解説

画像解析関連コンペティションの潮流

Trends on Competitions of Image Analysis and Related Areas

中山英樹

Abstract

統計的学習に基づく画像認識は，深層学習の驚異的な成功を背景に目覚ましい発展を続けている．特に近年は，単なる高精度化にとどまらず，次々に新しい高度な認識タスクが実現可能となっている．この分野では学習のためのデータセットが重要な役割を果たしているが，特にコンペティション形式で開催されるワークショップが近年の技術革新の大きな原動力となっている．本稿では，現在本分野において中心的存在となっている大規模画像認識コンペティションであるILSVRC を中心に，研究コミュニティで扱われてきたタスクの変遷と最新動向について，その他の画像系コンペティションの動向も含めながら紹介する．更に，現状で機械は一般的な画像から何を語れるレベルまでに達したのか，また，画像解析・認識の今後の展望についても触れる．

キーワード：画像認識，データセットと評価方法，コンペティション，ILSVRC

1. 統計的学習に基づく画像認識の発展

制約のない実世界画像の認識・理解は一般物体認識[(1)]と呼ばれ，古くから人工知能の究極的な目標の一つとされてきた．古典的には幾何的なルールや三次元モデルに基づくアプローチが模索されていたが，1990 年代前後から統計的パターン認識に基づくアプローチが大きな成功を収めるようになった．2000 年代に入るとより現代的な機械学習手法が多用されるようになり，現代に至るまでコンピュータビジョン分野における一大人気トピックとなっている．

より具体的に画像認識の目的を述べると，入力画像 x が与えられたときに，これを自動的に何らかのラベル y（例えばカテゴリー名や物体の座標など）へ変換する関数 $f{:}x\rightarrow y$ を得ることであると言える．統計的学習に基づくアプローチでは，入出力の多数の事例データから成る訓練データセット $\{x_i, y_i\}_{i=1}^{N}$（N はデータ数）から，そのような関数 f を帰納的に推定する．統計的画像認識手法はこれを実現するために，f としてある識別モデルや確率モデルを仮定し，そのパラメータを推定する一連の手続きを与えるものである．

さて，統計的学習に基づくシステムにおいては，認識手法の理論自体ももちろん大事であるが，それ以上に訓練データセットが重要である．いかに高度な数理的手法を用いようとも，訓練データセットが想定していない物事の認識は本質的に不可能であるため，所望のタスクに対して十分に網羅的かつ質の良い事例データを用意する必要がある．また，十分な汎化性能（テスト時の性能）を得るためには訓練データの量も同様に重要である．一般に，多数のパラメータを有する複雑なモデルほど表現能力は高いが，より多数の訓練データがなければ過学習に陥りやすく，高い性能を得ることは難しい．歴史をひも解くと，モデルとデータセットは常に一体となって発展してきており，その時々に最先端として注目されていた手法は，当時のデータの規模や計算機の速度に最も適合した複雑さを有するものであったと言える．現在深層学習として全盛期を迎えているニューラルネットワークも最初から現在のような成功を収めていたわけではなく，データ量や計算機パワーに乏しかった時代は十分な力を発揮できず，浅い識別モデルと特徴量に基づくアプローチがはるかに優勢であったことを忘れてはならないだろう．

本稿では，統計的学習に基づく画像認識を支えるデータセットの役割に注目する．特に，本分野においては共通のデータセットをベンチマークとして一斉に用いるコ

中山英樹　正員　東京大学大学院情報理工学系研究科創造情報学専攻
E-mail　nakayama@ci.i.u-tokyo.ac.jp
Hideki NAKAYAMA, Member (Graduate School of Information Science and Technology, The University of Tokyo, Tokyo, 113-8657 Japan).
電子情報通信学会誌　Vol.100 No.5 pp.373-380 2017 年 5 月

ンペティションが技術的ブレークスルーの原動力となっており，その歴史と現在の動向について議論する．

2. 画像認識のベンチマーキング

統計的学習により画像認識システムを構築するためにデータセットが原理的に不可欠であることは既に述べたとおりであるが，同時にデータセットは認識手法の良し悪しを定量的に評価するベンチマークとしての役割も担っている．評価においては，まず，人手によってラベルが付与された多数のデータ（画像）を，訓練用データ，検証用データ，テストデータの三つのサブセットに分ける．訓練用データ・検証用データによって認識システムの学習及びチューニングを行った後，テストデータの予測精度を測る．ここで，テストデータは答えのラベル（グランドツルース）を伏した状態で画像のみを与え，システムが予測した結果がラベルとどれだけ合致するかを評価する．用いるデータの条件をそろえることで定量的かつ客観的な評価基準を設けることができ，工学的に認識手法の研究開発を積み重ねていく上で非常に重要なプロトコルとなっている．画像認識手法や機械学習手法の研究論文を投稿する際は，幾つかの標準的なベンチマークデータセットにおいて他手法と比較を行い，定量的に優位性を示すことが求められる．

