深層畳み込みニューラルネットワークによる画像特徴抽出と転移学習

中山 英樹 †
† 東京大学 大学院情報理工学系研究科

Abstract

画像認識分野において，畳み込みニューラルネットワーク (CNN) は多くのタスクで驚異的な性能を達成し，注目を集めている．特に，ImageNet に代表される大規模物体認識データセットを用いて学習させた CNN の中間層から抽出される特徴は非常に汎用性が高く，さまざまなドメインで利用可能であることが示されている．本稿では，CNN の発展の歴史を概観したのち，CNN の特徴抽出器としての利用や，fine-tuning による転移学習の研究事例について紹介し，議論する．

1 はじめに

一般物体認識 (一般画像認識)[14] とは，制約のない実世界画像を言葉によって説明するタスクであり，古くから人工知能の究極的な目標の一つとされてきた (図 1)．現在，深層学習 (ディープラーニング) はさまざまな機械学習タスクで驚異的な性能を実現しているが，画像認識分野における躍進は研究業界のみならず広く一般に大きなインパクトを与えている．特に一般物体認識は，現在の深層学習および人工知能ブームの顔になっていると言っても過言ではないだろう．

深層学習の一般物体認識における大成功は，畳み込みニューラルネットワーク (CNN)[27] の構造がタスクに非常に良くはまったことに加え，質の良い大規模教師付きデータセットがいち早く整備され，研究コミュニティで共有されるようになったことに拠るところが大きい．これらのデータセットを用いて適切に学習させた CNN のパラメータは非常に汎用性が高く，強力な特徴抽出器として関連する他タスクへ転用可能であることが知られている．さらに，単に学習済みネットワーク (pre-trained network) を流用するだけでなく，これを初期状態としてさらに適用先タスクの訓練データで学習を進めることで，比較的少数の訓練データから極めて優れた性能が得られることが示されている．これらの転移学習法は，既に画像認識コミュニティにおいては必要不可欠な基本技術として確立しており，さまざまな pre-trained モデルがオープンソースで共有されている．以下では，これらの話題を中心に，CNN の歴史および最新動向について論じる．

2 画像認識における深層学習の歴史と発展

2.1 畳み込みニューラルネットワーク (CNN)

深層学習の手法は数多く提案されており，画像認識分野においても様々なアプローチが検討されてきたが，現在最も顕著な成功を収めているのは CNN である．CNN は古典的な多層パーセプトロンの延長にあるが，脳の視覚野の構造における知見 [18] を基に，ニューロン間の結合を局所に限定し層間の結合を疎にしていることを特徴とする．より具体的には，図 2 に示すように，画像の局所的な特徴抽出を担う畳み込み層と，局所ごとに特徴をまとめあげるプーリング層 (サブサンプリング層) を繰り返した構造となっている．畳み込みフィルタのパラメータは画像中のすべての場所で共有されるため，単純な全結合ネットワークに比べ大きくパラメータ数が減っている．また，プーリング層を交えることで，さらにパラメータ数を削減すると同時に，一般物体認識において必要不可欠である入力の平行移動に対する不変性を段階的に加えることができる．直感的には，入力の解像度を少しずつ落としながら異なるスケールで隣接する特徴の共起をとり，識別に有効な情報を選択的に上層へ渡していくネットワークであると解釈できる．このような畳み込み・プーリングの繰り返しによるアーキテクチャは日本発であり，福島らが開発した Neocognitron[10] が初出であった．その後，1990 年代に LeCun らによって誤差逆伝搬法による学習法が確立され [27]，現在にまで至る CNN の基本技術が確立された．

2000 年代に入ると，コンピュータビジョンのコミュニティでは一般物体認識の一大ブームが巻き起こり，CNN の応用も始まった [34, 33, 20]．しかし，この当時はデータ量・計算機パワーが共に十分ではなく，SIFT [30], HOG [4], bag-of-visual-words [3, 32] 等の経験的な特徴量ベースのシステムの方が優勢であった．例えば，一般物体認識で標準的なベンチマークとして長らく用いられてきた Caltech-101 [8] では数千枚程度の画像サンプルしか存在せず，入力から end-to-end で多数のネットワークパラメータを最適化することは困難であった．また，計算機の性能の面においても一般的な解像度の画像を扱うことは難しく，MNIST [26] (28×28 ピクセル) や CIFAR-10/100 [24] (32 × 32 ピクセル) 等の極

