G-15

確率的ディープラーニングによる画像カテゴリ認識における特徴量に関する一検討

A study of representations used for probabilistic deep learning in image category recognition

円道 滉一郎 †　江口 浩二 †
Endo Koichoro　Eguchi Koji

1. はじめに

近年，機械学習やパターン認識などの分野でディープラーニングの技術 [1] が注目を集めている．とりわけ，画像カテゴリ認識で成功的である事が報告されている [2]．

ディープラーニングの特徴的な点は，生データを直接与えるとそのデータの潜在的な特徴を自動的に抽出し，データの潜在特徴表現を学習する点である．ディープラーニングはニューラルネットワークを複数繋ぎ合わせて多層にしたものであり，これによりデータの潜在的な特徴の抽出精度を高めることが可能となる．ところで，従来のパターン認識において一般的とされていたアプローチは，経験則に基づく何らかの特徴抽出法で生データから抽出された局所特徴量を用いて，ベクトル量子化を行う事で得られた表現を Support Vector Machine(SVM)[3] などの分類器に通し，出力された結果を見て評価をするというものであった．

本研究ではディープラーニング，とりわけ確率的アプローチの Deep Belief Network(DBN)[4,20] に着目する．そして，DBN において有効な特徴について検証するため，画像データを生データと局所特徴量表現それぞれを用いて DBN によるカテゴリ認識を実際に行い，比較する事でその有用性を確かめる．次に，生データを入力としたモードと局所特徴量表現を入力としたモードの 2 つのモードを組み合わせ，マルチモーダルな DBN[5] を構築し，そのマルチモーダルな DBN を扱うとカテゴリ認識において認識精度が上がるのではないかと考え，その点について検証する．

ニューラルネットワークとして広く扱われているモデルに Restricted Boltzman Machine(RBM)[6,20] があり，RBM はデータの潜在的特徴の抽出を確率分布上で行える利便性が高いモデルとして知られている．RBM はバイナリデータを扱うモデルとして提案されており，この RBM を実数の連続データを扱う際に有効なように拡張したモデルとして Gaussian Restricted Boltzman Machine(GRBM)[7] が，文書を単語の頻出度合で表現した Bag of Words などスパースな離散データを扱う際に有効なように拡張したモデルとして Replicated Softmax Model[8] が提案されている．これらをさらに RBM で多層的に統合したモデルが我々の扱うマルチモーダルな DBN であり，構築したマルチモーダルな DBN を用いて学習を行い，カテゴリ認識の精度を調べる事で前述の検証を行う．

† 神戸大学, Kobe University

2. 関連研究

機械学習において，データと共にそのデータと関連のある学習すべき付随情報 (以下，ラベル) が与えられた上で，ラベルが与えられていないデータが与えられた時にそのデータに対応するラベルを予測する際の規則や関数などを学習する事を教師あり学習と呼ぶ．また，ラベルを伴わず，データのみを用いてデータの潜在的な特徴パターンを見つけようとする学習を教師なし学習と呼ぶ．ディープラーニングでは，まず与えられたデータのみを用いて教師なし学習を行う．教師なし学習を行うことでデータの潜在的な特徴を学習し，学習された潜在的特徴を用いて今度は教師あり学習を行い，その後に評価をする．このような手順を踏む事が有効であると知られている [9]．教師なし学習を行う手順を事前学習 (pre-training) と呼び，教師あり学習を伴う手順をファインチューニングと呼ぶ．本章では確率的ディープラーニングの事前学習にしばしば用いられる Restricted Boltzman Machine(RBM)[6,20] とそのいくつかの変形について述べる．次に確率的ディープラーニングの一実現方法である Deep Belief Network(DBN)[4,20] とそのファインチューニングについて述べる．

2.2 Restricted Boltzman Machine(RBM)

Restricted Boltzman Machine(RBM)[6,20] は確率分布を用いる生成モデルである．モデルの構成は 2 層構造のネットワークを成しており，1 つの層は入力データが与えられる可視層と呼ばれる層で，もう 1 つの層は可視層に与えられた入力データの特徴を表すデータが与えら

れる隠れ層と呼ばれる層である．可視層と隠れ層はそれぞれユニットを持っており，可視ユニットと隠れユニットと呼ばれる．可視ユニット同士，隠れユニット同士は結合せず，任意の可視ユニットと隠れユニットの間にのみ結合が制限されている完全 2 部グラフのネットワークとなっている．ここで，可視層の状態ベクトルを $\mathbf{v}$，隠れ層の状態ベクトルを $\mathbf{h}$ とし，可視ユニットを v_i，隠れユニット h_j で表すとすると，RBM のグラフィカルモデルは次のようになる．

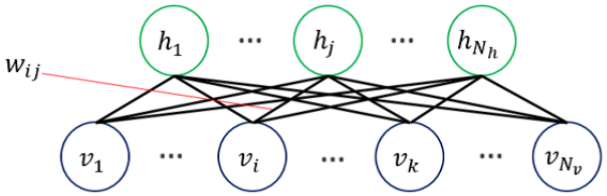

Fig 1: Graphical model of RBM.

