Keywords you should know. 第150回

知っておきたいキーワード

End2End

正会員 猪飼知宏†

† シャープ株式会社 研究開発本部
"End2End" by Tomohiro Ikai (Corporate R&D BU, Sharp Corporation, Chiba)
キーワード：End2End，オートエンコーダ，深層学習画像符号化，Video Coding for Machines

End2Endとは

End2End[1]とは，英語の「エンド（端）からエンド（端）まで」に由来し，技術的には「入力から出力までを一貫して処理や設計，評価する」技術のことです．当たり前のことにもみえますが，いくつかの例をとって，本質をつかんでいきます．多くの領域に適用されている考え方ですが，ここでは深層学習を用いた画像，音声，言語といったメディア処理の立場から説明したいと思います．

図1に示すように，従来の画像検出は，個別に設計された多段の処理から構成されていました．画像から特徴量抽出と，特徴量の分類・識別から構成され，さらに各々の中でまた複数の処理から構成されます．それに対して，End2End処理では，一貫して設計された多段の処理として，画像から直接分類結果を得ます．このように入力端に元の入力，出力端に結果を設定し，元の入力から結果をできるだけ一貫して設計，評価する点が，End2Endの特徴です．

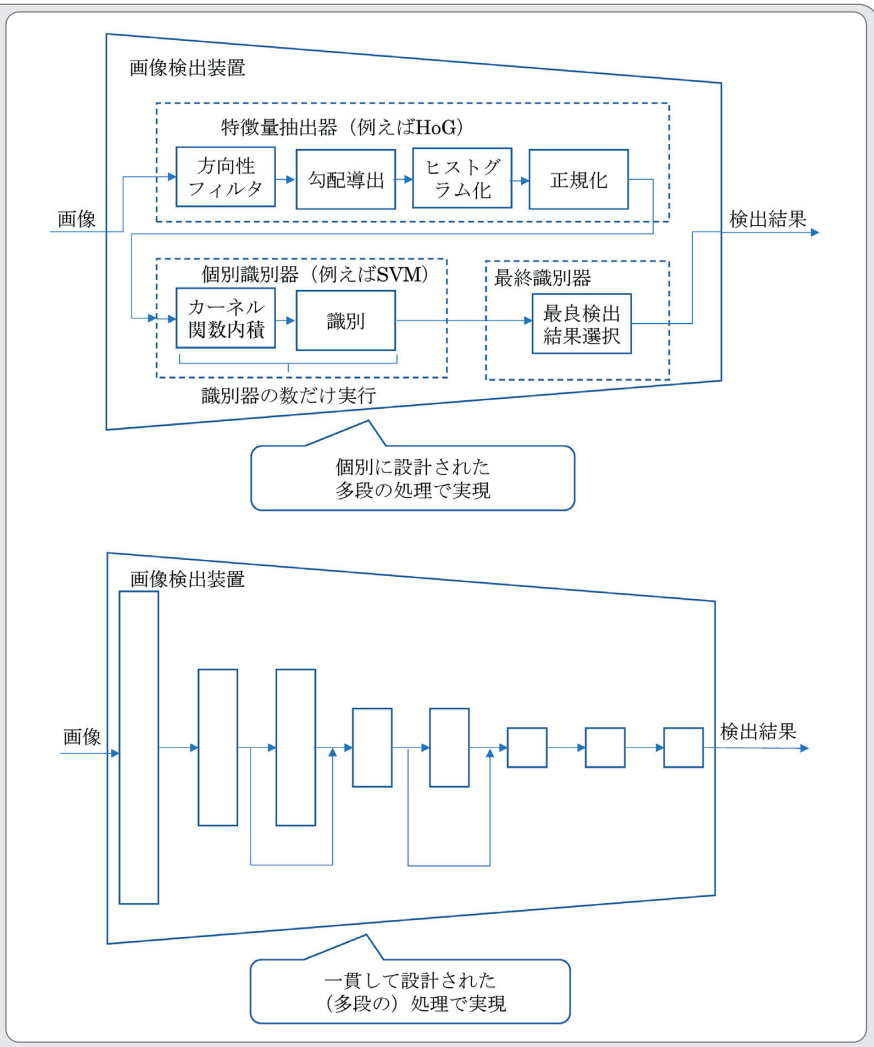

図1　従来型の画像検出（上）とEnd2End（下）の画像検出

映像情報メディア学会誌 Vol. 77, No. 1, pp. 58～61（2023）

画像，音声，言語＝人間には理解できない入出力関係

