研究開発法人 情報通信研究機構 脳情報通信融合研究センター脳機能計測研究室
Brain Imaging Technology Laboratory, CiNet, National Insitute of Information and Communications Technology

*3神戸大学大学院システム情報学研究科
Graduate School of System Informatics, Kobe University

This study proposes a novel semantic analysis method for movies using learning representation of a deep neural network. We employs the pre-trained convolutional neural network trained by natural images in Imagenet, and combine it with SVM. The proposed method uses the learning representation as a feature of the input image, and enables the target classification by SVM with minimum learning cost. The experimental result exhibits the effectiveness of the proposed method both in speed and accuracy, compared with a conventional method using high-speed SIN technique. Furthermore, it suggests the sixth layer is the optimal for the utilization of the learning representation in the used network.

1. はじめに

近年，YouTube やニコニコ動画といった動画サイトの普及により，多くの映像が選択，視聴できるようになった．このような映像検索サービスで，大量の映像の中から欲しい情報に自在にアクセスするためには，映像の内容に基づく検索やブラウジングが必要不可欠となる．このような要求を実現するためには，映像に対してアノテーションを与えた上で，テキストを利用した検索手法が考えられる．しかし，手作業によるアノテーションの付与には，多大な労力を要し，加えて作成する際の恣意性，主観性などの課題が残されている．これに対して，対象のデータを大量に集めたアーカイブと，内容に基づく正解データが付与された，コーパスを用いるアプローチが有効である．本研究の目的は，コーパスを用いるアプローチに対し Deep Learning の適用を検討することである．

従来の識別法では，図 1 の (a) のように，人の手で設計された (Hand-crafted) 特徴量 (Feature) をもとに，教師あり学習で識別をする手法が提案されている．このままでは，学習された概念のみの識別だけが可能である．これに対して Deep Learning では，図 1 の (b) のように，入力された画像から識別に至る階層的な処理過程を直接的に学習している．さらに，その階層的な処理過程の途中で，図 1 の (a) のような従来の特徴量に相当する構造が，学習表現 (learning representation) として自然に獲得されることが知られている．こうした学習表現を画像の特徴量として使用し，中間層以降のみを再学習すれば，学習されていない新たな概念への識別も可能になると考えられる (図 1(c))．本研究では，このような考え方に基づいて，Convolutional Neural Network (CNN) [LeCun 89, Fukushima 80] に SVM を組み合わせることによって, 動画像の意味解析を行う．

連絡先: 松本泰幸，神戸大学工学部情報知能工学科，
matsumoto@ai.cs.kobe-u.ac.jp

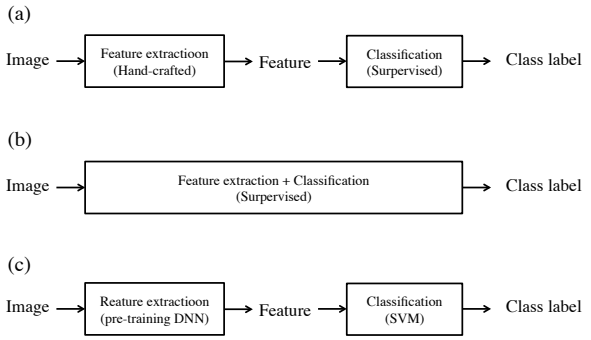

図 1: 画像認識のフロー．(a) 従来方法：設計した特徴量を使う．(b)CNN の教師あり学習．(c) 提案手法

2. 提案手法

本研究では，静止画像で事前学習された CNN モデルから，動画像の意味解析することを目的としている．事前学習で用いられた概念以外の識別のために，中間層の学習表現を利用し，線形 SVM で再学習を行う．これは，大規模データセットのみで学習済みの状態から，目的とする別のデータセットへ学習しなおす方法である，Fine-tuning を応用したモデルと考えられる．すなわち，SVM で中間層から先の Fine-tuning を行うことに相当している．

2.1 畳込みニューラルネットワークを用いた特徴抽出

一般に，Deep Learning の学習には大量のデータと，それに応じた学習時間が必要とされるため，動画像を認識するためのシステムを，ゼロから学習させることは現実的でない．そこで，すでに学習済みのネットワークの利用が考えられるが，本研究では，Deep Learning の CNN の計算処理として，C++で実装され，GPU に対応した Deep Learning ライブラリ caffe [Jia 14] を用いる．caffe では，リファレンスモデルとして大規模画像

