# Network-in-Network

# Abstract

我々は、受容野内の局所的なパッチに対するモデルの識別性を高めるために、"Network In Network"(NIN)と呼ばれる新しい深層ネットワーク構造を提案する。従来の畳み込み層では、線形フィルタと非線形活性化関数を用いて入力をスキャンしていました。その代わりに、より複雑な構造を持つマイクロニューラルネットワークを構築し、受容野内のデータを抽象化します。マイクロ・ニューラル・ネットワークのインスタンス化には，潜在的な関数近似である多層パーセプトロンを用いる．特徴量マップは、CNNと同様にマイクロネットワークを入力上でスライドさせることで得られ、次の層に供給されます。Deep NINは、上記のような構造を複数重ねることで実現できます。マイクロネットワークを用いて局所的なモデリングを強化することで、分類層では特徴マップの大域的な平均プーリングを利用することができ、従来の完全連結層に比べて解釈しやすく、オーバーフィッティングの可能性も低くなります。我々は、CIFAR-10およびCIFAR-100において、NINによる最先端の分類性能を実証し、SVHNおよびMNISTデータセットにおいても妥当な性能を示した。

# 1. 緒言

畳み込みニューラルネットワーク（CNN）[1]は，畳み込み層とプーリング層を交互に配置した構成になっている．**畳み込み層は，線形フィルタと下層の受容野の内積をとり，それに続いて入力の局所部分ごとに非線形活性化関数をとります．結果として得られる出力は特徴マップと呼ばれる。**

CNNの畳み込みフィルターは、基礎となるデータパッチに対する一般化線形モデル（GLM）であり、GLMでは抽象度が低いと主張している。抽象度とは、その特徴が同じ概念のバリエーションに対して不変であることを意味します[2]。GLMをより強力な非線形関数近似器に置き換えることで、局所モデルの抽象化能力を高めることができます。GLMは、潜在的な概念のサンプルが線形分離可能な場合、つまり、概念のバリエーションがGLMで定義された分離平面の片側にすべて存在する場合に、良好な抽象化を達成することができます。したがって、従来のCNNは暗黙のうちに潜在的な概念が線形分離可能であると仮定している。しかし、同じ概念のデータは非線形多様体上に存在することが多く、したがって、これらの概念を捉える表現は一般的に入力の高度な非線形関数となります。NINでは、GLMを一般的な非線形関数近似器である「マイクロネットワーク」構造に置き換えます。本研究では，マイクロネットワークのインスタンスとして多層パーセプトロン[3]を選択した。多層パーセプトロンは，普遍的な関数近似器であり，バックプロパゲーションによって学習可能なニューラルネットワークである。

その結果、mlpconv層と呼んでいる構造をCNNと比較したのが図1である。線形畳み込み層もmlpconv層も、局所受容野を出力特徴ベクトルにマッピングします。mlpconv層は、非線形活性化関数を持つ複数の完全連結層からなる多層パーセプトロン（MLP）を用いて、入力された局所パッチを出力特徴ベクトルにマッピングする。このMLPは，すべての局所受容野で共有されます。特徴量マップは、CNNと同様にMLPを入力上でスライドさせることで得られ、次の層に供給されます。NINの全体的な構造は、複数のMLPconv層を積み重ねたものです。深層ネットワーク全体を構成する要素であるマイクロネットワーク（MLP）をmlpconv層の中に入れていることから、「Network In Network」（NIN）と呼ばれています。

CNNの分類に従来の完全連結層を採用する代わりに、最後のmlpconv層からの特徴マップの空間平均を、グローバルアベレージプーリング層を経由してカテゴリーの信頼度として直接出力し、その結果のベクトルを

ソフトマックス層に入力する。**従来のCNNでは、完全連結層がブラックボックスのように機能するため、目的コスト層からのカテゴリーレベルの情報がどのように前の畳み込み層に戻されるのかを解釈することが困難でした。**対照的に、グローバルアベレージプーリングは、マイクロネットワークを使ったより強力なローカルモデリングによって可能になった、特徴マップとカテゴリーの間の対応関係を強制するので、より意味があり解釈しやすい。さらに、**完全連結層はオーバーフィッティングの傾向があり、ドロップアウト正則化に大きく依存していますが[4][5]、グローバルアベレージプーリングはそれ自体が構造的正則化であり、全体の構造に対するオーバーフィッティングを本質的に防止します。**