N_v, N_h は $\mathbf{v}, \mathbf{h}$ の次元数である．

通常，RBM では可視ユニット v_i と隠れユニット h_j には 0 または 1 の 2 値が入る．RBM はネットワーク全体からエネルギーが定義され，そのエネルギーによって確率が定義されるモデルであり，エネルギーを $E(\mathbf{v}, \mathbf{h})$ とすると，エネルギー関数と確率 $p(\mathbf{v}, \mathbf{h})$ は次の式で定義される．

$$p(\mathbf{v}, \mathbf{h}) = \frac{1}{Z(\theta)} \exp(-E(\mathbf{v}, \mathbf{h}))$$

$$Z(\theta) = \sum_{\mathbf{v}} \sum_{\mathbf{h}} \exp(-E(\mathbf{v}, \mathbf{h}))$$

$$E(\mathbf{v}, \mathbf{h}) = -\sum_{i=1}^{N_v} b_i v_i - \sum_{j=1}^{N_h} c_j h_j - \sum_{i=1}^{N_v} \sum_{j=1}^{N_h} v_i W_{ij} h_j$$

$\theta = (\mathbf{b}, \mathbf{c}, \mathbf{W})$ は RBM のパラメータである．$\mathbf{b}, \mathbf{c}$ はそれぞれ可視ユニット，隠れユニットの持つバイアスであり，$\mathbf{W}$ は可視ユニットと隠れユニット間の結合の重みである．$Z(\theta)$ は確率をすべて足し合わせた時に 1 にするための正規化定数である．エネルギーが小さい状態ほど起こる確率は高くなる．

RBM での学習はパラメータを可視ユニットの状態のみで推定する事である．そのため，パラメータ θ について最尤推定を行う．入力データとして学習サンプル $\mathbf{v}$ が N 個与えられたとし，それぞれを $(\mathbf{v}^1, \cdots, \mathbf{v}^N)$ とすると，対数尤度 $\mathcal{L}(\theta)$ は次のような式となる．

$$\mathcal{L}(\theta) = \sum_{n=1}^{N} \ln p(\mathbf{v}^n) = \sum_{n=1}^{N} \ln \sum_{\mathbf{h}} p(\mathbf{v}^n, \mathbf{h})$$

この対数尤度 $\mathcal{L}(\theta)$ を用いて，パラメータは $\arg\max_\theta \mathcal{L}(\theta)$ として推定される．

この対数尤度を最大化する事は KL 情報量を最小化する観点から以下に述べるように理解できる．KL 情報量 $\mathcal{D}[p||q] = \sum_x p(x) \ln \frac{p(x)}{q(x)}$ は 2 つの確率分布 $p(x), q(x)$ 間の近さを表している．まず，学習サンプルの集合を次の式で定義される経験分布 $q(\mathbf{v})$ によって生成されたものであるとする．

$$q(\mathbf{v}) \equiv \frac{1}{N} \sum_{n=1}^{N} \prod_{i=1}^{N_v} \delta(v_i, v_i^n)$$

ここで $\delta(x, y)$ はクロネッカーのデルタ関数である．経験分布 $q(\mathbf{v})$ を用いるとパラメータは $\arg\min_\theta \mathcal{D}[q(\mathbf{v})||p(\mathbf{v})]$ のように推定される．

$\mathcal{D}[q(\mathbf{v})||p(\mathbf{v})]$ をパラメータ θ について偏微分すると次の式のようになる．

$$\begin{aligned} \frac{\partial \mathcal{D}}{\partial \theta} &= \sum_{\mathbf{v}} \sum_{\mathbf{h}} \frac{\partial E}{\partial \theta} p(\mathbf{h}|\mathbf{v}) q(\mathbf{v}) - \sum_{\mathbf{v}} \sum_{\mathbf{h}} \frac{\partial E}{\partial \theta} p(\mathbf{v}, \mathbf{h}) \\ &= \left\langle \frac{\partial E}{\partial \theta} \right\rangle_{data} - \left\langle \frac{\partial E}{\partial \theta} \right\rangle_{model} \end{aligned}$$

ここで $\left\langle \frac{\partial E}{\partial \theta} \right\rangle_{data}$ と $\left\langle \frac{\partial E}{\partial \theta} \right\rangle_{model}$ はそれぞれ確率分布 $p(\mathbf{h}|\mathbf{v})q(\mathbf{v})$ と $p(\mathbf{v}, \mathbf{h})$ についての期待値を示す．以上より勾配を計算し，パラメータを逐次的に更新していくことが RBM の学習となる．