皆さんは，どうやって物を見て，音を聞き取り，話をしているのか理解していますか？「画像，音声，言語」はすべて多次元データであり，多次元データならではの大量のバリエーション（多様性）があります．さらに位置関係，前後関係，により意味が変化する多義性（さらなる組み合わせ爆発）もあります．図2に示すようにこれらメディアAB間の変換は人間にも理解できない自明でない入出力関係（図中の変換「???」）になり，入力と出力が多様な分，変換にも無数のパラメータが必要になります．このような場合に問題を分割せずにデータから学習できるEnd2Endはより力を発揮します．Diffusion Model（拡散モデル）等を用いてテキストから画像を生成する例は正に自明でない関係の代表例でしょう．

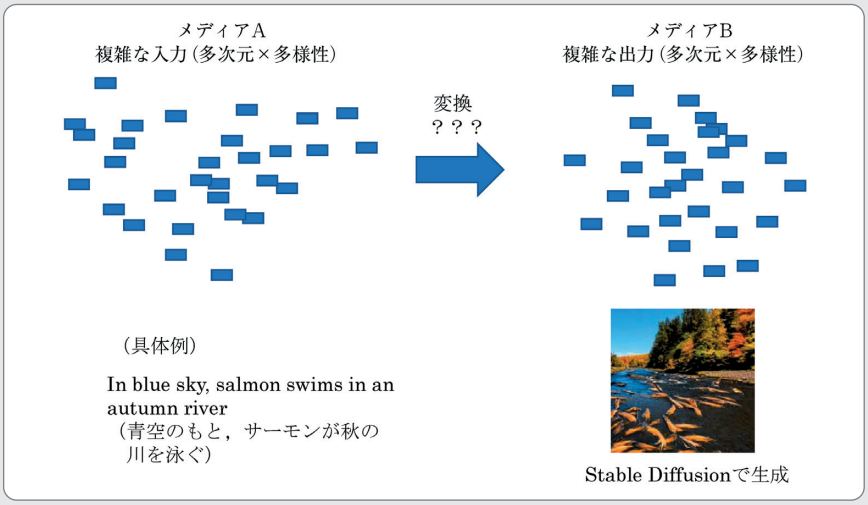

図2　多次元・多様性のある自明ではない関係

End2Endが高性能である理由

複雑な問題であるほど，中間的な要素の特定が困難で，部分問題に分割するのが難しくなります．部分問題の間で最適化ができない上に，処理の度に情報が失われますので，従来型は非効率になります．また複雑な問題＝多様性＝数ですから，「中間的な問題にも巨大なバリエーションの解答を解かせる」複雑さが生じます．End2Endでは，入力と出力の間の情報変化を無駄なく全体の学習が可能であるため高成能になります．

1980年代に任意の連続関数の近似能力（普遍性定理）が証明されたようにニュートラルネットワークはEnd2Endを実現するに足る表現能力があります．2010年代のネットワーク構造と学習上の成果，ResNet（Residual neural Network）等の差分接続，バッチ正則化等の安定性向上，CNN（Convolutional Neural Network）等のパラメータ共通化・近傍特徴，アテンション・Transformer等の大域的注視・系列間相関と，現在も続く半導体の進歩によって，今では1000億パラメータを超える巨大なネットワーク（＝巨大な演算器）でも学習可能になっています．現在では，どのようにデータから頑健な処理を学習するか，ラベルを必要としない教師なし手法などが課題になっています．表1に従来型処理とEnd2End型処理の比較をまとめました．