# 2. 畳み込み式ニューラルネットワーク

古典的な畳み込みニューロンネットワーク [1] は，交互に積み重ねられた畳み込み層と空間プーリング層から構成されています．畳み込み層では，線形畳み込みフィルタに非線形活性化関数（整流器，シグモイド，tanhなど）を適用して，特徴量マップを生成します．線形整流器を例にとると、特徴マップは次のように計算できる。

f_{i, j, k} = \max(w_{k}^T x_{i, j}, 0)

ここで、$(i;j)$はフィーチャーマップのピクセルインデックス、$x_{ij}$は位置$(i;j)$を中心とした入力パッチを表し、$k$はフィーチャーマップのチャンネルのインデックスとして使用されます。

潜在的な概念のインスタンスが線形分離可能な場合、この線形畳み込みは抽象化には十分である。しかし、優れた抽象化を実現する表現は、一般に入力データの高度な非線形関数である。従来のCNNでは、潜在的な概念のすべてのバリエーションをカバーするために、フィルタの過剰なセットを利用することでこの問題を解決していた[6]。つまり、同じ概念の異なるバリエーションを検出するために、個々の線形フィルタを学習することができる。しかし、1つの概念に対してあまりにも多くのフィルタを使用すると、前の層からのバリエーションのすべての組み合わせを考慮する必要がある次の層に余分な負担をかけることになる[7]。CNNのように，上位層のフィルタは元の入力のより大きな領域にマッピングされる．下の層の低レベルの概念を組み合わせることで、より高いレベルの概念を生成します。そのため、より高いレベルの概念に結合する前に、各ローカルパッチに対してより良い抽象化を行うことが有益であると主張する。

最近のmaxoutネットワーク[8]では，アフィン特徴マップに対する最大プーリングによって特徴マップの数を減らしている（アフィン特徴マップとは，活性化関数を適用しない線形畳み込みの直接の結果である）．一次関数に対する最大化により，任意の凸関数を近似できる区分的線形近似器が得られる．線形分離を行う従来の畳み込み層と比較して、maxoutネットワークは、凸集合内にある概念を分離することができるため、より強力である。この改良により、いくつかのベンチマークデータセットにおいて、maxoutネットワークは最高の性能を発揮している。

しかし、maxoutネットワークは、潜在概念のインスタンスが入力空間の凸集合内に存在するという事前条件を課しているが、これは必ずしも成り立たない。潜在的な概念の分布がより複雑な場合には、より一般的な関数近似を採用する必要があります。そこで我々は、各畳み込み層の中にマイクロネットワークを導入し、ローカルパッチのより抽象的な特徴を計算する「ネットワーク・イン・ネットワーク」という新しい構造を導入することで、これを実現しようとしている。

マイクロネットワークを入力に対してスライドさせる手法は，いくつかの先行研究で提案されている．例えば，SMLP（Structured Multilayer Perceptron）[9]では，入力画像の異なるパッチに共有の多層パーセプトロンを適用し，別の作品では，顔検出のためにニューラルネットワークベースのフィルタを学習している[10]．しかし，これらはいずれも特定の問題のために設計されたものであり，スライディングネットワーク

構造の1つの層しか含まれていません．NINは、より一般的な観点から提案されており、マイクロネットワークをCNN構造に統合し、あらゆるレベルの特徴をより良く抽象化することを追求しています。

# 3. Network in Network

まず、提案する「ネットワーク・イン・ネットワーク」構造の主要構成要素であるMLP畳み込み層とグローバルアベレージングプーリング層について、それぞれ3.1項と3.2項で説明します。続いて、NINの全体像を説明します（3.3項）。

## 3.1. MPL 畳み込み層

線形畳み込み層とmlpconv層の比較。線形畳み込み層は線形フィルタを含み、mlpconv層はマイクロネットワーク（本稿では多層パーセプトロンを選択）を含む。どちらの層も局所的な受容野を潜在的な概念の信頼値にマッピングする。