2.2.1 Gibbs Sampling

RBM では可視ユニット同士，隠れユニット同士に結合がないため，$\mathbf{h}$ が与えられた時の $\mathbf{v}$ の条件付き確率 $p(v_i = 1|\mathbf{h})$ と $\mathbf{v}$ が与えられた時の $\mathbf{h}$ の条件付き確率 $p(h_j = 1|\mathbf{v})$ は独立性があり，次のように解析的に計算できる．

$$p(v_i = 1|\mathbf{h}) = \sigma\left(b_i + \sum_j W_{ij} h_j\right)$$

$$p(h_j = 1|\mathbf{v}) = \sigma\left(c_j + \sum_i v_i W_{ij}\right)$$

$\sigma(x)$ はシグモイド関数で，$\sigma(x) = \frac{1}{1+\exp(-x)}$ である．$\left\langle \frac{\partial E}{\partial \theta} \right\rangle_{data}$ は positive phase と呼ばれるが，positive phase はこれらの条件付確率を用いて解析的に計算することが可能となる．

しかし，$\left\langle \frac{\partial E}{\partial \theta} \right\rangle_{model}$ は negative phase と呼ばれるが，この negative phase の期待値は $p(\mathbf{v}, \mathbf{h})$ の組について求める必要があるが，その値の組み合わせが膨大であるため，解析的に計算できない．後述する Contrastive Divergence(CD) 法 [10] が提案されるまで，negative phase を

近似的に求める方法として Gibbs Sampling[1] という方法が使われていた．Gibbs Sampling において N 個の学習サンプル $(\mathbf{v}^1, \cdots, \mathbf{v}^N)$ を元に，T 回サンプリングし，更新して得られた $(\mathbf{v}_T^1, \cdots, \mathbf{v}_T^N)$ から確率分布 $q_T(\mathbf{v})$ が次のように求められる．

$$q_T(\mathbf{v}) \equiv \frac{1}{N} \sum_{n=1}^{N} \prod_{i=1}^{N_v} \delta(v_i, \hat{v}_i^n)$$

ここで，$\hat{v}_i^n$ は T 回更新後の可視ユニットのサンプルである．そして，$p(\mathbf{v}, \mathbf{h}) \approx p(\mathbf{h}|\mathbf{v}) q_T(\mathbf{v})$ のように近似して negative phase を求める．

2.2.2 Contrastive Divergence(CD) 法

Gibbs Sampling において，サンプリングを T 回繰り返して得られた確率分布 $q_T(\mathbf{v})$ が $q_T(\mathbf{v}) \approx p(\mathbf{v})$ と近似するには，更新回数 T を比較的大きな値にする必要がある．しかし，T の値を大きくするほど計算量は非常に大きくなり，現実的であるとは言えない．そこで，用いられる方法が Contrastive Divergence(CD) 法 [10] である．CD 法では T の値が比較的小さくて済み，計算量を大きく削減できる．CD 法では前述した KL 情報量 $\mathcal{D}[q(\mathbf{v})||p(\mathbf{v})]$ を最小化するのではなく，次の式を最小化する．

$$\mathcal{D}[q(\mathbf{v})||p(\mathbf{v})] - \mathcal{D}[q_T(\mathbf{v})||p(\mathbf{v})]$$

その際，少ない T でも十分に良い精度が得られることが経験的に知られている．特に $T = 1$ であるような場合を CD-1 法と呼ばれ，本研究ではこの CD-1 法を利用した．

2.2.3 パラメータの更新

本研究では，CD-1 法を用いて得られた勾配を用いてパラメータを更新する．パラメータの更新方法の 1 つに確率的勾配降下法があり，ディープラーニングにおいて並列計算の効率の点でこの方法が良く使われる [10]．本研究ではこの方法を用いる．N 個の学習サンプルのそれぞれの勾配の総和を Δ_N とする時，確率勾配降下法による更新式は以下のようになる．

$$\theta^t = \theta^{t-1} - \eta \Delta_N$$

ここでパラメータ θ は $\theta = (\mathbf{b}, \mathbf{c}, \mathbf{W})$ であり，t は更新回数，η は学習率である．学習の精度を上げるためには，この学習率 η を適切なスケジュールで決定する必要がある．

2.2 Gaussian Restricted Boltzman Machine (GRBM)

画像データや音声データなど，現実世界における実際のデータの多くは連続値で表現され，2 値では表現されていない．そこで RBM における可視ユニットをガウシアンノイズを持つ線形ユニットに置き換える．置き換えたモデルは Gaussian Restricted Boltzman Machine(GRBM)[7] と呼ばれ，エネルギー関数は以下の式で与えられる．

$$E(\mathbf{v}, \mathbf{h}) = \sum_{i=1}^{N_v} \frac{(v_i - b_i)^2}{2\sigma_i^2} - \sum_{j=1}^{N_h} c_j h_j - \sum_{i,j} \frac{v_i}{W_{ij}} h_j$$