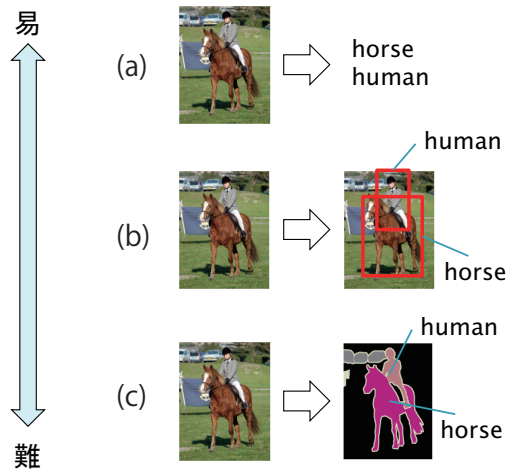

図 1　一般物体認識の主要なタスク．(a) 物体カテゴリ識別 (Categorization)．(b) 物体検出 (Detection)．(c) 物体領域抽出 (Semantic segmentation)．

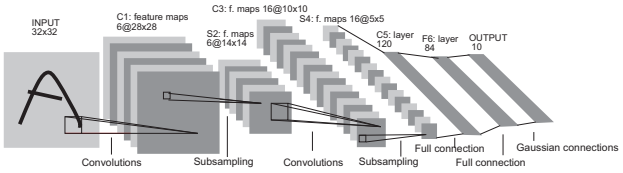

図 2　畳み込みニューラルネットワーク (LeNet-5)．図は [27] より引用．

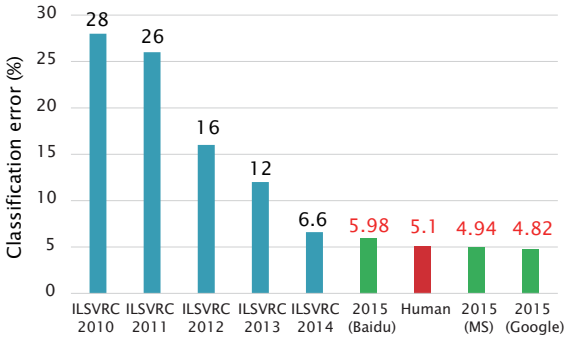

図 3　ILSVRC(1000 クラス物体識別タスク) のエラー率の推移．

端に小さい画像を用いた検証が主流であった．このため 2000 年代においては，CNN は現実的な方法としてはほとんど注目されず，停滞の時期にあったと言える．

2.2　ImageNet Large-scale Visual Recognition Challenge

2010 年代に入ると，状況は大きく変化しはじめた．特筆すべきは，ImageNet [5] という大規模教師付き画像データセットが公開されたことである．これは，自然言語処理分野で用いられる概念辞書である WordNet [9] に合わせ，網羅的に各概念のサンプル画像を収集したものである．種となる画像は既存のテキストベース画像検索エンジンを用いて収集し，クラウドソーシングによって人海戦術でアノテーションを行うことにより，大規模でありながら質の高い教師付きデータセットの構築に成功している．ImageNet は 2015 年 6 月現在，21841 クラス，14,197,122 枚ものアノテーション済み画像データを有する [1]．2010 年からは，ImageNet のデータの一部 (1000 クラス) を用いたコンペティション型ワークショップである ImageNet Large-scale Visual Recognition Challenge (ILSVRC) が毎年開催されており，2000 年代の研究の数百倍から数千倍もの規模のデータを自由に利用し，共通の土俵で競い合うことが可能になった．

また，同時期に，GPU 技術の進歩により計算機能力の著しい発達があったことも忘れてはならない．このように，学習データ量・計算機資源の双方において，深層学習が真の力を発揮する土壌が整いつつあった．

ブレークスルーとなったのは，2012 年の ILSVRC におけるトロント大学の Hinton らの躍進である．彼らは 8 層の CNN[2] を用い，1000 クラス識別のエラー率で，二位のチームに 10%以上もの差をつけて圧勝し [3]，世界中の研究者に極めて大きな衝撃を与えた [25]．2013 年の ILSVRC ではほぼすべてのシステムが CNN ベースに置き換わり，2014 年のコンテストでも更に大きく識別精度が改善している (図 3)．