潜在的な概念の分布に関する事前情報がない場合、潜在的な概念のより抽象的な表現を近似することができる普遍的な関数近似器を局所パッチの特徴抽出に用いることが望ましい。普遍関数近似器としては、放射状基底ネットワークや多層パーセプトロンがよく知られている。本研究では、2つの理由から多層パーセプトロンを選択した。まず、多層パーセプトロンは、バックプロパゲーションを用いて学習される畳み込みニューラルネットワークの構造と互換性がある。第二に、多層パーセプトロンはそれ自体が深層モデルとなりうるため、特徴再利用の精神に合致しています[2]。この新しいタイプの層は、本稿ではmlpconvと呼ばれており、MLPがGLMに代わって入力に対して畳み込みを行うものです。図1は、線形畳み込み層とmlpconv層の違いを示しています。mlpconv層で行われる計算は以下のようになります。

f^1_{i, j, k_1} = \max({w^1_{k_1}}^Tx_{i, j}+b_{k_1}, 0) \

f^n_{i, j, k_n} = \max({w^n_{k_n}}^Tx_{i, j}+b_{k_n}, 0)

ここで，nは多層パーセプトロンの層数である。多層パーセプトロンの活性化関数には、整流線形ユニットを使用しています。

クロスチャンネル（クロスフィーチャーマップ）プーリングの観点から見ると、式2は通常の畳み込み層にカスケード接続されたクロスチャンネルパラメトリックプーリングと同等である。各プーリング層は、入力された特徴マップに対して重み付き線形再結合を行い、その後、整流器線形ユニットを通過する。クロスチャンネルプーリングされた特徴マップは、次の層で何度も何度もクロスチャンネルプーリングされます。このカスケード接続されたクロスチャネル・パラメータ・プーリング構造により、クロスチャネル情報の複雑で学習可能な相互作用が可能になります。

**クロスチャンネルパラメトリックプーリング層は、1x1畳み込みフィルタを持つ畳み込み層に相当します。**このように解釈することで、NINの構造をわかりやすく理解することができます。

**maxout層との比較**：maxoutネットワークのmaxout層は、複数のアフィンフィーチャーマップに対してmaxプーリングを行う[8]。maxout層の特徴量マップは以下のように計算される。

$$f_{i, j, k}=\max_{m}(w^T_{k_m}x_{i, j})$$

一次関数のMaxoutは、任意の凸関数をモデル化できる区分的一次関数を形成します。凸関数の場合、特定の閾値以下の関数値を持つサンプルは、凸セットを形成します。したがって、ローカルパッチの凸関数を近似することで、maxoutは、サンプルが凸集合内にあるコンセプトの分離超平面を形成する機能を持っています（例：l2ボール、凸円錐）。Mlpconv層は、maxout層とは異なり、凸関数近似器がユニバーサル関数近似器に置き換えられており、潜在的な概念の様々な分布をモデル化する能力が高くなっています。

## 3.2. Global Average Pooling

従来の畳み込みニューラルネットワークは，ネットワークの下位層で畳み込みを行います．**分類のためには，最後の畳み込み層の特徴マップがベクトル化され，完全連結層とソフトマックス・ロジスティック回帰層に供給されます [4] [8] [11]．この構造は，畳み込み構造と従来のニューラルネットワーク分類器をつなぐものである．この構造では，畳み込み層を特徴抽出器として扱い，その結果得られた特徴を従来の方法で分類します．**

しかし，完全連結層ではオーバーフィッティングが起こりやすく，ネットワーク全体の汎化能力が低下する．Hintonら[5]は，学習時に完全連結層の活性度の半分をランダムにゼロにするレギュラライザとしてDropoutを提案した．これにより，汎化能力が向上し 汎化能力を向上させ，オーバーフィッティングを大幅に防止している[4].