このようなベンチマークデータセットは数多く作られているが，ここでは歴史上特に重要な役割を果たしてきたものを紹介する(注1)．

2.1 初期の小規模データセット

一般物体認識における最も基本的な問題であるカテゴリー識別においては，カリフォルニア工科大学が2004年に公開したCaltech-101[3]が最初にデファクトスタンダードとなったデータセットである．これは，Web上の画像を収集し，101種類の物体クラスと背景クラスについてそれぞれ31枚から800枚の画像データをラベル付けしたものである．これ以前の研究はほとんどが数クラスの物体認識にとどまっており，初めて現代的な一般物体認識の概念を評価データセットと合わせて確立した研究であると言える．2007年には対象クラス数を256クラスへと拡張したCaltech-256が登場しており[4]，Caltech-101と合わせて長い間評価に用いられてきた．これらのデータセットに含まれる画像は，色やテクスチャはバラエティに富んでいるものの，単一の物体がほぼスケールや向きをそろえて中心に位置している理想的なセットアップとなっており，現実の画像認識の状況からは掛け離れた簡単な問題であるとの批判も受けてきた．

このような問題意識から，もう一つの標準的なベンチマークデータセットとして，詳しく後述するPASCAL Visual Object Classes（VOC）[5,6]がコンペティションと合わせて開発されてきた．扱うクラス数は最大で20クラスと少ないが，複数の物体が様々な条件下で自然に映った画像を扱っており，Caltech-101と比較して，より一般的な状況下へタスクを近付けることを意識したデータセットとなっている．また，カテゴリー識別に加え物体検出タスク用のアノテーションもなされており，特に物体検出においては現在に至るまで最も中心的なベンチマークの一つとして用いられてきた．

上記以外にも，画像全体からのシーン[7,8]やイベントの認識[9]や，物体領域分割[10]など，様々なタスクがデータセットと合わせ提案されてきた．これらの初期のデータセットは基本的に各研究者レベルで画像の収集・ラベル付けをしたものであるため，最大でも数千枚程度の小規模なものであった．ここではまず認識手法の評価が目的であり，学習後に得られる識別システムの実用性は必ずしも考慮されていない[11]．

2.2 大規模データセット

真に有用な画像認識の実現には，実世界の広範なアピアランスを網羅する大規模なデータセットを構築する必要があるが，そのための膨大な人的リソースをどのように確保するかが課題であった．転機となったのは，Amazon Mechanical Turk（AMT）に代表されるクラウドソーシングの発達である[12]．インターネット経由で不特定多数の作業者を雇用し，画像ラベリング作業を依頼することにより，十分な資金があれば短期間で非常に大規模なデータセットを構築することが可能となった．また，アノテーション作業の内容は自由に設計できるため，研究の用途に応じて様々なデータセットを構築できる．現在最も大規模かつ標準的なデータセットが，Stanford大のグループが開発しているImageNet[13]である．ImageNetは，自然言語処理分野で提案された概念辞書であるWordNet[14]に合わせ，網羅的に各単語概念に対応する画像を収集したものであり，2016年11月現在，2万1,841クラス，1,400万枚ものアノテーション済み画像データを有する(注2)．

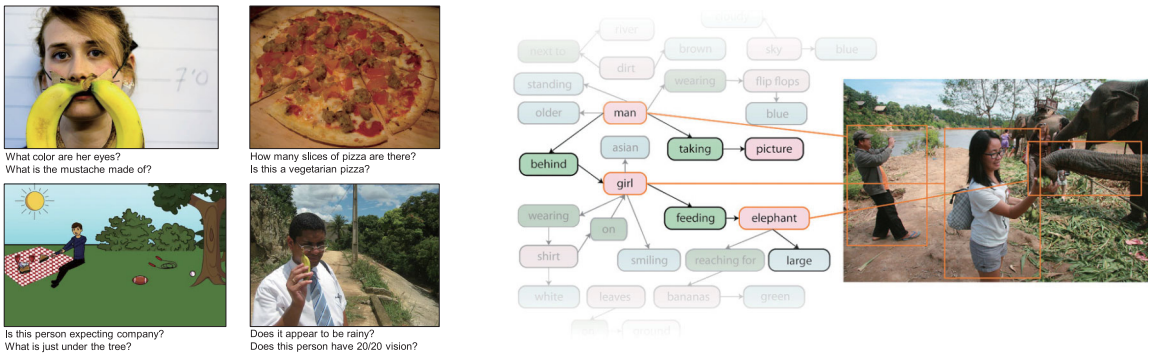
クラウドソーシングは既に現在のデータセット作成においては必要不可欠なツールとなっており，ほかにも多くの大規模画像データセットが誕生している．ImageNetと並び広く普及しているデータセットとして，Microsoft COCO[15,16]が挙げられる．こちらは約33万枚の画像から成りImageNetより数は少ないが，インスタンスレベルでの領域分割のための高精細な物体マスクや，自然言語による説明文など，より豊富なアノテー