2014 年以降は開発の主役が Web 系大企業に移り，それぞれしのぎを削っている状況である．ILSVRC 2014 では Google がエラー率 6.66% で優勝し [41]，その後 Baidu，Microsoft，Google がそれぞれ 5.98% [43]，4.94% [16]，4.82% [19] を達成している．同タスクにおける人間のエラー率は約 5.1% であるとの報告もあり，2012 年のブレークスルー後わずか二年ほどで人間レベルへ到達する驚異的な発展を遂げている．なお，これまでの ILSVRC の歴史は [39] に詳しく報告されているので，興味のある読者はぜひ参照されたい．

2.3　最新の研究動向

ILSVRC 2012 を境に，コンピュータビジョンの研究コミュニティでは従来の特徴量ベースのアプローチから急速に深層学習 (CNN) への乗り換えが進んだ．前述の通り，静止画のカテゴリ識別精度は既に人間と同レベルに至っていることから，現在は物体検出，物体領域抽出等のより難しい画像認識タスクへ焦点が移りつ

[1] http://www.image-net.org/

[2] その後，第一著者の Alex Krizhevsky にちなみ AlexNet と通称されるようになった．

[3] AlexNet を除くと，初回の ILSVRC 2010 からのエラー率の改善は 2~3% 程度に留まっており，頭打ちの傾向にあった．

つある (図 1). 物体検出においては，R-CNN [13] と呼ばれる手法が現在の主流である．前処理として物体の候補領域をあらかじめ多数取り出し，各候補領域について CNN で物体の有無を識別することで検出を行う．R-CNN は一枚の画像について数千個の領域候補画像を識別する必要があるため，GPU を用いても画像一枚あたりの認識に数十秒を要する極めて計算コストの大きい手法であったが，その後より効率のよい手法が多く提案されている [15, 12, 38, 36]．最新の手法では，物体候補領域の生成から識別・矩形抽出まで end-to-end で学習が行えるようになっている [38]．物体領域抽出はまだ発展途上であるが，やはり CNN をベースとしたものが中心的な役割を果たしている [29, 31]．

これらの従来的な一般物体認識タスクに加え，画像の自然言語による説明文生成 [22, 42] や，画像内容についての質疑応答 [11, 37]，動画像の認識・要約 [23, 6] など，さらに挑戦的なタスクも次々に取り組まれている．これらは，CNN を Recurrent neural network (RNN) [17] 等の別のネットワークと組み合わせることで実現されているが，いずれの場合もベースとなる CNN 自体の性能が極めて重要であることが知られている．

3 CNN を用いた転移学習

ILSVRC 2012 の結果は衝撃を持って受け入れられ，ワークショップ当日は活発な議論がなされた．その際の重要な問題提起の一つは，ImageNet により訓練された CNN はどの程度一般化できるのか，という問であった．すなわち，学習時 (ImageNet) と異なるタスク・データセットへの知識転移の実現可能性についての議論である．果たしてその後の研究トレンドは一気にこの方向へ動き，わずか半年ほどでその高い汎用性が立証されると共に，利用法が確立された．

3.1 Pre-trained ネットワークによる特徴抽出

CNN の最も簡単な利用方法は，学習済ネットワーク (pre-trained network) を固定し，純粋な特徴抽出器として用いる方法である (図 4)．すなわち，入力画像をフィードフォワードし，適当な中間層の出力する値をそのまま特徴ベクトルとして用いるものであり，利用者側は深層学習や CNN に関する知識がなくとも手軽にその恩恵を受けることができる．

Pre-trained network の利用においては，どの層から特徴抽出を行うかを考慮する必要がある．CNN では，入力に近い層から識別層に近づくにつれ，徐々に低次の視覚的特徴からデータセットに特化した意味的な特徴に構造化されることが知られている [44]．したがって，低すぎる層の特徴をとると CNN の高い識別的構造の恩恵を受けることができず，逆に高すぎる層の特徴を選ぶと学習時のデータセットに特化しすぎてしまい，転

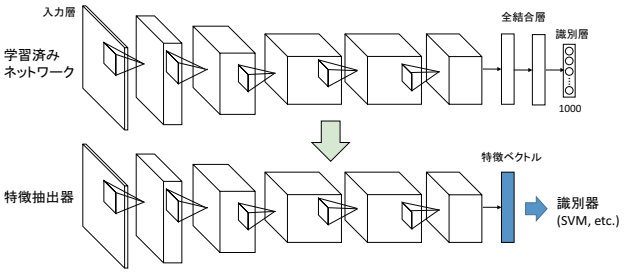

図 4　Pre-trained network を用いた特徴抽出.