本論文では、CNNにおける従来の完全連結層に代わるものとして、グローバルアベレージプーリングと呼ばれる別の戦略を提案する。**そのアイデアは、最後のmlpconv層で、分類タスクの対応するカテゴリごとに1つの特徴マップを生成することです。特徴マップの上に完全連結層を追加するのではなく、各特徴マップの平均値を取り、その結果のベクトルを直接ソフトマックス層に入力します。完全連結層と比較した場合のグローバルアベレージプーリングの利点は、特徴マップとカテゴリの間に対応関係を持たせることで、よりコンボリューション構造に近いものになることです。したがって、特徴量マップは、カテゴリ信頼度マップとして簡単に解釈できます。もう1つの利点は、グローバルアベレージプーリングでは最適化するパラメータがないため、この層でのオーバーフィッティングが回避されることです。さらに、グローバル・アベレージ・プーリングは、空間的な情報をまとめているため、入力の空間的な変換に対してより頑健である。**

グローバルアベレージプーリングは、特徴量マップが概念（カテゴリー）の信頼マップであることを明示的に強制する構造正則化として見ることができます。これは、GLMよりも信頼マップの近似を行うmlpconv層によって可能になります。

## 3.3. Network In Network の構造

---

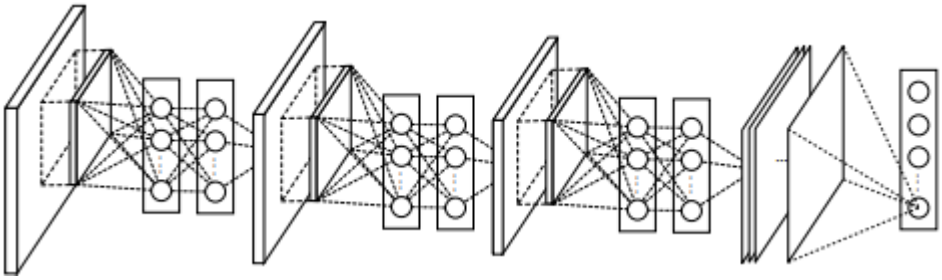

Fig2. Network In Networkの全体的な構造です。本論文では、NINには3つのMLPCONVレイヤーと1つのグローバルアベレージプーリングレイヤーの積層が含まれている。

---

NINの全体的な構造は、mlpconvレイヤーのスタックで、その上にグローバルアベレージプーリングと目的コストのレイヤーがあります。CNNやmaxoutネットワークのように、mlpconv層の間にサブサンプリング層を追加することもできます。図2は、3つのmlpconv層を持つNINを示しています。各mlpconv層の中には、3層のパーセプトロンがあります。NINもマイクロネットワークも、層の数は柔軟で、特定のタスクに合わせて調整することができます。

# 4. 実験の様子

## 4.1. 概要

NINを4つのベンチマークデータセットで評価した．CIFAR-10 [12]，CIFAR-100 [12]，SVHN [13]，MNIST [1]の4つのベンチマークデータセットでNINを評価しました．これらのデータセットに用いたネットワークは，3層のmlpconv層を積み重ねたものであり，すべての実験において，mlpconv層の後には，入力画像を2倍にダウンサンプリングするspatial max pooling層が続く．レギュラライザーとして，最後のmlpconv層を除くすべての層の出力にドロップアウトを適用した．特に断りのない限り，実験で使用したすべてのネットワークは，ネットワークの最上位にある完全連結層の代わりに大域平均プーリングを使用しています．また、Krizhevskyら[4]が使用した重み減衰も適用されています。図2は，本節で使用するNINネットワークの全体構造を示したものである．各パラメータの詳細な設定は補足資料に記載しています。Alex Krizhevsky[4]が開発した超高速cuda-convnetコードを用いてネットワークを実装しました。データセットの前処理，トレーニングセットと検証セットの分割は，すべてGoodfellowら[8]に従った．

Krizhevskyら[4]が用いた学習手順を採用しました。すなわち，重みと学習率の適切な初期化を手動で設定します．ネットワークの学習には，サイズ128のミニバッチを使用します．学習プロセスは，初期の重みと学習率から始まり，学習セットの精度が向上しなくなるまで続け，その後，学習率を10段階で下げていく．この手順を1回繰り返し，最終的な学習率が初期値の1%になるようにする．