(注1) なお，2006年以前のデータセットについては，文献(2)に詳しく議論されており，是非参照されたい．
(注2) http://www.image-net.org/

(a) ImageNet

(b) Microsoft COCO

(c) MIT Places



(d) VQA　　(e) Visual Genome

図1　各データセット　図はそれぞれ文献(26)，(15)，(16)，(18)，(23)，(25)から引用．

ションがなされている．その他，SUN[17]やMIT Places[18]（シーン認識），Sports-1M[19]やActivity-Net[20]（動画像カテゴリー認識），MPII-MD[21]やMSR-VTT[22]（動画像説明文生成），VQA[23]やVisual7W[24]（画像質問応答）など，各種新タスクのための大規模データセットが次々に開発されている．2016年には，ImageNetを開発したStanford大のグループが，画像中の複数物体の関連性や属性をグラフ形式で極めて細かくアノテーションしたVisualGenome[25]と呼ばれる次世代データセットを公開し，注目を集めている．図1に各データセットの画像例を示す．

3. コンペティション

ベンチマークデータセットは通常，各研究者が自由なタイミングで各自の手法を評価するために用いるが，しばしばデータセット作成者によりコンペティションが開催されている．コンペティションでは，基本的に同じタイミングでデータの配布や，テスト結果の提出と評価が行われるため，よりスコア競争の色合いが強いものとなり，最先端の認識技術の見本市となっている．同時に，コンペティションで得た知見を次のデータセット作成や拡張へフィードバックし，画像認識のコミュニティとしてより優れた方向性を見いだしていくことも重要な目的である．ここでは，画像認識分野において中心的な存在となってきた二つのコンペティションを解説する．

3.1 PASCAL Visual Object Classes Challenge

PASCAL VOC challenge[5),6)]は，ヨーロッパの研究コミュニティであるPASCAL（Pattern Analysis, Statistical Modelling and Computational Learning）が主催するコンペティションであり，2005年から2012年まで，ECCVやICCVなどの有力な国際会議のサテライトワークショップとして毎年開催されてきた．前章で述べたとおり，できるだけ現実的な画像認識の問題を解くことを念頭に，画像共有サイトであるFlickrの画像を利用してデータセットを構築している．また，明確なデータの区分を与え評価ソフトウェアを配布するなど，厳密な定量評価が可能なように工夫されている．当初は10クラスのカテゴリー識別タスクからスタートしたが，後に20クラスへと拡張され，物体検出タスクも始まった．これらのほかにも動作識別，セマンティックセグメンテーション，人体部位のレイアウト推定など，年によって新しいテストタスクも実施されている．また，コンペティション終了後も比較評価ができるように，2007年まではテストデータのグランドツルース，2008年以降は評価サーバを終了後に公開している．

VOC challengeを土壌に，HOG[26)]やbag-of-visual-words（BoVW）[27)]など多くの特徴量や，deformable part model[28)]などの優れた物体検出手法が確立され，初期の分野への貢献は極めて大きいものであったと言える．しかしながら，データやクラス数は少ないためモデルの工夫には限界があり，2010年以降には多くの手法が似たような構成へと収束し，その役割を終えることとなった．

3.2 ImageNet Large-Scale Visual Recognition Challenge（ILSVRC）

ImageNetのデータの一部を用い，2010年から始まった大規模画像認識コンペティションがILSVRC[29)]であり，米国の複数の大学が共同で開催している．PASCAL VOC challengeの成功を受け，データのセットアップや，コンペティション終了後のテストサーバの提供など多くの面でそのスタイルを踏襲した運営になっている．