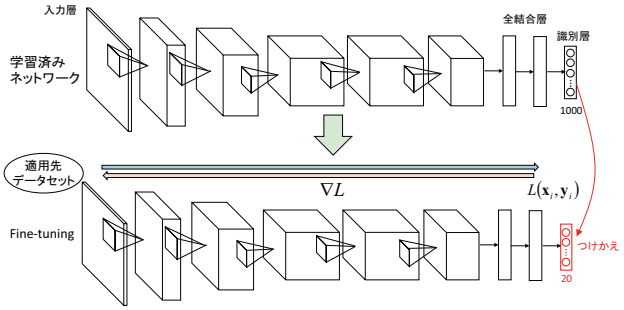

図 5　Pre-trained network に fine-tuning を加える場合の流れ.

移学習の性能が下がってしまうおそれがある．経験的には識別層の一つ二つ手前の全結合層を用いることが多い．

3.2 Fine-tuning による転移学習

Pre-trained network をより積極的に活用する転移学習法として，対象タスクのデータセットを用いて更にネットワークの学習をすすめる fine-tuning のアプローチも広く用いられている．図 5 に示すように，pre-trained network の識別層だけを対象タスクのものにつけかえる．その他の部分は学習済みのパラメータを初期値として用い，誤差逆伝搬法による学習を進める．一般に，CNN の学習は初期値依存性が強く，特に訓練データが少ない場合はできるだけよい初期値を得ることが，過学習を防ぎよい学習結果を得るために重要である．対象タスクに関連した pre-trained network を適切に選択し初期値として用いて fine-tuning を行うことで，フルスクラッチから学習するよりも格段によい結果を得られる場合が多い．

なお，一般に深層学習において fine-tuning という言葉は，教師なし事前学習でネットワークを初期化したあと教師あり学習を進めるプロセスのことを指すが，現在画像認識の文脈においてはここで述べたように他の大規模データセットを用いた教師付き学習による初期化を指す場合が多いことに注意されたい [4].

[4] 現在，CNN では教師なし事前学習はほとんど用いられなくなっている．

表 1　PASCAL VOC 2007 における，ImageNet pre-trained network (AlexNet) を用いた転移学習アプローチの比較 ([1] より引用)．特徴抽出器としてのみ利用した場合 (Pre-trained feature)，Fine-tuning を行った場合，およびフルスクラッチで CNN の学習行った場合の各検出成功率 (%) を示す．

Scratch	40.7
Pre-trained feature	45.5
Fine-tuning	**54.1**

表 2　PASCAL VOC 2007 における，さまざまな ImageNet pre-trained network をベースに fine-tuning を行った際の検出成功率 (%)([36] より引用)．物体検出手法はいずれも R-CNN を用いている．

AlexNet	58.5
Small VGG	60.2
VGG-16	**66.0**

3.3　事例紹介

Pre-trained network から得られる特徴量の利用については [7, 35] で詳しく調査され，ILSVRC のデータで学習した CNN から得られる特徴量は，物体認識・詳細画像カテゴリ識別・ドメイン適応・画像検索などのさまざまなタスクで非常に有効に働くことが報告されている．Fine-tuning のアプローチは [13] によって提案され，その後 [1] によって詳しく調査された．表 1 にその結果の一部を引用する．ここでは，AlexNet を pre-trained network として用い，物体検出のデータセットである PASCAL VOC 2007 (20 クラス，約 5 千枚の画像データセット) における検出成功率 (mAP) を示している．このように，フルスクラッチからターゲットのデータで学習した場合や，pre-trained network の特徴量のみを用いた場合に比べ，fine-tuning が非常によい精度を達成していることが分かる．

Pre-trained network を用いた転移学習では，元のモデル自体の性能も最終的なネットワークの精度に大きな影響を与える．2015 年現在は，ILSVRC 2014 でそれぞれ第二位，第一位であった Oxford visual geometry group の 16 層 CNN (VGG-16) [40]，Google の GoogLeNet[5] [41] がよく用いられるようになっている．表 2 に，[36] で報告された結果の一部を引用する．ターゲットのデータセットは同じく PASCAL VOC 2007 であり，R-CNN による物体検出においてベースとなる pre-trained network を変えて検出精度を評価したものである．この結果が示す通り，AlexNet から VGG-16 にモデルを差し替えるだけで大きく精度向上していることが分かる．今後も，ILSVRC のカテゴリ識別タスクにおける CNN の進化に伴い，標準的に用いられる pre-trained network は随時置き換わっていくものと思われる．

