解説

エンドツーエンド深層学習のフロンティア

Frontier of End-to-End Deep Learning

西田京介　井島勇祐　田良島周平

Abstract

本稿では，ニューラルネットワークの基礎的な知識を有する読者を想定し，エンドツーエンド深層学習の最新動向について解説する．エンドツーエンド深層学習は，タスクに対して適切な構造を持つ一つのニューラルネットワークにより入出力関係を直接学習するものであり，従来取り組まれてきた，人間により設計された特徴量による学習に比べて高い精度を実現できた例が多数報告されている．本稿では，まず自然言語処理，音声処理，画像処理を中心に研究の動向について紹介した後に，各分野の動向に基づいて，エンドツーエンド学習がどのようなケースで有効に働くかについて考察を述べる．

キーワード：深層学習，エンドツーエンド，ニューラルネットワーク

1. はじめに

近年の人工知能技術の目覚ましい発展の要因の一つとして，エンドツーエンド（end-to-end）深層学習の実現が挙げられる．エンドツーエンド深層学習とは，入力データが与えられてから結果を出力するまで多段の処理を必要としていた機械学習システムを，様々な処理を行う複数の層・モジュールを備えた一つの大きなニューラルネットワークに置き換えて学習を行うものである．自動運転を例に取ると，非エンドツーエンドのアプローチでは，物体認識，レーン検出，経路プランニング，ステアリング制御など，人間が設定した複数個のサブタスクを解く必要があるところ，エンドツーエンド学習では図1のように車載カメラから取得した画像から直接ステアリング操作を学習する[(1)]．

このような学習が可能になった背景には，深層学習の発展と，学習データ量の増加が挙げられる．まず，深層学習の発展について，従来のアプローチと比較して説明する．従来は，タスク全体の入力データを出力データに変換するにあたって，データの中間的な表現，すなわち，多段に接続される各モジュールの入出力を人間が設計する必要があった．このような中間表現の設計は，特徴設計（feature engineering）と呼ばれ，精度に大きく影響する．その一方で，タスクに対して適切な構造を持つ多層のニューラルネットワークを用いてタスク全体の入出力を直接学習すると，人間が設計するよりも優れた様々な粒度の中間表現を自動的に獲得できる場合がある．このような能力の獲得は表現学習と呼ばれ，深層学習を採用するメリットとして知られている[(2)]．次に，クラウドソーシングの発展や，Web サービス・IoT サービスの普及等によって，大規模データセットの収集・作成が容易になったことが，エンドツーエンド深層学習の発展につながっていると考える．その一方で，現状全てのタスクがエンドツーエンド学習で取り組まれているわけではなく，例えば自動運転については個別にタスクを解くアプローチも依然として多く取り組まれている．本稿では，まず2.において各分野におけるエンドツーエンド深層学習の動向について説明した後，3.においてその可能性と課題について論じる．

2. 各分野における動向

2.1 自然言語処理

従来の自然言語処理においては，入力テキストのトー

西田京介　正員　日本電信電話株式会社 NTT メディアインテリジェンス研究所
E-mail nishida.kyosuke@lab.ntt.co.jp
井島勇祐　正員　日本電信電話株式会社 NTT メディアインテリジェンス研究所
E-mail ijima.yusuke@lab.ntt.co.jp
田良島周平　日本電信電話株式会社 NTT メディアインテリジェンス研究所
E-mail tarashima.shuhei@lab.ntt.co.jp
Kyosuke NISHIDA, Yusuke IJIMA, Members, and Shuhei TARASHIMA, Nonmember（NTT Media Intelligence Laboratories, NIPPON TELEGRAPH AND TELEPHONE CORPORATION, Yokosuka-shi, 239-0847 Japan）.
電子情報通信学会誌　Vol.101 No.9 pp.920-925 2018年9月