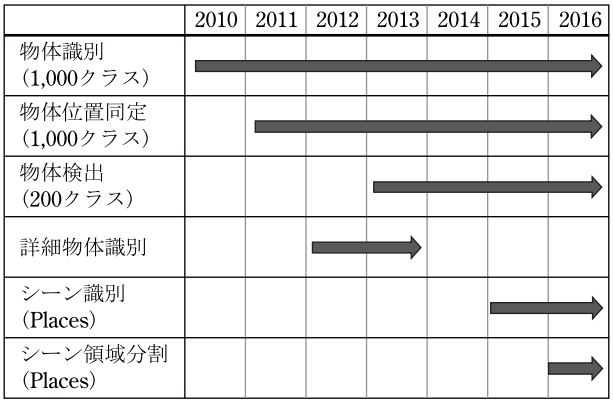


図2 ILSVRCで実施されてきた認識タスク シーン識別，シーン領域分割はMIT Placesが主催．

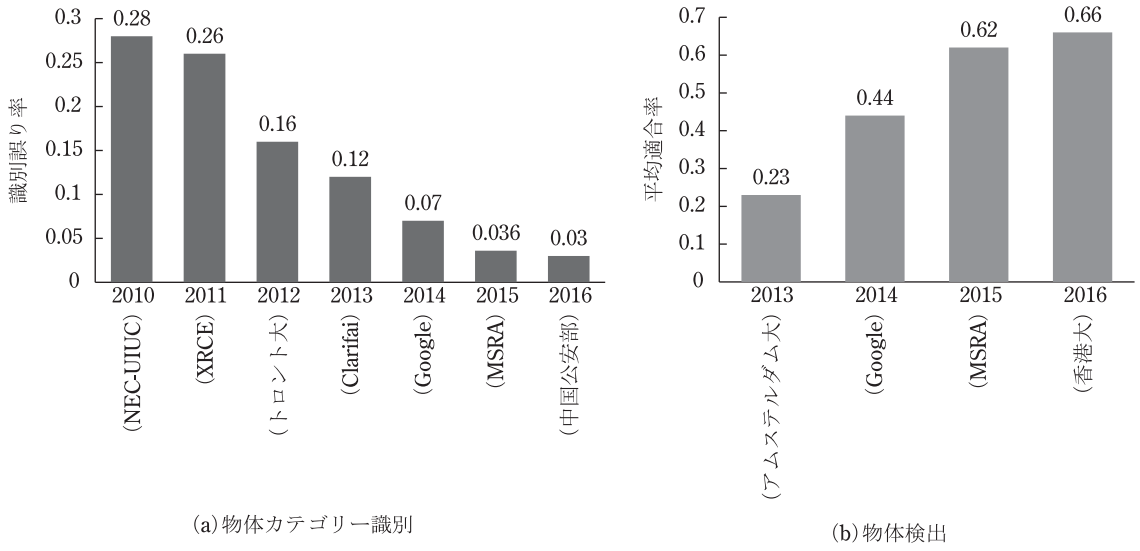
初回は，1,000クラスのカテゴリー識別が題材となった．データセットは120万枚の訓練画像，15万枚のテスト画像（現在は10万枚）から成り，PASCAL VOC Challengeと比較して文字通り桁違いに大規模なものであった．その後，カテゴリー識別は物体位置同定タスク（localization）へと一般化され，物体検出，動画像物体検出等次々に新しいタスクが追加されている．図2に，各年のILSVRCで実施されてきたタスクをまとめた．参加者の数も，2010年の35チームから始まり2016年は172チームを数えるまでに至っており，画像認識研究者のみならず，広く一般に訴求力のある一大イベントへと成長している．

3.2.1 メインタスクにおける歴史と進歩

図3に，1,000クラス物体カテゴリー識別と，200クラス物体検出タスクのスコアの推移をまとめた．前者では，トップ5の誤り率（出力した上位5クラスにグランドツルースのクラスが入っていれば正答とみなす），後者ではMean Average Precisionが評価指標となっている．以下に，この二つのタスクを中心に，各年のILSVRCにおけるトレンドをまとめる．

2010年，2011年は，深層学習のブレーク以前に研究されてきた画像特徴量の成熟期であり，上位チームのシステムの多くはこれらを徹底的に活用したものであった．特に，大規模な訓練データからの学習を可能とするために，距離計量を基に適切なユークリッド空間へ埋め込まれたBoVW表現[30)~32)]と，オンライン線形識別器を組み合わせるアプローチがトレンドであった．

2012年は深層学習のブレークスルーとなった年である．この年，トロント大学のHinton教授らのグループ



(a)物体カテゴリー識別　　(b)物体検出

図3　ILSVRC における優勝スコアの推移

は，8層の畳込みニューラルネットワーク（CNN）を用い，1,000クラス識別の誤り率で2位以下のチームに10%以上もの大差を付けて圧勝し，世界中の研究者に極めて大きな衝撃を与えた[33]．この後，画像認識のアプローチは，従来の特徴量ベースからニューラルネットワークへ一気に移行することになる．