σ_i は可視ユニット v_i に対するガウシアンノイズの標準偏差である．これは多変量正規分布において，分散共分散行列を対角要素が σ_i^2 で，非対角要素が 0 であるような対角行列とであると仮定した時の分布に，各ユニットに入力されたデータが従っているとして学習している．この場合条件付き確率 $p(v_i|\mathbf{h})$ と $p(h_j = 1|\mathbf{v})$ は次のようになる．

$$p(v_i|\mathbf{h}) = \mathcal{N}\left(v_i; b_i + \sum_j W_{ij} h_j, 1\right)$$

$$p(h_j = 1|\mathbf{v}) = \sigma\left(c_j + \sum_i v_i W_{ij}\right)$$

ここで $\mathcal{N}\left(v_i; b_i + \sum_j W_{ij} h_j, 1\right)$ は正規分布関数であり，v_i が正規分布 $\mathcal{N}\left(b_i + \sum_j W_{ij} h_j, 1\right)$ に従うという事である．

GRBM における学習の流れは尤度を表す条件付き確率が正規分布に基づく形に置き換わる事を除いて，前節で述べた通常の RBM と同等である．しかし，仮定しているモデルが RBM と GRBM では異なるため，得られたパラメータは本質的に異なる．

2.3 Replicated Softmax Model

Replicated Softmax Model[8] は文書を文書内の単語の語順に関係なく単語の出現回数で表現した Bag of Words などのスパースな離散データをモデリングするのに役立つことが知られている．Replicated Softmax Model のグラフィカルモデルは RBM と同様である．語彙数が K と与えられ，単語 k の出現回数を v_k とし，文書を $\mathbf{v} = (v_1, \cdots, v_K)$ で表したデータを Replicated Softmax Model の入力データとして扱うとする．潜在ユニットの数を F，文書内の単語数を D とすると，Replicated Softmax Model のエネルギー関数は次のように与えら

れる．

$$E(\mathbf{v},\mathbf{h}) = -\sum_{k=1}^{K} b_k v_k - D\sum_{j=1}^{F} c_j h_j - \sum_{k=1}^{K}\sum_{j=1}^{F} v_k W_{kj} h_j$$

ここで，$\theta = (\mathbf{b},\mathbf{c},\mathbf{W})$ はパラメータである．そして，条件付き確率は次のように与えられる．

$$p(v_k = 1|\mathbf{h}) = \frac{\exp\left(b_k + \sum_{j=1}^{F} h_j W_{kj}\right)}{\sum_{q=1}^{K} \exp\left(b_q + \sum_{j=1}^{F} h_j W_{qj}\right)}$$

$$p(h_j = 1|\mathbf{v}) = \sigma\left(c_j + D\sum_i v_k W_{kj}\right)$$

Replicated Softmax Model の可視ユニットはソフトマックスユニットと呼ばれ，順序性を持たない K 個の状態の値に対してサンプリングする際に適当である．可視ユニットに関して，RBM では条件付き確率を用いて 1 回サンプリングしていたのに対し，Replicated Softmax Model では条件付き確率 $p(v_k = 1|\mathbf{h})$ を用いて D 回サンプリングされる．すなわち，RBM の各ユニットが 0 か 1 のどちらかの値しかとらなかったのとは違い，K 個のユニットの各状態は 1 より大きい値を取ることが出来るが，その合計は D でなければならない．Replicated Softmax Model における学習の流れは RBM と同様である．

2.4 Deep Belief Network (DBN)

Deep Belief Network(DBN)[4,20] は可視層の上に階層的に隠れ層を積み重ねていく事で構成される．DBN のグラフィカルモデルは次のようなものである．

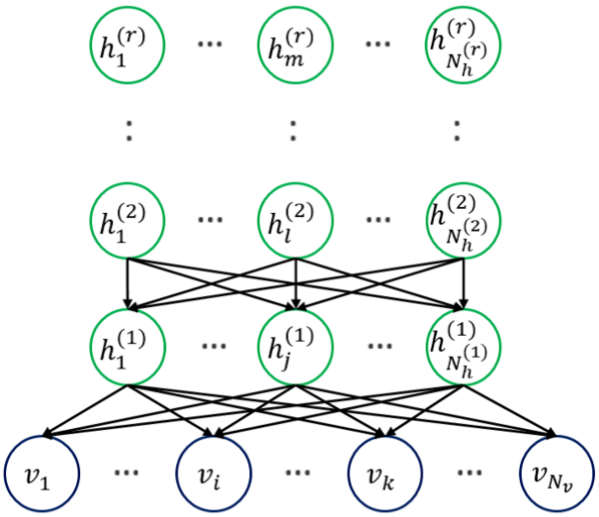

Fig 2: Graphical model of DBN.