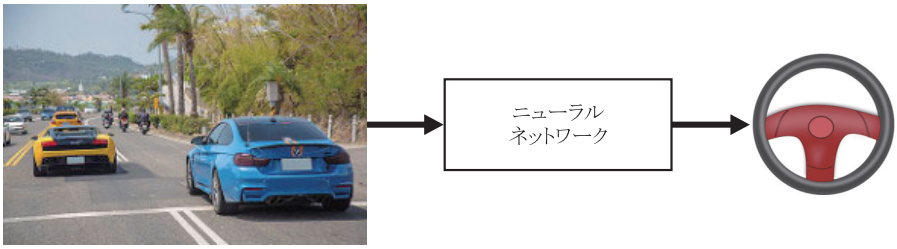


図1　自動運転におけるエンドツーエンド深層学習の概要　物体認識や歩行者検出などの個別モジュールを用いず，車載カメラの画像からステアリング操作を直接出力するニューラルネットワークを学習する．

クナイズ（形態素解析）に加えて，係り受け解析や品詞解析，固有表現抽出などの学習済みツールの出力を補助的な特徴量とし，機械翻訳や質問応答などのタスクの学習に利用していた．近年の深層学習による自然言語処理では，トークナイズのみ別途実施して，他の中間表現についてはエンドツーエンド学習によって暗に獲得する場合が多い．これは，Word2Vec[3]などの単語の分散表現を獲得する技術の発展により，単語の意味については大規模コーパスから学習した方が高い精度が得られる場合が多いためである．その一方で，トークナイズも不要とすべく，文字ベースで自然言語処理タスクに取り組むニューラルネットワーク[4]も増加している．

自然言語処理分野において，エンドツーエンド深層学習が成功した代表例は機械翻訳である．従来のフレーズベースの統計機械翻訳においては，元文と翻訳文のペアのコーパスデータから，文内の単語の対応（アラインメント）の推定，推定結果に基づくフレーズテーブル（翻訳モデル）の構築，言語モデルの構築，翻訳モデルと言語モデルからの翻訳候補の探索，など複数のステップに分割し，それぞれのステップに対して個別の研究が行われていた．これに対して，ニューラル機械翻訳では一つのニューラルネットワークにより翻訳を行うモデルが提案され，統計ベースの方法よりも高い精度が報告されている．

図2に示すニューラル機械翻訳モデルの例では，エンコーダ—デコーダ（符号器—復号器）と呼ばれる入力された元文をRNN[用語]により固定長のコンテキストベクトルに変換（エンコード）し，コンテキストベクトルを基にRNNにより翻訳文を生成（デコード）するモデルを連結した一つのネットワークとして扱う[5]．まず，エンコーダは長さTの元文の入力ベクトル系列$\boldsymbol{x}=\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_T\}$を受け取り，RNNを用いて隠れ状態系列$\{\boldsymbol{h}_1,\cdots,\boldsymbol{h}_T\}$に変換する．そして，入力長分の隠れ状態ベクトルを，非線形変換qにより固定長ベクトル$\boldsymbol{c}$に変換する．デコーダは，エンコーダが出力したコンテキストベクトル$\boldsymbol{c}$と，$t-1$番目までの出力単語系列$(y_1,\cdots,y_{t-1})$を用いてt番目の出力単語y_tを予測する．このようなネットワーク構造とすることで，対訳文のペアから直接学習を行うことができる．

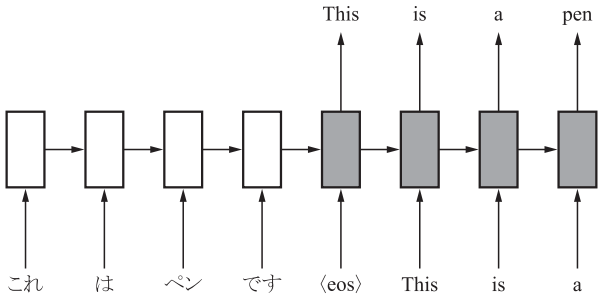


図2　エンコーダ—デコーダモデルによる機械翻訳の最も単純な例　二つのRNNを連結したネットワークによりエンドツーエンドに翻訳を行う．

更に，ニューラル機械翻訳の精度を大きく改善した要因として，アテンション機構の導入が挙げられる[6]．アテンション機構では，入力側のコンテキストベクトルを一つの固定長ベクトル$\boldsymbol{c}$に変換するのではなく，翻訳文の単語位置tに応じて異なるベクトル$\boldsymbol{c}_t=\sum_{i=1}^{T}\alpha_{ti}\boldsymbol{h}_i$を計算する．ここで，$\alpha_{ti}$はアテンションの重みを表しており，翻訳文の$t$番目の要素を生成する際に，入力文に対してどのような重み付けを行うかを決定する役割を担っている．これは，t番目の単語を生成する際に入力文のどの単語に着目すればよいかという入力文と翻訳文のソフトなアライメント（単語の対応付け）情報も同時に学習していると捉えることができる．このように，エン