2013年のILSVRCではほぼ全てのシステムがCNNベースに置き換わった．1,000クラス識別タスクにおいてはモデルにそれほど大きな進歩は見られず，従来の画像認識研究者にとってはキャッチアップの年であったと言える．一方，新しく開始された物体検出タスクにおいては，物体候補領域とCNNによる特徴抽出を活用するRCNN[34]について紹介があり，deformable part modelを大きく打ち破る結果が報告されるなど，物体検出においても深層学習ベースに大きな展開を見せ始めている．

2014年には，モデルの深さが精度向上に重要であることが分かり，Googleが22層，Oxford VGGグループが19層のネットワークを構築し，好成績を上げている．このようなより深いモデルの学習を可能とするbatch normalization[35]，leaky ReLU[36]，ネットワーク初期化法[36]などの研究が進展したのもこの年の特徴であろう．物体検出においても，RCNNとこれらの強力な識別ネットワークを合わせて用いることで，2013年度の約2倍の精度を達成している．この頃から主なプレーヤがGoogleやMicrosoftなどのIT系大企業に移り，非常に力を入れてスコア競争を行うようになってきた．また，2014年以降はNVIDIAがコンペティション参加者にGPGPU計算環境を無償で貸与するサービスを開始するなど，深層学習を中心にアカデミアのみならず産業界も非常に大きな盛り上がりを見せるようになった．

2015年には，Microsoft Research Asiaが152層のCNNを用い，1,000クラス識別において誤り率3.57%を達成した．鍵となったのは，畳込み層をバイパスする結合を有するresidual network（ResNet）[37]と呼ばれる構造であり，従来よりも更に深いモデルの学習が可能となった．同タスクにおける人間の誤り率は約5.1%であるとの参考値も示されており，2012年のブレークスルー後僅か3年ほどで人間レベルを上回る驚異的な発展を遂げたことになる．

2016年は，中国公安部がResNetやその改良版を駆使したシステムにより識別率2.99%を達成し優勝した．しかしながら，上位5チームの誤り率の差は0.3%未満となっており，必ずしも有意差を認められる結果であるとは言い難い．この年は，前年のResNetの成功を受け，層をスキップする結合の工夫が多くなされたが，結果的には層の深さ・認識精度共に頭打ちの傾向となった．これまで最も重要なベンチマークであった1,000クラス識別タスクであるが，より強力なモデルを用いた精度向上を行うためにはもはやデータが少な過ぎる可能性があり，より多クラス・大規模なベンチマークが必要とされる時期に移ってきたと考えられる．

3.2.2　ILSVRC のもたらした変化

ILSVRCでは毎年同じタスクがなされているものの，常にその在り方は議論され，必要に応じて新しいレギュレーションを追加してきた．例えば，深層学習のブレーク以前は，提供されている訓練用データ以外のデータを学習に利用することは，フェアな評価を妨げるものとして認められていなかった．しかしながら，深層学習では多量の外部データで事前学習を行うことは極めて一般的であったため，ベンチマーク内の訓練用データのみにこだわることは時代遅れとなってきた．これを受け，外部

データを用いる場合・用いない場合とでエントリを分け，別の部門として評価を行うようになっている．また，ILSVRC のデータで学習されたネットワークはオープンソース上で共有されるようになり，他のデータセット上で fine-tuning を行う転移学習も現在では分野内で定石の一つとして確立している．このように，コンペティションでの議論や試行錯誤を通じ，分野における文化や常識もダイナミックに変化していることは興味深い．

3.3 その他の画像・動画像認識コンペティション

ImageNet と同様のコンペティションは数多く開催されており，近年はますます増加傾向にある．これらは新しいデータセットの作成と一体になっていることが多く，2.2 も併せて参照されたい．画像認識において ImageNet と並び注目されているのは，Microsoft COCO[15),16)] を用いた COCO challenge[注3] であり，インスタンスレベルの物体領域分割，画像説明文生成等のタスクが実施されている．また，MIT Places[18] を用いた大規模シーン認識コンペティションである Places2 challenge[注4] も，1,000 万枚規模の学習データを用いる新しい識別タスクとして注目されている，なお，COCO challenge, Places chalenge は 2015 年から ILSVRC と併催され，同じワークショップで発表されるようになっている．ほかにも，THUMOS challenge[注5]，ActivityNet challenge[注6] といった動画像からの多クラス動作識別や，VQA challenge[注7]（画像質問応答），LSMDC[注8]（動画像説明文生成）など新しいタスクを扱ったコンペティションが多数始まっており，いずれも CVPR 等の国際会議のサテライトワークショップとなっている．