## 4.2. CIFAR-10

CIFAR-10データセット[12]は，10クラスの自然画像から構成されており，合計50,000枚のトレーニング画像と10,000枚のテスト画像がある．各画像は，サイズが32x32のRGB画像である．このデータセットには，Goodfellowらがmaxoutネットワーク[8]で使用したのと同じグローバルコントラスト正規化とZCAホワイトニングを適用している．トレーニングセットの最後の10,000枚の画像を検証データとして使用した．

---

| Method | Test Error |
|---|---|
| Stochastic Pooling [11] | 15.13% |
| CNN + Spearmint [14] | 14.98% |
| Conv. maxout + Dropout [8] | 11.68% |
| **NIN + Dropout** | **10.41%** |
| CNN + Spearmint + Data Augmentation [14] | 9.50% |
| Conv. maxout + Dropout + Data Augmentation [8] | 9.38% |
| DropConnect + 12 networks + Data Augmentation [15] | 9.32% |
| **NIN + Dropout + Data Augmentation** | **8.81%** |

Table1. 様々な手法によるCIFAR-10のテストセットのエラーレート。

---

この実験では，各 mlpconv 層の特徴マップの数は，対応する maxout ネットワークと同じ数に設定されています．検証セットを用いて2つのハイパーパラメータ（局所受容野サイズと重みの減衰）を調整します。その後、ハイパーパラメータを固定し、トレーニングセットと検証セットの両方を用いて、ネットワークを一から再トレーニングします。その結果得られたモデルをテストに使用します。このデータセットでのテストエラーは10.41%で，最先端の手法に比べて1%以上改善されています．従来の手法との比較を表1に示す。

---

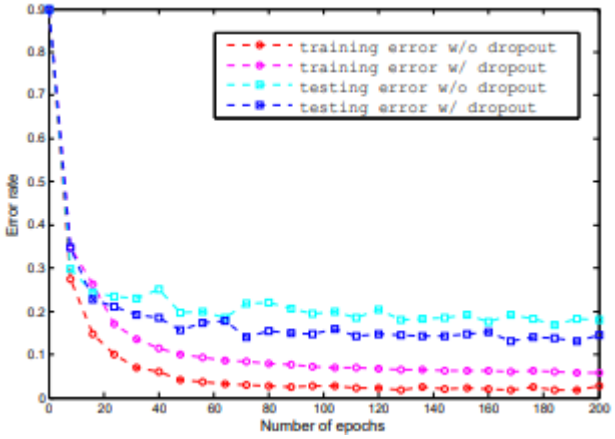


Fig4. mlpconv層の間のドロップアウトによる正則化効果。学習の最初の200エポックにおける，ドロップアウトのあるNINとないNINの学習誤差とテスト誤差を示す．

---

今回の実験では、NINのmlpconv層の間にドロップアウトを使用することで、モデルの汎化能力が向上し、ネットワークの性能が向上することがわかりました。図3に示すように、mlpconv層の間にドロップアウト層を導入することで、テストエラーが20%以上減少しました。この観察結果は、Goodfellowら[8]と一致しています。そこで、本稿で使用するすべてのモデルにおいて、mlpconv層の間にドロップアウトを追加しました。CIFAR-10データセットにおいて、ドロップアウト正則化なしのモデルは14.51%のエラーレートを達成し、これは正則化を用いた過去の多くの技術（maxoutを除く）を既に上回っている。ドロップアウトなしのmaxoutの性能は利用できないため、本稿ではドロップアウト正則化バージョンのみを比較している。

また，過去の研究との整合性をとるために，CIFAR-10データセットにおいて，平行移動と水平反転の拡張を行った上で，我々の手法を評価した．その結果、テストエラーは8.81%となり、最先端の性能を達成することができました。

## 4.3. CIFAR-100

CIFAR-100データセット[12]は、CIFAR-10データセットと同じサイズとフォーマットですが、100のクラスを含んでいます。そのため、各クラスの画像数はCIFAR-10データセットの10分の1しかありません。CIFAR-100ではハイパーパラメータの調整は行わず、CIFAR-10データセットと同じ設定を使用します。唯一の違いは、