ここで DBN について RBM で階層的に隠れ層を積み重ねていった場合を例に説明する．r 番目の隠れ層を $\mathbf{h}^{(r)}$，隠れユニット数を $N_h^{(r)}$，バイアスを $c_j^{(r)}$，$r-1$ 番目の隠れ層と r 番目の隠れ層の間の結合の重みを $\mathbf{W}^{(r)}$ で表すとすると，隠れ層が R 層ある DBN エネルギーは次のようになる．

$$E(\mathbf{v},\mathbf{h}) = -\sum_{i=1}^{N_v} b_i v_i - \sum_{i=1}^{N_v}\sum_{j=1}^{N_h^{(1)}} v_i W_{ij}^{(1)} h_j^{(1)} - \sum_{r=1}^{R}\sum_{j=1}^{N_h^{(r)}} c_j^{(r)} h_j^{(r)} - \sum_{r=2}^{R}\sum_{i=1}^{N_h^{(r-1)}}\sum_{j=1}^{N_h^{(r)}} h_i^{(r-1)} W_{ij}^{(r)} h_j^{(r)}$$

DBN も原理的には RBM と同様にこのエネルギーから学習を行うが，計算量が膨大であるため厳密な学習は望めない．そこでいくつかの近似学習の方法が必要となる．近似学習の基本戦略の一つとして貪欲学習 [4，12] がある．貪欲学習の概要としては，まず可視層と 1 層目の隠れ層について RBM を用いる．その際，2 層目以降の隠れ層は考えない．次に，可視層と 1 層目の隠れ層の間のパラメータは固定したままにし，1 層目の隠れ層を擬似的な可視層とみなして 1 層目の隠れ層と 2 層目の隠れ層について，その他の層は考えず RBM を用いる．この際，1 層目の隠れ層に入力されるデータには条件付き確率 $p(h_j^{(1)} = 1|\mathbf{v})$ からサンプリングされたデータを利用する．3 層目以降の隠れ層に対しても同様の方法で RBM を用いて学習していく事を繰り返すことによって DBN が学習される．実際の DBN では，1 層目と 2 層目を GRBM で学習し，2 層目と 3 層目を RBM で学習するなどして変更が加わる事がある．このように教師なし学習を行う過程を事前学習と呼ぶ．

2.5 ファインチューニング

ファインチューニングを行う前に，前節で述べた事前学習を行った結果として得られたパラメータを利用し，条件付き確率 $p(\mathbf{h}|\mathbf{v})$ などを用いて，入力データを作成したモデルの隠れ層の最上位層の表現に変換する．本研究の目的である画像のカテゴリ認識を実現するため，多クラス分類を行う必要がある．多クラス分類する際，隠れ層の最上位層の表現を用いて多変量ロジスティック回帰 [13] を適用する．

多クラス分類におけるファインチューニングは，事前学習を行った結果として得られたパラメータを利用して求めたモデルの隠れ層の最上位層の表現を説明変数として，多変量ロジスティック回帰 [13] を用いて学習を行い，その予測誤差が最小となるように各種パラメータを更新する過程である．

3. ディープラーニングの画像の局所特徴量を用いた表現への適用

先ほど述べた文書を対象として提案されている Replicated Softmax Model[8] の入力データに画像の局所特徴量を用いた表現を適用させるためには，局所特徴量を用いて画像をベクトル量子化し，Bag of Visual Words の表現を得る必要がある．本章では，画像から Bag of Visual Words の表現を用いて Replicated Softmax Model の入力データを得る過程について述べる．そして，最後に画像の複数表現がディープラーニングにどのように適用されるのかについて述べる．

3.1 局所特徴量の抽出

局所特徴記述子には様々な種類があるが，本研究では SIFT[16] を用いた．SIFT は画像のスケール変化や回転に強く，照明変化に対しても比較的頑強であり，広く受け入れられている．SIFT には画像の特徴点を抽出するアルゴリズムがあり，画像中の特徴変化の大きい点を特徴点として抽出することが可能である．しかし，画像のカテゴリ認識などの問題において，特徴点の抽出にはしばしば dense sampling という方法が用いられ，有効な手段である事が示されている [14]．この dense sampling とは等幅の格子を仮定し，その交点を特徴点と見なす方法である．本研究において特徴点を dense sampling を用いて抽出している．そして，抽出された特徴点において，SIFT 記述子を計算する．