画像以外のメディア（音声，テキスト等）も積極的に利用するマルチメディアの分野においては，ImageCLEF[注9]，MediaEval[注10] などが歴史あるコンペティションとして挙げられる．また，ACM Multimedia というマルチメディアの難関国際会議では，会議自体が Grand Challenge というユニークなプログラムを持っており，毎年企業から挑戦的なテーマとデータを募り，コンペティション形式のワークショップを実施する．2016 年は，MSR-VTT[22] を用いた動画像説明文生成タスクがその一つとして取り上げられた[注11]．米国 NIST が主催する TRECVID[注12] も歴史あるコンペティションであり，非常に大規模な動画像アーカイブから画像・音声・テキスト等の情報を駆使し，実用的な検索や認識の精度を競うものである．データの規模が他のワークショップと比較しても桁外れに大きく，一般的な画像認識の手法をナイーブに適用することは困難な場合が多く，スケーラビリティに優れたシステムが数多く提案されてきた．

4. コンペティションの意義と弊害

このように，大規模なデータセットとコンペティションは一体となって開発されてきたことを述べたが，そもそもデータセット自体が既に定量的評価を可能としているにもかかわらず，更に大きな労力を割いてコンペティションを行う意義はなんであろうか．個人的には，大きく分けて以下の 3 点が重要なメリットではないかと考える．

（1） 厳密な一発勝負の評価が行える

一般的なベンチマークデータセットではテストデータのグランドツルースも公開されているため，真の意味でテストデータに対する性能を評価することは難しい．例え各試行ではテストデータを隔離していたとしても，手法開発からテストに至る一連の流れを何度も試行錯誤すれば，結果的にテストデータへのチューニングを行っていることと変わらなくなるためである．これをどこまで厳密に考えるかは各研究者のモラルに依存しているのが現状であり，論文でのスコアは非常に高かった手法がコンペティションでは低調に終わることも珍しくない．

また，通常の研究論文におけるベンチマーク評価は常に「後出しじゃんけん」の形となり，それまでの最高スコアを目標に試行錯誤が行われた後に論文投稿が行われる．このため，後発の手法が数値上スコアが良くなること自体は必然であり，真に先行研究より優れた手法なのか，後発であるがゆえの単なるチューニングの結果であるのか判断が難しい場合がある．これらの問題に対し，コンペティションでは同じタイミングで一斉に一発勝負の評価をするため，よりフェアで意味のあるスコア比較が可能である．

（2） 本当に「使える」技術がスポットを浴びる

学術論文として発表される手法は，技術的な新規性が求められるために得てして過度に複雑なものとなりやすく，実際には現実的な問題では利用困難なことも多い．また，分野内で標準的となっているアプローチや，著名な研究グループが発表する研究が注目されやすい傾向にある（本来あってはならないことだが）．このため，通常の論文査読をベースとしたエコサイクルからは，真に有用性の高い技術が必ずしも発見できない場合がある．

（注 3） http://mscoco.org/dataset/
（注 4） http://places2.csail.mit.edu/challenge.html
（注 5） http://www.thumos.info/
（注 6） http://activity-net.org/challenges/2016/
（注 7） http://www.visualqa.org/challenge.html
（注 8） https://sites.google.com/site/describingmovies/
（注 9） http://www.imageclef.org/
（注 10） http://www.multimediaeval.org/mediaeval2016/
（注 11） http://ms-multimedia-challenge.com/challenge
（注 12） http://www-nlpir.nist.gov/projects/trecvid/

このようなバイアスを取り去り，どんな方法であろうと，また無名の研究者であってもフェアにチャンスが与えられ，客観的に評価される場としてコンペティションは重要である．例えば，かつて画像認識では特徴量ベースの方法がもてはやされ，ニューラルネットワークはほとんど相手にされず論文が通らない時代があったが，ILSVRC で劇的な勝利を収めたことにより有無を言わさず業界の常識を塗り替えることとなった．

（3） タスク自体のメタ評価が可能

先に述べたとおり，コンペティションは様々な新しいタスクを試す実験場としての役割も持っている．各タスクに同じタイミングで取り組んだ研究者が一堂に会し，成果や知見・反省点を共有し，次の目標としてどのようなタスクやデータが適切か方向性を探ることで，研究分野全体を前進させていくループを構成することができる．

上記のようなメリットの一方で，コンペティションには弊害もある．参加者にとってコンペティションで上位成績を収めることには分かりやすい広告価値があるため，しばしば学術的興味を見失い，コンペティションで勝つこと自体が目的化しやすい(注13)．スコア向上には得てして既存手法の単なる組合せやチューニングに注力することが有効な場合が多いが，そのような参加チームばかりでは得るものの少ない退屈なイベントとなってしまうであろう．