用　語　解　説

RNN　リカレントニューラルネットワーク（Recurrent Neural Network）の略．主に系列データ処理に利用．内部状態及びフィードバックループを持つことが特徴．

CNN　畳込みニューラルネットワーク（Convolutional Neural Network）の略．主に画像処理に利用．局所特徴を抽出する畳込み層と位置・回転に不変性を与えるプーリング層を持つ．

ドツーエンド学習においても，従来人間が設計してきた特徴量に相当するような情報を表現可能なネットワークを利用することが精度向上には重要であり，このようなネットワークアーキテクチャの検討はアーキテクチャ設計（architecture engineering）などと呼ばれる．

ほかにも，対話システム[7]，質問応答[8]など，エンドツーエンド深層学習により精度向上したタスクが多く存在し，ネットワークアーキテクチャの検討や，更なるデータセットの整備が進んでいる．

2.2 音声処理

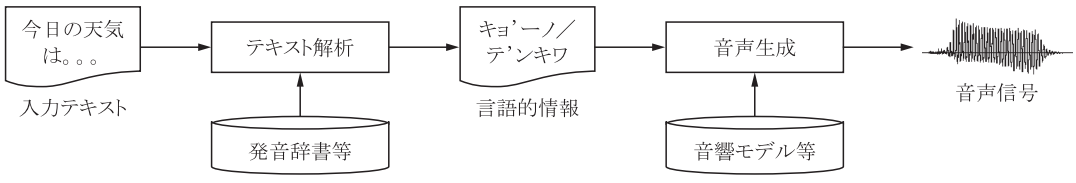
音声からテキストへ変換する音声認識（ASR：Automatic Speech Recognition），任意のテキストから対応する音声を生成するテキスト音声合成（TTS：Text-to-Speech synthesis）等に代表される音声処理においても，エンドツーエンド深層学習による性能向上が報告されている．本節では，エンドツーエンド深層学習により大きな性能向上が得られているテキスト音声合成に焦点を当てる．

テキスト音声合成は任意のテキストから音声を生成する処理であるが，言い換えれば，離散値の時系列情報であるテキストと連続値の時系列情報である音声信号（若しくは音声信号から得られる音響特徴量）との対応付けの問題と捉えることができる．この問題に対し，従来のテキスト音声合成では，テキストと音声信号との対応付けを直接学習するのではなく，テキストと中間情報である言語的情報（日本語の場合，読み，アクセント，間の位置等）との対応付けを行うテキスト解析部，言語的情報と音声信号（音響特徴量）との対応付けを行う音声生成部から構成され，この二つの構成要素を独立に学習している（図3(a)）．各要素の詳細については，過去の小特集を参照頂きたい[9],[10]．

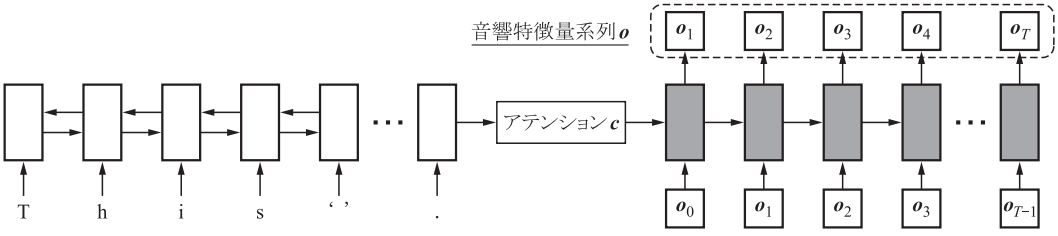
これに対し，エンドツーエンド音声合成[11]～[13]では，テキストと音声信号との対応付けを単一のニューラルネットワークで直接学習する．図3(b)にエンドツーエンド音声合成のモデル構造の例を示す．これは前節で述べたニューラル機械翻訳モデルと同様にアテンション機構[6]付のエンコーダ―デコーダモデルで構成されている．まず，エンコーダでは，文字単位に分割された入力テキストを中間表現であるコンテキストベクトルへ変換する．デコーダでは，エンコーダで得られたコンテキストベクトルを基に音声信号（音響特徴量）を生成する．