3.2 Bag of Visual Words

Bag of Visual Words はカテゴリ認識において広く普及している画像特徴量表現で，情報検索や自然言語処理で使われる Bag of Words のアナロジーである．単語の出現回数を考える Bag of Words と同様に，Bag of Visual Words は画像の局所特徴をヒストグラムにしたもので表現する [14，15]．これにより画像の局所特徴量を文書内の単語に対応させ，画像を文書のようなスパースな離散表現として扱うことが可能になる．

3.3 ディープラーニングにおける画像の複数表現の適用

本研究において，生データをディープラーニングに適用させる場合，生データは実数の連続値なので，前章で述べた Gaussian Restricted Boltzman Machine(GRBM)[7] の入力として直接扱って学習し，次に GRBM の隠れ層の表現を前章で述べた Restricted Boltzman Machine(RBM)[6,20] の入力として扱って学習するという手順を踏む．また，本研究で局所特徴量表現としている前節までに述べた Bag of Visual Words の表現をディープラーニングに適用させる場合，Bag of Visual Words の表現はスパースな離散表現であるため，前章で述べた Replicated Softmax Model の入力として扱って学習し，次に Replicated Softmax Model の隠れ層の表現を RBM に入力として扱って学習するという手順を踏む．さらに，これらの学習から得られた潜在特徴表現を RBM で多層的に統合したモデルが画像の複数表現に基づくマルチモーダルな DBN[5] である．本研究では，マルチモーダルな DBN における特徴表現に関する比較検討を行うことを目的とし，その点が [5] とは異なる．

4. 実験

本章では，生データと局所特徴量表現それぞれについて潜在的特徴を抽出した後にその特徴を統合させて扱った時の認識精度について検証するための実験に関して述べる．検証は入力データに生データ，局所特徴量表現，またその両者を用いてディープラーニングで学習し，結果を比較する事で行う．尚，本研究でのディープラーニングによる学習は事前学習を行った後，得られたパラメータを用いてロジスティック回帰を行うことで予測を行うという過程で行われる．

4.1 データセット

本研究では CIFAR-10 をデータセットとして使用する．CIFAR-10 は 32 × 32 ピクセルの画像 6 万枚からなるデータセットである．各画像には飛行機や鳥など 10 種類の物体の内 1 つが写っており，どの物体が写っているのかがクラス番号として与えられている．1 つのクラスについて画像は 6000 枚用意されており，訓練データとして 5000 枚，テストデータとして 1000 枚使用する．すなわち，合計で訓練データは 5 万枚，テストデータは 1 万枚となる．クラスと物体の対応関係を以下の表に示す．

Table 1: Classes and the corresponding objects.

class	object
0	airplane
1	automobile
2	bird
3	cat
4	deer
5	dog
6	frog
7	horse
8	ship
9	truck

CIFAR-10 の画像を生データとして扱う時，ディープラーニングにおいて画像の明るさを正規化する事で学習の性能が上がることが知られているので，前処理として Global Contrast Normalization(GCN)[17] と ZCA Whitening[18] を行い画像を正規化する．本研究では CIFAR-10 の画像データに GCN と ZCA Whitening を行ったものを生データとして使用した．

CIFAR-10 の画像から局所特徴量表現を取り出す際，まず格子幅を 2 × 2 ピクセルとし，局所特徴記述子のスケールを 2，4，8 ピクセルの内ランダムに選択し dense sampling を行った．ここで，局所特徴記述子は SIFT である．格子幅を 2 × 2 ピクセルとしているので，dense sampling によって 32 × 32 ピクセルの各画像から特徴点が 256 個取れる．CIFAR-10 の 6 万枚の画像全てに dense sampling を行った後，256 × 6 万個の特徴点から局所特徴量が得られ，これをベクトル量子化するために K-means アルゴリズムを用いる．K-means において K=1000 とした．1000 個の Visual Word に各局所特徴量が置き換えられ，各画像について 256 個の特徴点が得られているので，256 回置き換えられた Visual Word に投票したヒストグラムを局所特徴量表現として扱う．

4.2 カテゴリ認識

まず事前学習における手順について述べる．生データを用いる際は，CIFAR-10 の訓練データに GCN[17] と ZCA Whitening[18] を行ったものを入力とする．入力が連続値のため，GRBM を用いてまず学習する．GRBM において可視ユニットの数はピクセルの数 (32 × 32) と画像の色情報である RGB 値に GCN と ZCA Whitening を行った 3 値を掛け合わせた 3072 個である．隠れユニットの数は 1000 個とし，学習率を 0.001 とした．そして，次に GRBM の隠れ層の表現を入力とし，RBM で学習する．ここで，可視ユニットの数は GRBM の隠れユニットの数と同じ 1000 個であり，RBM の隠れユニットの数は 1000 個，学習率は 0.001 とした．以上より 1 層目と 2 層目の間が GRBM で，2 層目と 3 層目の間が RBM で構成される DBN により学習したと解釈することが可能である．この DBN を以下に図示する．

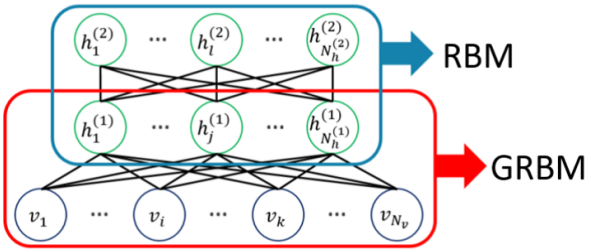

Fig 3: DBN with raw data.