5. まとめと今後の展望

クラウドソーシングを基盤としたアノテーション技術の向上により，大規模な教師付き画像データセットが次々と作られるようになった．一方で，深層学習の確立により，複雑な入出力関係をエンドツーエンドにモデル化し学習させることが可能となった．このように，学習のためのデータと手法の両輪が発展し，これを支える計算機環境も大幅に向上したことで質的な変化が起き，画像認識分野において現在カンブリア爆発のような状態を生んでいると言える．コンペティションはその重要な舞台装置となっており，適切な目標設定ができれば，驚くほどの速さで分野の技術向上をもたらすことができる．ImageNet はその最も顕著な成功例であると言える．

しかしながら，次々と提案される新しいタスクについて，その問題設定が本当に良いものなのかは常に考える必要がある．例えば，画像説明文生成タスクでは，表面的には高度な説明文が生成されているように見えるものの，言語モデルによって生成される文の 60〜70% が訓練データ中に存在し(38)，本質的には単純な最近傍探索（すなわち検索）に近いのではないかという指摘もある．事実，最近傍法に基づく手法もコンペティションでは驚くほど良い性能を出しており，単純にこのタスクのスコアを上げることが真の意味で画像の詳細な理解に近付いていると言えるかどうかは議論の余地があるだろう．このような知見も，より地に足の着いた画像理解を行うために，画像中の複数物体の属性や関連性の詳細なアノテーションを有する Visual Genome 等の次世代のデータセットを作り出すモチベーションになっていると考えられる．今後，これを利用したコンペティションも増えてくると予想される．

コンペティションに参加する場合は，単にスコアレースに終始するだけではなく，そのタスクの背後にある哲学は何か，本質的に困難問題は何か，そのために次は何が必要になるか，というサイエンスの営みを意識しながら取り組むことが重要であろう．

（注 13） 2015 年には，あるグループが不正な方法で評価サーバへの提出を非常に多い回数行い，事実上テストデータへのチューニングを行っていたことが判明し，1 年間コンペティションへの参加を禁止される事態となった．評価方法の意味を十分に理解しないまま目先のスコア向上にとらわれたことが原因であろう．

文　献

（1） K. Grauman and B. Leibe, Visual object recognition, Morgan & Claypool Publishers, 2011.

（2） J. Ponce, T.L. Berg, M. Everingham, D.A. Forsyth, M. Hebert, S. Lazebnik, M. Marszalek, C. Schmid, B.C. Russell, A. Torralba, C.K.I. Williams, J. Zhang, and A. Zisserman, “Dataset issues in object recognition,” Toward Category-Level Object Recognition, vol. 4170, pp. 29-48, Springer-Verlag Berlin Heidelberg, 2006.

（3） 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,” Comput. Vis. Image Underst., vol. 106, no. 1, pp. 59-70, 2007.

（4） G. Griffin, A. Holub, and P. Perona, “Caltech-256 object category dataset,” Technical report, California Institute of Technology, no. CNS-TR-2007-001, pp. 1-20, 2007.

（5） M. Everingham, L.V. Gool, C.K.I. Williams, and J. Winn, “The PASCAL visual object classes（VOC）challenge,” Int. J. Comput. Vis., vol. 88, no. 2, pp. 303-338, 2010.

（6） M. Everingham, S.M.A. Eslami, L. Van Gool, C.K.I. Williams, J. Winn, and A. Zisserman, “The Pascal visual object classes challenge : A retrospective,” Int. J. Comput. Vis., vol. 111, no. 1, pp. 98-136, 2014.

（7） A. Oliva, W. Hospital, and L. Ave, “Modeling the shape of the scene : A holistic representation of the spatial envelope,” Int. J. Comput. Vis., vol. 42, no. 3, pp. 145-175, 2001.

（8） S. Lazebnik, C. Schmid, and J. Ponce, “Beyond bags of features : Spatial pyramid matching for recognizing natural scene categories,” Proc. IEEE CVPR, 2006.

（9） L.J. Li and L. Fei-Fei, “What, where and who? Classifying event by scene and object recognition,” Proc. IEEE ICCV, 2007.

（10） J. Winn, A. Criminisi, and T. Minka, “Object categorization by learned univeral dictionary,” Proc. IEEE ICCV, vol. 2, pp. 1800-1807, 2005.

（11） A. Torralba and A.A. Efros, “Unbiased look at dataset bias,” Proc. IEEE CVPR, pp. 1521-1528, 2011.

（12） A. Sorokin and D. Forsyth, “Utility data annotation with Amazon Mechanical Turk,” Proc. the 1st IEEE Workshop on Internet Vision at CVPR, pp. 1-8, 2008.

（13） J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet : A large-scale hierarchical image database,” Proc. IEEE CVPR, pp. 248-255, 2009.