最後のmlpconvレイヤーが100個のフィーチャーマップを出力することです。CIFAR-100のテスト誤差は35.68%となり、データ増強なしの現在の最高性能を1%以上上回ることができました。性能比較の詳細を表2に示します。

| Method | Test Error |
|---|---|
| Learned Pooling [16] | 43.71% |
| Stochastic Pooling [11] | 42.51% |
| Conv. maxout + Dropout [8] | 38.57% |
| Tree based priors [17] | 36.85% |
| **NIN + Dropout** | **35.68%** |

Table2. 様々な手法によるCIFAR-100のテストセットのエラーレート。

## 4.4. ストリートビューのハウスナンバー

SVHNデータセット[13]は，630,420枚の32x32カラー画像から構成されており，トレーニングセット，テストセット，および追加セットに分かれている．このデータセットの課題は，各画像の中央に位置する数字を分類することである．トレーニングとテストの手順は，Goodfellowら[8]に従っています．すなわち，トレーニングセットからクラスごとに400個のサンプルを選択し，追加セットからクラスごとに200個のサンプルを検証に使用します．トレーニングセットと追加セットの残りの部分はトレーニングに使用されます。検証セットはハイパーパラメータを選択する際のガイダンスとしてのみ使用され、モデルの学習には使用されません。

| Method | Test Error |
|---|---|
| Stochastic Pooling [11] | 2.80% |
| Rectifier + Dropout [18] | 2.78% |
| Rectifier + Dropout + Synthetic Translation [18] | 2.68% |
| Conv. maxout + Dropout [8] | 2.47% |
| NIN + Dropout | 2.35% |
| Multi-digit Number Recognition [19] | 2.16% |
| **DropConnect** [15] | **1.94%** |

表3. 様々な手法のSVHNに対するテストセットのエラーレート。

データセットの前処理は、再びGoodfellowら[8]に準じ、局所的なコントラストの正規化を行った。SVHNで用いた構造とパラメータはCIFAR-10で用いたものと類似しており、3つのmlpconv層とそれに続くグローバルアベレージプーリングで構成されています。このデータセットでは、テストエラー率は2.35%となりました。この結果を、データを補強しなかった手法と比較し、その比較結果を表3に示します。

## 4.5. MNIST

MNIST [1] データセットは，28×28サイズの手書きの数字0～9で構成されています．このデータセットには，60,000枚のトレーニング画像と10,000枚のテスト画像が含まれています．このデータセットでは、CIFAR-10で用いたのと同じネットワーク構造を採用しています。しかし、各mlpconvレイヤーから生成される特徴量マップの数は少なくなっています。MNISTはCIFAR-10に比べてシンプルなデータセットなので、必要なパラメータも少なくて済みます。このデータセットを用いて、我々の手法をデータ増強なしでテストする。その結果を、畳み込み構造を採用した過去の作品と比較し、表4に示す。

| Method | Test Error |
|---|---|
| 2-Layer CNN + 2-Layer NN [11] | 0.53% |
| Stochastic Pooling [11] | 0.47% |
| NIN + Dropout | 0.47% |
| **Conv. maxout + Dropout [8]** | **0.45%** |

表4. 様々な手法のMNISTに対するテストセットのエラーレート。

---

MNISTは非常に低いエラーレートに調整されているため、現在のベスト（0.45%）と同等の性能（0.47%）を達成しました。

## 4.6.

グローバルアベレージプーリング層は、ベクトル化された特徴マップを線形変換するという点で、完全連結層と似ています。違いは、変換行列にあります。グローバルアベレージプーリングでは、変換行列は前置されており、同じ値を共有するブロック対角要素でのみ非ゼロになります。完全連結層では、密な変換行列を持つことができ、その値はバックプロパゲーションによる最適化の対象となります。グローバルアベレージプーリングの正則化効果を調べるために、グローバルアベレージプーリング層を完全連結層に置き換え、モデルの他の部分はそのままとした。このモデルを，完全連結線形層の前にドロップアウトがある場合とない場合で評価した．両方のモデルをCIFAR-10データセットでテストし，性能を比較した結果を表5に示す．

---

| Method | Testing Error |
|---|---|
| mlpconv + Fully Connected | 11.59% |
| mlpconv + Fully Connected + Dropout | 10.88% |
| mlpconv + Global Average Pooling | 10.41% |