認識のコンテスト ImageNet Large-scale Visual Recognition Challenge (ILSVRC) で，2012 年にトップとなった CNN の画像分類モデル [Krizhevsky 12] が利用できる．この画像分類モデルは，ILSVRC2012 のデータセットを利用して学習済みであり，あらかじめ決められた 1000 のカテゴリへ分類することが可能となっている（図 2）．しかしながら，1000 カテゴリとして用意された概念は必ずしも適切なものでなく、他のデータセットにそのまま適用することが困難な場合もある。例えば，動物や物の品種への偏りが強く，人や動作を表す概念は含まれていない．逆に，動物，例えば犬であれば，その犬種まで詳細な分類が可能となっている．

一方，Deep Learning は学習過程において、入力データの普遍的な特徴 (ここでは自然画像の一般特徴) が，より入力に近い層に学習表現として蓄積される，表現学習とよばれる現象を持つことが知られている [Bengio 13]．そこで本研究では二つの疑問を提起する．

- 静止画像の学習によって得られた学習表現を，一般的な動画像分類課題に適用可能か
- また，分類精度は学習表現を取り出す層に依存するのか

前者に関しては，同規模のモデルを新たな学習データを揃えて作成すれば，対象とする概念の識別は可能となるが，膨大な学習データと学習時間が必要とされるため，効果的ではない．そこで，リファレンスモデルが自然画像によって学習した学習表現を，CNN の中間層から取り出し，SVM で新たな概念の再学習を行なえば，目標とする分類が可能になると考えられる．

本研究での中間層の学習表現とは，畳み込み層およびプーリング層が交互に接続された，図 2 における CNN の第 6 層または第 7 層の出力である．この場合，第 6 層で出力される特徴量ベクトルを利用した場合と，第 7 層の場合で，精度がどのようか変化するのかという疑問が後者となる．最終層では 1000 次元であるが，第 6 層，第 7 層ではいずれも 4096 次元の特徴量ベクトルが出力される．入力層に近いほど汎用的な学習表現となるとしても，事前学習された概念の表現学習を，異なる概念の識別へ利用する場合，どの程度の汎用性をもつ学習表現が最適であるかを実験から考察する．

2.2 時間的マックスプーリング

動画像では，1 ショットに複数のフレームが含まれるために，2.1 節の構造から得られる特徴量ベクトルは，フレーム数分だけ得られることになる．そこで，ショット内フレームの特徴量ベクトルに対して，時間的なマックスプーリング (Max-Pooling) を適用して，ショットのベクトル表現としている．これにより，ショットとして様々な特徴情報をもつベクトルが生成され，全フレームの特徴量の学習を避けた上で，効率的に学習ができると考えられる．

具体的には，特徴量ベクトルを $\vec{x}_{ij} = (x_{1j}, x_{2j}, \ldots, x_{Mj})$ として表すと，ショット内の全フレーム特徴量ベクトルは $\vec{x}_{ij} = (x_{i1}, x_{i2}, \ldots, x_{MN})$ と表され，マックスプーリングの出力 x' は式 (1) から求めることができる．なお，M は特徴量の個数，本研究では 4,096 個であり，N はフレーム数である．

$$x'_i = \max(x_{i1}, x_{i2}, \ldots, x_{iN}) \tag{1}$$

1 ショットは数秒程度の長さを持つため，対象概念に対応する画像には，その画像的特徴 (様々な向き、大きさ、位置) の変化したものが，個々のフレームに出現している．この時，特徴量ベクトルの空間が，もし充分にスパースであれば，それぞれの特徴に対応するベクトルは独立に近い状態である．このため，フレーム間でのマックスプーリングは，ショット内の対象概念に対する，様々な特徴情報を集積することに対応している．例えば，人物の動画像においては，横顔や正面の顔などの特徴情報が集積することで，より人物と判定される．一方，飛行機の動画像であれば，遠くに小さく見える画像や，近くで大きく見える画像などの特徴情報の集積によって，より飛行機と判定されることになる．