(14) C. Fellbaum, WordNet : An Electronic Lexical Database, MIT Press, 1998.
(15) T.-Y. Lin, M. Maire, S. Belongie, L.D. Bourdev, R.B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C.L. Zitnick, "Microsoft COCO : Common objects in context," Proc. ECCV, pp. 740-755, 2014.
(16) C. Rampf, B. Villone, and U. Frisch, "Microsoft COCO captions : Data collection and evaluation server," arXiv preprint, arXiv : 1504.00325, 2015.
(17) J. Xiao, J. Hays, K.A. Ehinger, and A. Torralba, "SUN database : Large-scale scene recognition from abbey to zoo," Proc. IEEE CVPR, pp. 3485-3492, 2010.
(18) B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," Proc. NIPS, pp. 487-495, 2014.
(19) A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," Proc. IEEE CVPR, pp. 1725-1732, 2014.
(20) F.C. Heilbron, V. Escorcia, B. Ghanem, J.C. Niebles, and U. Norte, "ActivityNet : A large-scale video benchmark for human activity understanding," Proc. IEEE CVPR, pp. 961-970, 2015.
(21) A. Rohrbach, M. Rohrbach, N. Tandon, and B. Schiele, "A dataset for movie description," Proc. IEEE CVPR, pp. 3202-3212, 2015.
(22) J. Xu, T. Mei, T. Yao, and Y. Rui, "MSR-VTT : A large video description dataset for bridging video and language," Proc. IEEE CVPR, 2016.
(23) A. Agrawal, J. Lu, S. Antol, M. Mitchell, C.L. Zitnick, D. Batra, and D. Parikh, "VQA : Visual question answering," Proc. IEEE ICCV, pp. 2425-2433, 2015, http://arxiv.org/abs/1505.00468
(24) Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei, "Visual7W : Grounded question answering in images," Proc. IEEE CVPR, pp. 4995-5004, 2016.
(25) R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalanditis, L.-J. Li, D.A. Shamma, M.S. Bernstein, L. Fei-Fei, Y. Kalantidis, L.-J. Li, D.A. Shamma, M.S. Bernstein, and F.-F. Li, "Visual Genome : Connecting language and vision using crowdsourced dense image annotations," arXiv preprint, arXiv : 1602.07332, 2016.
(26) N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," Proc. IEEE CVPR, pp. 886-893, 2005.
(27) G. Csurka, C.R. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," Proc. ECCV Workshop on Statistical Learning in Computer Vision, 2004.
(28) P.F. Felzenszwalb, R.B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models," IEEE Trans. Pattern Anal. Mach. Intell., vol. 32, no. 9, pp. 1627-1645, 2009.
(29) 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," Int. J. Comput. Vis., vol. no. 3, pp. 1-42, 2015.
(30) F. Perronnin, J. Sánchez, and T. Mensink, "Improving the Fisher kernel for large-scale image classification," Proc. ECCV, part. IV, pp. 143-156, 2010.
(31) J. Wang, J. Yang, K. Yu, F. Lv, T.S. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," Proc. IEEE CVPR, pp. 3360-3367, june 2010.
(32) H. Nakayama, T. Harada, and Y. Kuniyoshi, "Global Gaussian approach for scene categorization using information geometry," Proc. IEEE CVPR, pp. 2336-2343, 2010.
(33) A. Krizhevsky, I. Sutskever, and G.E. Hinton, "ImageNet classification with deep convolutional neural networks," Proc. NIPS, pp. 1097-1105, 2012.
(34) R. Girshick, J. Donahue, T. Darrell, U.C. Berkeley, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," Proc. IEEE CVPR, pp. 580-587, 2014.
(35) S. Ioffe and C. Szegedy, "Batch normalization : Accelerating deep network training by reducing internal covariate shift," Proc. ICML, vol. 37, pp. 448-456, 2015.
(36) 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.
(37) K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," Proc. IEEE CVPR, pp. 770-778, 2016.
(38) J. Devlin, H. Cheng, H. Fang, S. Gupta, L. Deng, X. He, G. Zweig, and M. Mitchell, "Language models for image captioning : The Quirks and what works," Proc. ACL, pp. 100-105, 2015.

（平成 28 年 12 月 3 日受付　平成 28 年 12 月 26 日最終受付）

中山（なかやま）　英樹（ひでき）（正員）

2006 東大・工・機械情報卒．2011 東大大学院情報理工学系研究科知能機械情報学専攻博士課程了．博士（情報理工学）．2008～2011 日本学術振興会特別研究員（DC1）．2012 から東大大学院情報理工学系研究科創造情報学専攻講師．2015 から産総研人工知能研究センター客員研究員を兼務．マルチメディアを中心としたデータマイニング，機械学習手法と応用アプリケーションの研究に従事．PRMU 研究奨励賞，情報処理学会全国大会奨励賞，計測自動制御学会 SI 部門賞若手奨励賞等を各受賞．