なお，ImageNet を用いた pre-trained network は何にでも転用可能なわけではなく，ターゲットとするタスクが ImageNet のカバーする領域に関連するものでなければ必ずしもよい結果は得られないことに注意が必要である．例えば，シーン認識タスクは ImageNet が対象とする物体認識とはやや離れた関係にあるため，他のタスクに比べ ImageNet ベースの pre-trained network による転移学習の効果が薄いことが報告されている [7, 45][6]．

3.4　実践方法

現在，画像認識分野における深層学習の標準的な OSS である Caffe[21] には model zoo というモデル共有の枠組みが用意されており，多くの研究者がそれぞれの手法で構築した学習済みネットワークを公開している．前述の AlexNet はもちろんのこと，network-in-network [28]，VGGnet [40], GoogLeNet [41] 等の最新の成果も次々と共有されており，自由に利用することが可能である [7]．Pre-trained network を用いた特徴抽出は，スクリプトを実行するだけで容易に行える環境が整っている．Fine-tuning を行う際は CNN の学習に関するノウハウが多少必要となるが，基本的にモデル自体は流用するためネットワーク構造に関するハイパーパラメータの設定は必要なく，フルスクラッチの学習と比較すると容易である．筆者の経験上，主に学習率の設定さえ気をつければ，十分によい結果が得られる場合が多い．

4　まとめと今後の展望

本稿では，画像認識分野における深層学習の歴史と最新動向について俯瞰し，特に CNN の特徴抽出器としての利用や，fine-tuning による転移学習について中心的に紹介を行った．これらは既に手軽に利用可能な技術として確立しており，深層学習研究におけるロールモデルの一つになっていると言える．

このような CNN の一般物体認識における驚異的な成功を受け，画像認識こそが深層学習に最も向いたタスクであると見られる向きも少なくない．しかしながら筆者の意見では，これはタスクの性質というよりも，静止画は比較的クラウドソーシングによるアノテーショ

[5] 初代 CNN である LeNet [27] にちなんで付けられた名前である．

[6] シーン認識に特化した Places [45] と呼ばれる大規模データセットを MIT が公開しており，ImageNet 同様，数百万枚のラベル付き画像と学習済みネットワークが入手可能になっている．

[7] これらのモデルは，Torch7 [2] 等の他の OSS でも利用できる．

ンが容易なため，ImageNet のような質のよい大規模教師付きデータセットがいち早く登場したことに負うところが大きいと考える．事実，現在研究業界で華々しい成果をあげているシステムは，元を辿ると何らかの形で ImageNet や同規模のデータを利用しているものが大半であり，そこから外れるもの (例えば歩行者検出，動画像認識等) では既存の特徴量と比較して必ずしも優れた成果をあげていないことに注意する必要があるだろう．

現在，画像認識分野はもとより，人工知能に関わるあらゆる分野においてより高度なタスクの実現へ向けて期待が高まっているが，手法に関する議論は盛んに為される一方で，それを支えるデータについては必ずしも十分に注意が払われていないように感じる．手法とデータは常に車の両輪の関係にあり，両者をバランスよく発展させることが，深層学習が次のブレークスルーを起こせるか否かの鍵であると考える．

参考文献

[1] P. Agrawal, R. Girshick, and J. Malik. Analyzing the Performance of Multilayer Neural Networks for Object Recognition. In *Proc. ECCV*, 2014.

[2] R. Collobert, K. Kavukcuoglu, and C. Farabet. Torch7: A matlab-like environment for machine learning. *BigLearn, NIPS Workshop*, pages 1–6, 2011.

[3] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Proc. ECCV Workshop on Statistical Learning in Computer Vision*, 2004.

[4] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. In *Proc. CVPR*, 2005.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proc. CVPR*, pages 2–9, 2009.

[6] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term Recurrent Convolutional Networks for Visual Recognition and Description. In *Proc. IEEE CVPR*, 2015.

[7] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. In *Proc. ICML*, 2014.

[8] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Journal of Computer Vision and Image Understanding*, 106(1):59–70, 2007.

[9] C. Fellbaum. *WordNet: An electronic lexical database*. MIT Press, 1998.

[10] K. Fukushima. Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):93–202, 1980.

[11] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, L. Angeles, and W. Xu. Are You Talking to a Machine? Dataset and Methods for Multilingual Image Question Answering. *arXiv preprint arXiv:1505.05612*, 2015.

[12] R. Girshick. Fast R-CNN. *arXiv preprint arXiv:1504.08083*, 2015.