エンドツーエンド音声合成の利点は，音声合成器の構築コストの低減であると考えられる．従来のテキスト音声合成は，テキスト解析部，音声生成部について，言語ごとに専門家による規則，発音辞書，学習データ（テキスト，音声の収集とそれらに対する人手でのアノテーション）の整備が必要であるため，新しい言語のテキスト音声合成器の構築コストは非常に高い．それに対し，エンドツーエンド音声合成では，学習データは大量のテキストとそれを発声した音声のみであるため，従来と比べて学習データの整備コストを大幅に低減することが可能である．

一方，実用的なテキスト音声合成では，テキストとして特殊な発音を持つ人名，地名等の固有名詞が入力された場合でも，正しい発音の音声の生成ができる，若しくは発音誤りがあっても辞書等で修正ができることが求められるが，エンドツーエンド音声合成ではそのような保証がない．また，現在報告されているエンドツーエンド音声合成は入力される文字のバリエーションが数十種類である英語（アルファベット，記号）が主であるため，バリエーションが数千から数万種類である日本語，中国



(a) 従来のテキスト音声合成の例



(b) エンドツーエンド音声合成の例

図3 テキスト音声合成の例

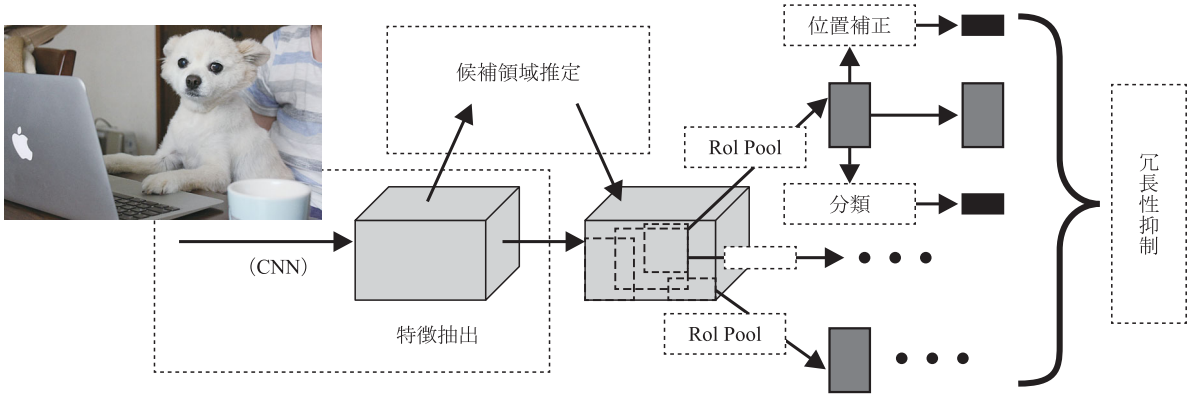


図4　エンドツーエンド物体検出の例

語等で頑健に動作するかは検証されておらず，言語によらず頑健に動作する枠組みの検討も必要であると考えられる．

本節では音声処理のエンドツーエンド深層学習の例としてテキスト音声合成を解説したが，ほかにも音声認識[14]，感情認識[15]等の単一のタスクに加えて，音声認識と機械翻訳を単一のニューラルネットワークで最適化するエンドツーエンド音声翻訳[16]等の複数のタスクを同時に処理する手法も提案されており，更に多様なタスクでの検討が進むと考えられる．

2.3　画像処理

画像分野でも，複雑なタスクを相対的に容易なサブタスク群に分解し，個々を独立に解く従来のアプローチの多くが，深層学習ベースの手法に置換されつつある．本節ではエンドツーエンド深層学習の適用が積極的に進んでいる画像タスクの例として，物体検出と自己位置推定を取り上げる．