2.3 SVM の学習

CNN の出力の特徴量ベクトルに対し，特定の特徴があるかないかの，２クラス分類問題を効率良く解く学習機械として，SVM を用いている．SVM の実装には C-support vector classification (C-SVC) [Chang 11] を用いた libSVM を使用する．CNN の中間層における出力には，事前に学習されたカテゴリへの識別を構成するための中間表現が含まれているため，SVM で新たな概念の再学習を行えば，目標とする分類が可能となる．本研究では，SVM をアノテーションごとに，概念が存在するか否かの判定を行うために適用する．さらに各ショットに対し，２値出力のみでは順位付けが困難であるため，SVM の実数出力値をシグモイド関数で $[0, 1]$ の範囲の値に変換した値を利用している．このようにすれば，対象概念への適合度を算出できることになる．

3. 評価実験

事前学習されたリファレンスモデルの再利用性と，提案手法の分類精度と計算時間の観点からの有効性を示すための評価実験を行った．実験には，学習済みリファレンスモデルとして，ILSVRC 2012 のデータセットを用い，識別する動画像として，TREC Video Retrieval Evaluation (TRECVID) 2012 の web 動画を利用する．

TRECVID は，NIST 主催の動画像の意味解析における国際型ワークショップである．また，TRECVID のタスクの一つに SIN (Semantic indexing) がある．SIN は，色，エッジ，動きといった特徴量に基づいて，特定の概念が映っているショットと映っていないショットを分類する問題である．具体的には，機械学習のアプローチを用いて，概念が映っている，もしくは映っていないとラベル付けされたショット（学習例）から，両者を判別するための識別器を学習する．そして，識別器を用いて，未知のショット（テスト例）中の概念を高精度に認識することが評価となる．

本研究では，SIN の参加者に提供されるデータセット，197,000 個 (400,238 ショット) の学習用映像，8,263 個 (145,634 ショット) のテスト映像を利用する．表 1 に示した 15 種類の概念を認識対象とする．各映像データには，それぞれアノテーションとして学習用映像のショットに概念の有無を表すラベルデータが付与されている．このラベルデータは，TRECVID の参加者が，インターネットを介した協調型映像アノテーションに参加し，ショット中に概念が映っているかどうか検証して作成されたものである．

本実験環境には，大規模演算処理の高速化を行うために GPU (NVIDIA 社製 Tesla K40) を搭載したワークステーションを用いている．CPU は Intel 社製 Xeon E5-2687W 2.5GHz，メモリは 16GB である．認識結果の評価に関しては，TRECVID の評価基準に従って，テスト用映像のショットを SVM の出力値が高い順にランク付けしたときの，上位 2,000 ショットに対する “平均精度 (AP: Average Precision)” で評価する．AP

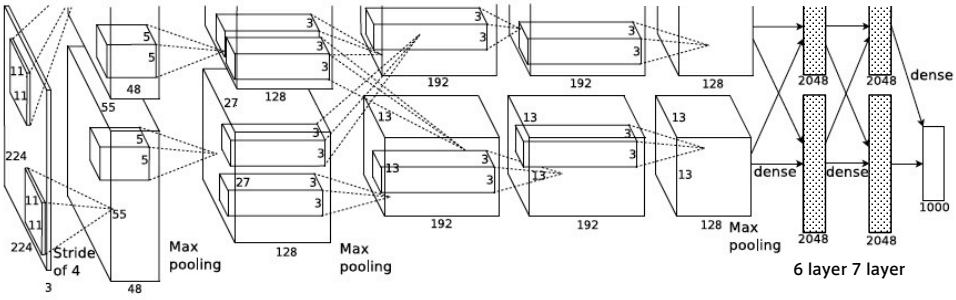

図 2: リファレンスモデルにおける特徴量ベクトルの出力構造 [Krizhevsky 2012 より一部改変].