局所特徴量表現を用いる際は，CIFAR-10 の訓練データを Bag of Visual Words の表現にしたもの，すなわち訓練データを局所特徴量表現にしたものを入力とする．Bag of Visual Words の表現に変換したことにより，画像を文書として扱えるようになったので，事前学習において文書データに対して提案されている Replicated Softmax Model を使用し，学習する．この Replicated Softmax Model において，単語数は各画像の特徴点の数である 256 個であり，K-means において K=1000 としているので可視ユニットの数は 1000 個である．隠れユニットの数は 500 個とし，学習率は 0.001 とした．そして，次に Replicated Softmax Model における隠れ層の表現を入力とし，RBM で学習する．ここで，可視ユニットの数は Replicated Softmax Model の隠れユニットの数と同じ 500 個であり，隠れユニットの数は 1000 個，学習率は 0.001 とした．以上より，1 層目と 2 層目の間が Replicated Softmax Model で，2 層目と 3 層目の間が RBM で構成される DBN により学習したと解釈することが可能である．この DBN を以下に図示する．

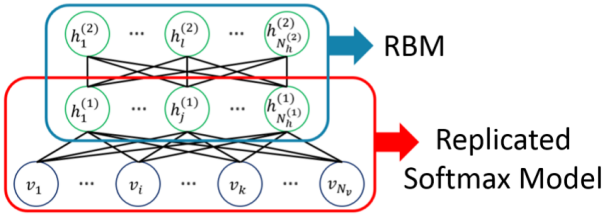

Fig 4: DBN with local feature representations.

マルチモーダルな DBN によるカテゴリ認識を行う際は，生データ，局所特徴量表現を用いて事前学習を行う際に用いたそれぞれの DBN の最上位層の表現を取り出し，それらを組み合わせた表現を入力として RBM を行う．これは一方を生データを入力としたモード，もう一方を局所特徴量表現を入力としたモードとした 2 つのモードを持つマルチモーダルな DBN による学習と考える事が出来る．RBM において，可視層は局所特徴量表現の際に用いた DBN の最上位層のユニット 1000 個と生データの際に用いた DBN の最上位層のユニット 1000 個を結合させたものとなっており，可視ユニットの数は 2000 個となる．そして，RBM の隠れユニットの数を 1000 個，学習率を 0.001 とした．このマルチモーダルな DBN を以下に図示する．

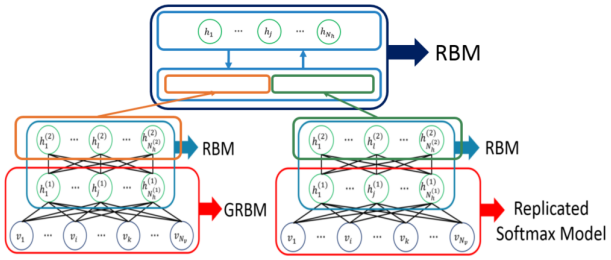

Fig 5: Multimodal DBN.

以上のモデルを用いて事前学習した後に，それぞれにおいて，テストデータを用いてファインチューニングをすることでカテゴリ認識を行う．カテゴリ認識の認識精度は F 値を用いて評価する．以下に，それぞれのモデルの各クラスにおける認識精度の検証閣下を F 値で評価した表を示す．

Table 2: F-measure with various DBNs.

class	raw data	local feature representations	multimodal
0	0.495	0.418	0.443
1	0.501	0.286	0.406
2	0.319	0.224	0.323
3	0.329	0.221	0.283
4	0.355	0.294	0.334
5	0.368	0.295	0.358
6	0.487	0.371	0.443
7	0.476	0.249	0.376
8	0.591	0.453	0.532
9	0.463	0.343	0.365

以上より，生データを入力とした DBN，特徴量表現を入力とした DBN，マルチモーダルな DBN のそれぞれを用いて学習を行った結果，F 値の平均は 0.4384，0.3154，0.3863 となった．