まず物体検出は，入力画像中に任意の数写り込む物体個々のクラスと位置（多くの場合，方形）を認識するタスクである．方法論は幾つか存在するが，タスク全体を位置検出と領域の分類とに分解して解くアプローチは，直感的でかつ高性能なことが多く，広く用いられている．従来は各サブタスクが独立に解かれてきた．しかし昨今では，各要素処理がニューラルネットワーク化され，それらがエンドツーエンドに最適化されている．

図4はエンドツーエンド物体検出の一例である．まず入力画像は，画像分類用途に事前学習されたCNN[用語]の中間層まで順伝搬され，特徴マップが生成される（特徴抽出）．この共通の特徴マップから，物体の写っていそうな領域の候補を複数抽出する（候補領域推定）[17]とともに，領域ごとの特徴マップを獲得する（RoI Pool）[18]．最終的に，得られた領域特徴から物体の分類と方形位置の補正がなされる（分類・位置補正）．各モジュールはいずれも微分可能であり，学習データ所与の下，全体はエンドツーエンドに最適化される[注1]．

エンドツーエンド物体検出は従来の手法に比べ，検出精度だけでなく，テスト処理のスループットも極めて高い[19]．これは，画像当り一度のみ実行される特徴抽出処理が，その他の処理の多くで共有されていることに起因している．エンドツーエンド深層学習の導入によるサブタスク間でのパラメータ共有が，処理効率化につながった好例であると言えよう．また最近では，アテンション機構を用いて冗長な候補領域を抑制する手法[20]等も提案されている．通常は後処理とみなされるサブタスクもエンドツーエンドの中に取り込むことで，従来の物体検出が内包するヒューリスティックに設定されてきたパラメータの排除も進みつつある．

次に画像からの自己位置推定は，単一画像あるいは映像を入力として，撮影したカメラの位置姿勢（ポーズ）を推定するタスクである．拡張現実（AR）やナビゲーション，地図構築に応用されることが多い．自己位置推定もまた，単一画像を入力としたカメラ姿勢推定，画像ペア間を入力としたカメラ姿勢変化推定，映像全体を考慮したカメラ姿勢推定と3Dマップ構築，自己位置を正確かつ効率的に推定するための行動の提案等から構成される複合的なタスクである．従来はそれらを構成する処理が個々に設計・最適化されてきたが，エンドツーエンド深層学習は，それらの統合的最適化を実現しつつある．

例えば文献(21)では，事前学習されたCNNのファインチューニングにより，単一の画像入力から6自由度のポーズを回帰する手法が提案されている．この手法では，学習データを屋内外シーンを捉えた2Dの画像群から3Dマップを構築することで，各2D画像のポーズを自動付与して構築することが特徴的であり，従来の特徴設計に基づく方法が不得手なケース（e.g. ぶれ・動物体

（注1）　正確には，候補領域推定とそれ以外は交互に，分類誤差と位置回帰誤差を同時に最小化するマルチタスク学習で最適化される．

の存在・天候の変化）における優位性が報告されている．また文献(22)では，時間的コンテキストを考慮した，画像ストリームからのエンドツーエンド自己位置推定手法が提案されている．この手法では，未知環境であっても頑健にポーズ推定をするため，画像ペアの動きベクトルを推定するCNN[23]を流用している点が興味深い．更に最近では，例えば最小ステップで自己位置を特定するための行動の獲得[24]など，他のサブタスク群のエンドツーエンド化の実現も提案されつつある．これらの融合が進めば，全てのサブタスクが統合的に最適化される「完全な」エンドツーエンド自己位置推定の実現も，そう遠い話ではないのかもしれない．

2.4 その他

（1） 自動運転

従来はmediated perceptionと呼ばれる物体認識やレーン検出などのサブタスクを解くアプローチが主流であったが，画像を入力として直接ステアリング制御を予測するエンドツーエンドモデルが近年増加している[25]．エンドツーエンドモデルでは予測したステアリング操作の根拠を示すことが難しいが，Kimらはアテンション機構を備えたエンコーダ―デコーダにより，精度を犠牲にすることなく解釈性を備えたモデルを提案している[1]．また，カリフォルニア大学バークレー校から1万時間を超える大規模データセットが近年公開されたように[26]，データ量の問題についても解決の兆しがある．今後のエンドツーエンド学習に基づく自動運転の進展が期待される．