表 1: 認識対象とした 15 種類の概念とその正例数，負例数

認識対象	正例数	負例数
Airplane Flying	460	29540
Bicycling	391	29609
Boot Ship	866	29134
Computers	1399	28601
Female Person	11063	18937
Instrumental Musician	3078	26922
Landscape	4406	25594
Male Person	15000	15000
Nighttime	2251	27749
Scene Text	3303	26697
Singing	4666	25334
Sitting Down	2481	27519
Stadium	784	29216
Throwing	289	29711
Walking Running	5607	24396

は，情報検索の分野で開発された評価尺度で，実際に概念が映っているショットが上位にランク付けされているほど高くなるようになっている．総合的な評価指標として 15 種類の概念に対する AP の平均を MAP(Mean Average Precision) として表す．

まず予備実験として，動画像からフレームの切り出しに関して実験を行った．フレーム数が増えると，その分データ量が増えるために，精度が向上すると考えられる．そこで，フレーム数による精度比較を図 3 に示す．1fps と 3fps とで，各カテゴリについて AP の差は出ていない．これより，切り出すフレーム数に精度は依存していないことが分かる．したがって，1 秒の動画像から 1 フレーム (1fps) として切り出している．なお，この実験のみ正例と負例の数を同じにしている．

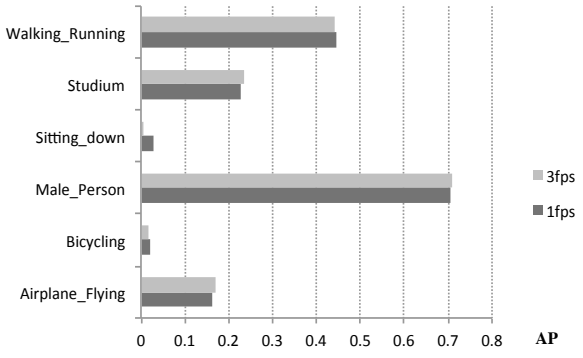

図 3: フレーム数による精度比較.

本研究の精度の比較手法として TRECVID2012 の SIN (light) 部門で 1 位に認識精度を達成した，高速化手法に基づく手法を採用する [白浜 13]．この手法は，行列演算に基づいて，大量の学習例間の類似度（カーネル値）を一括して計算する，高速な識別器の学習・テスト手法，および大量の記述子に対する確率密度を一括して計算する，高速な特徴量抽出手法 (STD-RGB-SIFT) からなる．この手法との精度と計算時間の比較から，本手法の有効性を示す．

3.1 再利用性の結果

まず，リファレンスモデルをオリジナルの状態で用いた分類結果を示す．TRECVID の正例画像に対して分類を行った時の結果を図 4 に示す．認識対象としたのは，上から Airplane-Flying, Bicycling, MalePerson で，出力結果は左から順位，概念，適合確率となっている．飛行機や自転車など，事前に学習された概念に対しては，対象の部品や種類といった，固有名詞までの概念が識別結果として現れている．逆に，“人” などの学習されていない概念，例えば “Male Person” に対しては，“Windsor tie” や “coat” といった服装を対象とした識別結果のみとなっている．つまり，事前に学習されている限られた概念へ識別されてしまい，所望する “男性” のような概念への識別が行われていないことが分かる．このことから，事前に学習された概念以外の識別に，リファレンスモデルを再利用することは，極めて困難であることが示唆される．

図 4: 正例画像をリファレンスモデルの 1000 カテゴリへ分類した結果.

3.2 精度比較の結果

つぎに，新たな概念の識別を可能とするために，中間表現を再学習させたときの識別結果を示す．図 5 に，第 6 層と第 7 層の特徴ベクトルを用いた TRECVID の各概念に対する認識精度，および比較手法である STD-RGB-SIFT の認識精度を示す．縦軸には認識対象の 15 種類の概念を並べ，横軸には AP の値をプロットしている．