[13] R. Girshick, J. Donahue, T. Darrell, U. C. Berkeley, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proc. IEEE CVPR*, 2014.

[14] K. Grauman and B. Leibe. *Visual object recognition.* Morgan & Claypool Publishers, 2011.

[15] K. He, X. Zhang, S. Ren, and J. Sun. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *arXiv preprint arXiv:1406.4729*, 2014.

[16] K. He, X. Zhang, S. Ren, and J. Sun. Delving Deep into Rectifiers : Surpassing Human-Level Performance on ImageNet Classification. *arXiv preprint arXiv:1502.01852*, 2015.

[17] S. Hochreiter and J. Schmidhuber. Long Short-term Memory. *Neural Computation*, 9(8):1–32, 1997.

[18] D. Hubel and T. Wiesel. Receptive fields of single neurones in the cat's striate cortex. *The Journal of physiology*, 148:574–591, 1959.

[19] S. Ioffe and C. Szegedy. Batch Normalization : Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv preprint arXiv:1502.03167*, 2015.

[20] K. Jarrett, K. Kavukcuoglu, M. A. Ranzato, and Y. Lecun. What is the best multi-stage architecture for object recognition? In *Proc. IEEE ICCV*, 2009.

[21] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe : Convolutional Architecture for Fast Feature Embedding. In *ACM Conference on Multimedia*, pages 675–678, 2014.

[22] A. Karpathy and L. Fei-Fei. Deep Visual-

Semantic Alignments for Generating Image Descriptions. In *Proc. IEEE CVPR*, 2015.

[23] A. Karpathy and T. Leung. Large-scale Video Classification with Convolutional Neural Networks. In *Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

[24] A. Krizhevsky. *Learning multiple layers of features from tiny images*. Master's thesis, Toronto University, 2009.

[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *Proc. NIPS*, 2012.

[26] Y. LeCun. The MNIST database of handwritten digits.

[27] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proc. of the IEEE*, pages 2278–2324, 1998.

[28] M. Lin, Q. Chen, and S. Yan. Network in network. In *Proc. ICLR*, 2014.

[29] J. Long, E. Shelhamer, and T. Darrell. Fully Convolutional Networks for Semantic Segmentation. In *Proc. IEEE CVPR*, 2015.

[30] D. Lowe. Object recognition from local scale-invariant features. In *Proc. IEEE ICCV*, pages 1150–1157, vol.2, 1999.

[31] M. Mostajabi, P. Yadollahpour, and G. Shakhnarovich. Feedforward semantic segmentation with zoom-out features. In *Proc. IEEE CVPR*, 2015.

[32] F. Perronnin, J. Sánchez, and T. Mensink. Improving the Fisher kernel for large-scale image classification. In *Proc. ECCV*, 2010.

[33] M. A. Ranzato, F. J. Huang, Y. L. Boureau, and Y. LeCun. Unsupervised learning of invariant feature hierarchies with applications to object recognition. In *Proc. IEEE CVPR*, 2007.

[34] M. A. Ranzato, C. Poultney, S. Chopra, and Y. LeCun. Efficient learning of sparse representations with an energy-based model. In *Proc. NIPS*, 2006.

[35] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. CNN Features off-the-shelf : an Astounding Baseline for Recognition. *arXiv preprint arXiv:1403.6382*, 2014.

[36] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You Only Look Once : Unified , Real-Time Object Detection. *arXiv preprint arXiv:1506.02640*, 2015.

[37] M. Ren, R. Kiros, and R. Zemel. Exploring Models and Data for Image Question Answering. *arXiv preprint arXiv:1505.02074*, 2015.

[38] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN : Towards Real-Time Object Detection with Region Proposal Networks. *arXiv preprint arXiv:1506.01497*, 2015.

[39] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, pages 1–42, 2015.

[40] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recoginition. In *Intl. Conf. on Learning Representations (ICLR)*, 2015.

[41] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going Deeper with Convolutions. In *Proc. IEEE CVPR*, 2015.

[42] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and Tell : A Neural Image Caption Generator. In *Proc. IEEE CVPR*, 2015.

[43] R. Wu, S. Yan, Y. Shan, Q. Dang, and G. Sun. Deep Image : Scaling up Image Recognition. *arXiv preprint arXiv:1501.02876*, 2015.

[44] M. D. Zeiler and R. Fergus. Visualizing and Understanding Convolutional Networks. In *Proc. ECCV*, 2014.

[45] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning Deep Features for Scene Recognition using Places Database. In *Proc. NIPS*, 2014.