---

表5に示すように，ドロップアウト正則化を行わない完全連結層では，最悪の性能（11.59%）となった．これは，正則化を適用しない場合，完全連結層が学習データに過剰適合するためと予想される．完全連結層の前にドロップアウトを追加すると，テストエラーが減少した（10.88%）．大域平均プーリングは、3つのうちで最も低いテストエラー（10.41%）を達成した。

次に、グローバルな平均プーリングが従来のCNNに対して同じ正則化効果を持つかどうかを調べます。Hintonら[5]によって記述された従来型CNNをインスタンス化します。このCNNは3つの畳み込み層と1つのローカル接続層で構成されています。ローカル接続層では16個の特徴マップを生成し，これをドロップアウト付きの完全接続層に供給する．比較を公平にするために、ローカル接続層の特徴量マップの数を16から10に減らしています。これは、グローバルアベレージプーリング方式では、各カテゴリに1つの特徴量マップしか認められていないからです。そして、ドロップアウト＋完全連結層をグローバルアベレージプーリングに置き換えることで、グローバルアベレージプーリングを用いた同等のネットワークを作成する。これらの性能は，CIFAR-10データセットでテストされた．

このCNNモデルでは、完全連結層では17.56%のエラーレートしか達成できません。ドロップアウトを追加すると、Hintonら[5]の報告と同様の性能（15.99%）が得られます。このモデルで完全連結層をグローバルアベレージプーリングに置き換えると、エラーレートは16.46%となり、ドロップアウトなしのCNNと比べて1%改善されます。レギュラライザーとしてのグローバルアベレージプーリング層の有効性が改めて証明されました。ドロップアウト正則化の結果よりもわずかに悪いですが、カテゴリの信頼度マップをモデル化するためにアクティベーションを修正した線形フィルタを必要とするため、グローバルアベレージプーリングは線形畳み込み層には負荷がかかりすぎるのではないかと考えています。

## 4.7. NINの視覚化

NINの最後のmlpconv層の特徴マップを、グローバルアベレージプーリングによって、カテゴリーの信頼度マップになるように明示的に強制する。これは、NINのmlpconvなど、より強力な局所受容野モデリングでのみ可能なことである。この目的がどの程度達成されているかを理解するために、CIFAR-10の学習済みモデルの最後のmlpconv層から特徴マップを抽出し、直接視覚化します。