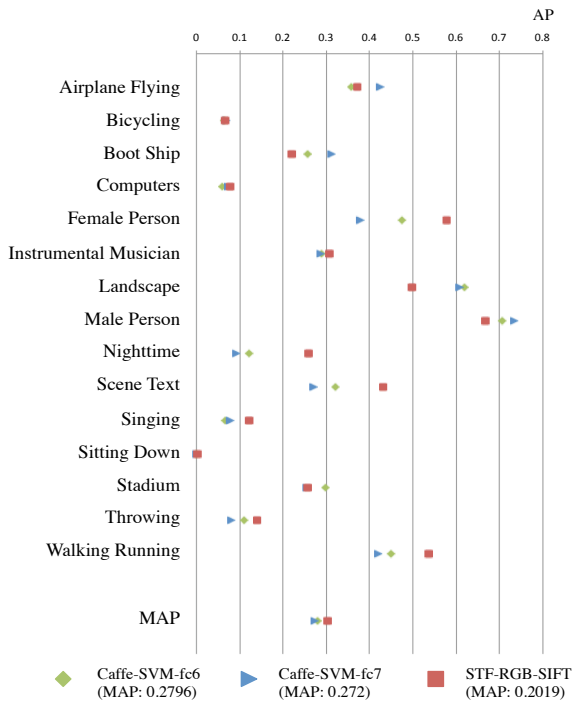

図 5: 比較手法と各出力層による精度の比較結果．

第 6 層を利用した識別結果の AP は，“Male Person” が最も大きく，“Sitting Down” が最も小さくなり，MAP は STD-RGB-SIFT の認識精度とほとんど変わらない結果を得ている (図 5)．特に “Landscape”，“MalePerson”，“Stadium” に関しては，比較手法を大きく上回っている．このように，比較手法と同精度の結果を得たことから，静止画像の学習によって得られた学習表現は，一般的な動画像分類課題に適用可能であり，本手法の精度の観点からの有効性が示唆される．

第 7 層を利用した識別結果の AP は，多くの概念では第 6 層の結果を下回る結果が見られ，全体としても下回る結果となっている．これは，利用した中間表現が最終層に近すぎるため，事前学習された概念に特化された，汎用性を持たないものになっていると考えられる．よって，第 6 層の中間表現を用いて，再学習を行うのが最適であると示される．

3.3 計算時間の結果

計算時間について比較手法との差を示す．比較手法では全 545,872 ショットから特徴量を抽出するために，50 プロセス並列で 1 ヶ月弱の時間を要したのに対し，本手法では caffe と時間的マックスプーリングを適用しているため，1 台の GPU マシンにより 4 日程度の時間で識別を完了した．精度を落とさずに，効果的な抽出が行えることから，本手法の計算時間の観点からの有用性が示唆される．

4. 結論

本稿では，リファレンスモデル（ILSVRC 2012）の再利用に関して，一般的な動画像分類課題に適用は難しいことから，中間層から出力された多次元の特徴量ベクトルを SVM により再学習を行い，目的とする認識対象に分類する手法を提案した．第 6 層の中間表現を利用した識別では，認識対象としたいくつかの概念で，比較手法を上回る精度が得られ，MAP では同精度を達成した．

一方で，計算速度に関しては，比較の先行研究に対して十分速く，競争力のあるものであることが分かる．以上から，本手法の精度と計算時間の観点からの有用性が示唆される．また，提起した疑問にもあるように，第 6 層の中間表現を利用したほうが，第 7 層よりも精度が高いため，この特徴量の汎化性は出力層に左右される上で，リファレンスモデルの再利用に関して有効であると考えられる．今後は，6 層以前の中間表現を利用した識別と，特徴量ベクトルを利用した映像の意味解析で，より正確な分類が可能であるか，検証する予定である．

参考文献

[Bengio 13] Bengio, Y., Courville, A., and Vincent, P.: Representation learning: A review and new perspectives, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 35, No. 8, pp. 1798–1828 (2013)

[Chang 11] Chang, C.-C. and Lin, C.-J.: LIBSVM: A library for support vector machines, *ACM Transactions on Intelligent Systems and Technology*, Vol. 2, pp. 27:1–27:27 (2011)

[Fukushima 80] Fukushima, K.: Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position, *Biological Cybernetics*, Vol. 36, pp. 193–202 (1980)

[Jia 14] Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., and Darrell, T.: Caffe: Convolutional Architecture for Fast Feature Embedding, *arXiv preprint arXiv:1408.5093* (2014)

[Krizhevsky 12] Krizhevsky, A., Sutskever, I., and Hinton, G. E.: ImageNet Classification with Deep Convolutional Neural Networks, in Pereira, F., Burges, C., Bottou, L., and Weinberger, K. eds., *Advances in Neural Information Processing Systems 25*, pp. 1097–1105, Curran Associates, Inc. (2012)

[LeCun 89] LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D.: Backpropagation Applied to Handwritten Zip Code Recognition, *Neural Comput.*, Vol. 1, No. 4, pp. 541–551 (1989)

[白浜 13] 白浜 公章, 上原 邦昭：行列演算に基づく高速かつ厳密な大規模映像データ処理, 映像情報メディア学会誌, Vol. 67, No. 7, pp. J241–J251 (2013)