（2） マルチモーダル処理

画像やテキストなど複数種類のデータを理解する必要があるタスクにおいては，適切な中間表現の設計が非常に難しく，エンドツーエンド学習の適用が非常に有効な場合が多い．例えば，画像についてテキストで質問応答を行う視覚的質問応答のタスクでもCNNによる画像の理解とRNNによる質問・回答の理解・生成をエンドツーエンドに訓練可能なモデル[27]が提案されている．その一方で，個別に大量のデータから学習した物体認識・シーン分析などのモジュールが出力した結果を利用することで，視覚的質問応答においてエンドツーエンドアプローチよりも高い精度が得られた例が最近報告されている[28]．

3. おわりに：発展の可能性と課題

本稿では，エンドツーエンド学習に関する動向について示した．エンドツーエンド学習は，大量の学習用データが存在する状況においては非常に有効な手段である．特に，入出力関係が複雑なタスクであり，適切なサブタスクへの分割及びサブタスクごとの学習データの準備が難しい場合には大きな優位性を持つ．ただし，エンドツーエンドアプローチにおいても，従来の特徴設計のようにネットワークアーキテクチャの設計は重要であり，単純にRNN，CNNを適用するだけでは，大量のデータがあったとしても高い精度を得られない場合が多い．また，ネットワークのサイズや学習方法等を定義するパラメータチューニングはエンドツーエンド学習でも依然として必要であり，大きな課題となっている．

また，自分の解きたいタスクの学習データが十分にある場合でも，他の関連するタスクの学習データが多く存在する場合は，事前学習，あるいはマルチタスク学習や転移学習の利用により更に精度が改善される例が報告されている．特に，画像を扱うネットワークにおいては，ImageNetなどにより事前学習したCNNを利用することが有効である．

その他，モジュール化して個別に学習を行うアプローチにおいては，中間表現が人間が設計したものとなるため，結果の解釈性について利点がある．データの重要な部分に注目するアテンション機構の利用による可視化などについては研究が進展しているが，解釈性については深層学習分野全体の課題となっており，今後の研究の進展が期待される．

文　　献

（1） J. Kim and J.F. Canny, “Interpretable learning for self-driving cars by visualizing causal attention,” ICCV, pp. 2961-2969, 2017.

（2） Y. Bengio, A.C. Courville, and P. Vincent, “Representation learning : A review and new perspectives,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 35，no. 8，pp. 1798-1828, 2013.

（3） T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” NIPS, pp. 3111-3119, 2013.

（4） Y. Kim, “Convolutional neural networks for sentence classification,” EMNLP, pp. 1746-1751, 2014.

（5） K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, “Learning phrase representations using rnn encoder-decoder for statistical machine translation,” EMNLP, pp. 1724-1734, 2014.

（6） D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” ICLR, 2015.

（7） I.V. Serban, A. Sordoni, Y. Bengio, A.C. Courville, and J. Pineau, “Building end-to-end dialogue systems using generative hierarchical neural network models,” AAAI, pp. 3776-3784, 2016.

（8） S. Sukhbaatar, A. Szlam, J. Weston, and R. Fergus, “End-to-end memory networks,” NIPS, pp. 2440-2448, 2015.

（9） 匂坂芳典，“コーパスベース音声合成技術の動向［I］―コーパスベース音声合成の過去・現在・将来―，” 信学誌，vol. 87, no. 1, pp. 64-69, Jan. 2004.

(10) 小林隆夫，“小特集にあたって(〈小特集〉音声合成に関する研究の動向)，” 音響誌，vol. 67, no. 1, pp. 15-16, 2010.

(11) Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R.J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q.V. Le, Y. Agiomyrgiannakis, R. Clark, and R.A. Saurous, “Tacotron : Towards end-to-end speech synthesis,” INTERSPEECH, pp. 4006-4010, 2017.