5. まとめ

本研究では，ディープラーニングで学習を行う事により，ラベルが与えられていない様々な表現を持つデータのラベルを予測する事が可能であると示した．実験において，入力に生データを用いた時の F 値の平均 (0.4384) が局所特徴量表現を用いた時の F 値の平均 (0.3154) を約 39 %も上回った事が確認されたが，局所特徴量表現にも一定の有用性はある事が確認された．しかし，ディープラーニングでカテゴリ認識を行う場合は，生データを直接用いて学習する事が有効であると考えられる．マルチモーダルにした時の F 値の平均が 0.3863 であり，生データのみを用いた場合に比べて下がってしまったのは，生データを用いた場合と局所特徴量表現を用いた場合でそれぞれ認識しやすいクラスが異なるため，マルチモーダルにした時に互いに悪影響が出たのではないかと考えられる．どのデータを用いても最も認識しやすかった画像は ship のクラスが与えられたもので，画像全体の配色や写る物体の形状が似ており，特徴的なものであったので，学習しやすいものとしにくいものがある事が確認された．

今後の展望としては，ディープラーニングの他に機械学習の分野で注目されているトピックモデルの入力として，ディープラーニングで学習したデータの潜在的な特徴表現を扱う方法 [19] について検討することなどが挙げられる．

参考文献

[1] Bengio, Yoshua. ”Learning deep architectures for AI.” Foundations and trendsR in Machine Learning 2.1 (2009): 1-127.

[2] Le, Quoc V. ”Building high-level features using large scale unsupervised learning.” Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013.

[3] Burges, Christopher JC. ”A tutorial on support vector machines for pattern recognition.” Data mining and knowledge discovery 2.2 (1998): 121-167.

[4] Bengio, Yoshua, et al. "Greedy layer-wise training of deep networks." Advances in neural information processing systems 19 (2007): 153.

[5] Srivastava, Nitish, and Ruslan R. Salakhutdinov. "Multimodal learning with deep boltzmann machines." Advances in neural information processing systems. 2012.

[6] Hinton, Geoffrey E. "A practical guide to training restricted boltzmann machines." Neural Networks: Tricks of the Trade. Springer Berlin Heidelberg, 2012. 599-619.

[7] Cho, KyungHyun, Alexander Ilin, and Tapani Raiko. "Improved learning of Gaussian-Bernoulli restricted Boltzmann machines." Artificial Neural Networks and Machine Learning?ICANN 2011. Springer Berlin Heidelberg, 2011. 10-17.

[8] Hinton, Geoffrey E., and Ruslan R. Salakhutdinov. "Replicated softmax: an undirected topic model." Advances in neural information processing systems. 2009.

[9] Erhan, Dumitru, et al. "Why does unsupervised pre-training help deep learning?." The Journal of Machine Learning Research 11 (2010): 625-660.

[10] Carreira-Perpinan, Miguel A., and Geoffrey E. Hinton. "On contrastive divergence learning." Proceedings of the tenth international workshop on artificial intelligence and statistics. NP: Society for Artificial Intelligence and Statistics, 2005.

[11] Ngiam, Jiquan, et al. "On optimization methods for deep learning." Proceedings of the 28th International Conference on Machine Learning (ICML-11). 2011.

[12] Hinton, Geoffrey, Simon Osindero, and Yee-Whye Teh. "A fast learning algorithm for deep belief nets." Neural computation 18.7 (2006): 1527-1554.

[13] Cawley, Gavin C., Nicola LC Talbot, and Mark Girolami. "Sparse multinomial logistic regression via bayesian l1 regularisation." Advances in neural information processing systems 19 (2007): 209.

[14] Fei-Fei, Li, and Pietro Perona. "A bayesian hierarchical model for learning natural scene categories." Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. Vol. 2. IEEE, 2005.

[15] Csurka, Gabriella, et al. "Visual categorization with bags of keypoints." Workshop on statistical learning in computer vision, ECCV. Vol. 1. No. 1-22. 2004.

[16] Lowe, David G. "Distinctive image features from scale-invariant keypoints." International journal of computer vision 60.2 (2004): 91-110.

[17] Hansen, Bruce C., and Edward A. Essock. "Anisotropic local contrast normalization: The role of stimulus orientation and spatial frequency bandwidths in the oblique and horizontal effect perceptual anisotropies." Vision research 46.26 (2006): 4398-4415.

[18] Krizhevsky, Alex, and Geoffrey Hinton. "Learning multiple layers of features from tiny images." Computer Science Department, University of Toronto, Tech. Rep 1.4 (2009): 7.

[19] Wan, Li, Leo Zhu, and Rob Fergus. "A hybrid neural network-latent topic model." International Conference on Artificial Intelligence and Statistics. 2012.

[20] 八木康史, and 斎藤英雄. "コンピュータビジョン最先端ガイド 6." アドコム・メディア，12 (2013): 89-121.