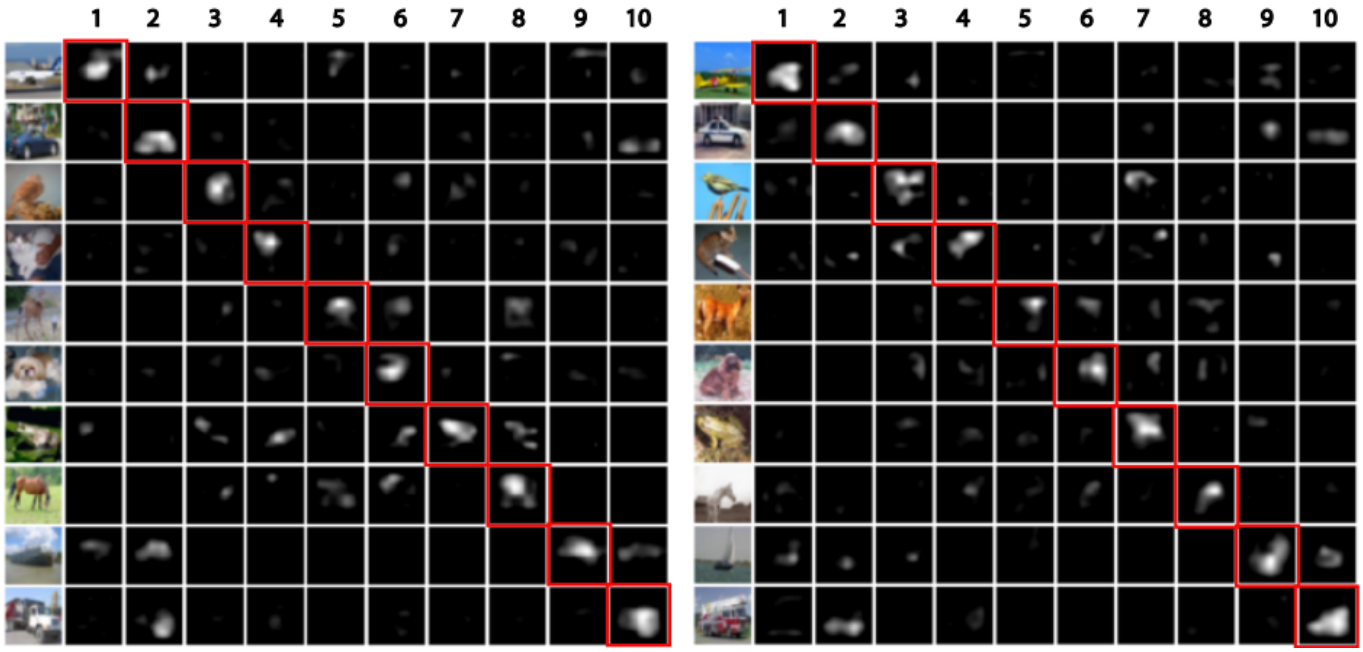


Fig4. 最後のmlpconvレイヤーの特徴マップを視覚化したもの。上位10%の活性化のみを だけを表示しています。特徴量マップに対応するカテゴリーは 1.飛行機 2.自動車、3.鳥、4.猫、5.鹿、6.犬、7.蛙、8.馬、9.船、10.トラック。フィーチャーマップ は，入力画像のグランドトゥルースに対応する特徴量マップが強調されている．左のパネルと右のパネル は別の例です。

図4は，CIFAR-10テストセットから選択した10個のカテゴリに対応する，いくつかの例示画像とその特徴マップを示している．入力画像のグランドトゥルースカテゴリに対応する特徴マップでは，最大の活性化が観察されることが予想されますが，これはグローバルアベレージプーリングによって明示的に強制されています。グランドトゥルースカテゴリの特徴マップでは、最も強い活性化が、元の画像のオブジェクトの同じ領域にほぼ現れていることが観察されます。これは、図4の2行目にある車のように、構造化されたオブジェクトに特に当てはまります。なお，カテゴリ用の特徴マップは，カテゴリ情報のみを用いて学習されたものである．オブジェクトのバウンディングボックスを使って細かいラベルを付ければ，より良い結果が期待できる．

この可視化により、NINの有効性が改めて示されました。これは、mlpconvレイヤーを使った、より強力なローカル再受容場モデリングによって実現されています。そして、グローバルな平均プーリングにより、カテゴリーレベルの特徴マップの学習が行われます。さらに、一般的な物体の検出に向けて、さらなる検討を行うことができます。Farabetら[20]のシーンラベリング作業と同じように、カテゴリレベルの特徴マップに基づいて検出結果を得ることができる。

この可視化により、NINの有効性が改めて示されました。これは、mlpconvレイヤーを使った、より強力なローカル再受容場モデリングによって実現されています。そして、グローバルな平均プーリングにより、カテゴリーレベルの特徴マップの学習が行われます。さらに、一般的な物体の検出に向けて、さらなる検討を行うことができます。Farabetら[20]のシーンラベリング作業と同じように、カテゴリレベルの特徴マップに基づいて検出結果を得ることができる。

# 5. 結言

我々は、分類タスクのために「Network In Network」（NIN）と呼ばれる新しい深層ネットワークを提案した。この新しい構造は、多層パーセプトロンを用いて入力を畳み込むmlpconv層と、従来のCNNの完全連結層の代わりとなるグローバルアベレージプーリング層で構成されています。Mlpconv層は局所的なパッチをよりよくモデル化し、グローバルアベレージプーリング層はグローバルにオーバーフィッティングを防ぐ構造的正則化の役割を果たします。NINのこれら2つのコンポーネントを用いて、CIFAR-10、CIFAR-100、SVHNの各データセットで最先端の性能を実証しました。また、特徴量マップの可視化により、NINの最後のmlpconv層からの特徴量マップがカテゴリの信頼度マップであることを示し、NINによる物体検出の可能性を示唆した。