(12) W. Ping, K. Peng, A. Gibiansky, S.O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, “Deep voice 3 : 2000-speaker neural text-to-speech,” CoRR, 1710.07654, 2017.

(13) J. Shen, R. Pang, R.J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, R.A. Saurous, Y. Agiomyrgian-

nakis, and Y. Wu, “Natural tts synthesis by conditioning wavenet on mel spectrogram predictions,” CoRR, 1712.05884, 2017.
(14) A. Graves and N. Jaitly, “Towards end-to-end speech recognition with recurrent neural networks,” ICML, pp. 1764-1772, 2014.
(15) G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M.A. Nicolaou, B. Schuller, and S. Zafeiriou, “Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network,” ICASSP, pp. 5200-5204, IEEE, 2016.
(16) A. Bérard, O. Pietquin, L. Besacier, and C. Servan, “Listen and translate : A proof of concept for end-to-end speech-to-text translation,” NIPS Workshop on end-to-end learning for speech and audio processing, 2016.
(17) S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN : towards real-time object detection with region proposal networks,” NIPS, pp. 91-99, 2015.
(18) J. Dai, Y. Li, K. He, and J. Sun, “R-FCN : Object detection via region-based fully convolutional networks,” NIPS, pp. 379-387, 2016.
(19) B. Singh, H. Li, A. Sharma, and L.S. Davis, “R-FCN-3000 at 30 fps : Decoupling detection and classification,” CVPR, 2018（to appear）.
(20) H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei, “Relation networks for object detection,” CVPR, 2018（to appear）.
(21) A. Kendall, M. Grimes, and R. Cipolla, “PoseNet : A convolutional network for real-time 6-DOF camera relocalization,” ICCV, pp. 2938-2946, 2015.
(22) S. Wang, R. Clark, H. Wen, and N. Trigoni, “DeepVO : Towards end-to-end visual odometry with deep recurrent convolutional neural networks,” ICRA, pp. 2043-2050, 2017.
(23) A. Dosovitskiy, P. Fischer, E. Ilg, P. Häusser, C. Hazırbaş, and V. Golkov, “FlowNet : Learning optical flow with convolutional networks,” ICCV, pp. 2758-2766, 2015.
(24) D.S. Chaplot, E. Parisotto, and R. Salakhutdinov, “Active neural localization,” ICLR, 2018.
(25) M. Bojarski, D.D. Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L.D. Jackel, M. Monfort, U. Muller, J. Zhang, X. Zhang, J. Zhao, and K. Zieba, “End to end learning for self-driving cars,” CoRR, 1604.07316, 2016.
(26) H. Xu, Y. Gao, F. Yu, and T. Darrell, “End-to-end learning of driving models from large-scale video datasets,” CVPR, pp. 3530-3538, 2017.
(27) M. Malinowski, M. Rohrbach, and M. Fritz, “Ask your neurons : A neural-based approach to answering questions about images,” ICCV, pp. 1-9, 2015.
(28) I. Ilievski and J. Feng, “Multimodal learning and reasoning for visual question answering,” NIPS, pp. 551-562, 2017.

（平成 30 年 3 月 23 日受付　平成 30 年 4 月 12 日最終受付）

西田　京介（正員）

2008 北大大学院情報科学研究科博士課程了．博士（情報科学）．2006–2009 JSPS 特別研究員．2009 日本電信電話株式会社入社．現在，NTT メディアインテリジェンス研究所主任研究員（特別研究員）．2017 DBSJ 上林奨励賞受賞．言語処理に関する研究開発に従事．

井島　勇祐（正員）

2015 東工大大学院総合理工学研究科博士課程了．博士（工学）．2009 日本電信電話株式会社入社．現在，NTT メディアインテリジェンス研究所研究主任．2016 本会音声研究会研究奨励賞，2018 日本音響学会粟屋潔学術奨励賞各受賞．音声合成に関する研究開発に従事．

田良島　周平

2011 東大大学院新領域創成科学研究科修士課程了．修士（環境学）．2011 日本電信電話株式会社入社．現在，NTT メディアインテリジェンス研究所研究員，東大大学院情報理工学系研究科博士課程在籍．画像認識の研究開発に従事．