表1　従来型処理とEnd2End型処理

	構　成	利　点	欠　点
従来型処理	個別に設計された多段の処理から構成	・処理が明確　・再利用が容易 ・物理的な制約・関係を入れやすい	・性能に一定の限界がある
End2End処理	一貫して設計された多段の処理から構成	・高い性能を実現 ・データから直接学習できる ・入力と出力の関係（中身）がわからなくても学習できる	・結果に至る理由が不明確 ・設計（学習）が大変 ・再利用しにくい ・大量のデータが必要

End2Endの具体例

図3（左）に示すようにEnd2Endは入出力関係を実現しており，特に要素数やモーダルの違う表現なども実現することができます．End2Endの面白い実例は「オートエンコーダ」です．オートエンコーダはデータサーズを減少させた中間表現（潜在表現）を求めるもので，主成分分析，t-SNE（t-distributed Stochastic Neighbor Embedding）と同じく次元圧縮の一種です．図3（右）のように，オートエンコーダと呼ばれるデータから中間表現を求めるネットワークと，オートデコーダと呼ばれる中間表現から元のデータを直列に結び，両者を同時に学習します．

☞

End2End変換の例

カテゴリ	入出力関係（処理）	例	備考
回帰・判別・分類	画像から数値｜音声から数値	物体検出｜話者識別	
同じモーダル	画像から画像｜音声から音声	イラスト化｜日英翻訳	
別のモーダル	音声からテキスト｜テキストから音声	文字おこし｜音声合成	
	テキストから画像｜画像からテキスト	画像生成｜画像要約	DALL-E、Stable Diffusion
	画像から音声｜音声から画像	音声ナビ	
マルチモーダル	画像と音声から要約テキスト		
中間表現	画像から特徴量｜音声から特徴量	音響特徴量	オートエンコーダ
動作，環境インタラクション	画像・音声からハンドル操作	自動運転	強化学習
	盤面（駒・石・牌配置）から最善手	将棋・囲碁・麻雀	AlphaGo

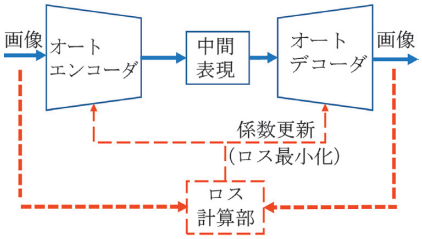

図3　多様なEnd2End変換の例（左）とオートエンコーダによる中間表現学習（右）

従来型処理とEnd2End

画像，音声，言語でも，End2Endを必要としない関係もあります．例えば，画像の回転，拡大・縮小，立体の射影変換，基本的なレンズの歪みは，アフィン変換で記述可能であり，従来型の処理で高精度に実現できます．音の数値化に必要な，音響特徴量のメルスペクトル導出も従来型を用いますし，言葉の数値ベクトル表現もWord2Vecなどで，独立して学習されたコーパスから得られた表現を用います．また，顔や人体をリアルタイムで検出する場合でも低レベル特徴量と識別器の組み合わせも現役です．解析用途では理解が容易な決定木が常に有力な選択肢であり，データによってはランダムフォレストやXGBoost（eXtreme Gradient Boosting）なども高性能を実現します．

End2End画像圧縮

End2Endの最新動向の一つは画像圧縮（画像符号化）です．JPEG，MPEG規格に代表される符号化技術では，図4（上）のように，予測と変換，量子化，エントロピー符号化，フィルタ処理を別個に設計し利用しています．例えば画像を1/30，映像を1/300に圧縮するためには，予測により周囲との冗長性を除去し，変換で予測で残る画素間の相関を除去し，エントロピー符号化で確率的な偏りに応じて無駄のない短い符号を割り当てます．理想的には「予測と変換，変換と符号化」を一度に設計したいところですが，容易ではありませんでした．深層学習によるEnd2Endはそれを可能にしました．つまり，図4（下）のように「画像を圧縮して符号を出し符号から復号画像に戻す」処理全体を一度につなげ，全体を設計することができています．2016年の誕生当初はDCT画像符号化を超えたことが評価されましたが，ついに最近，End2End画像圧縮は静止画で最新規格のVVC（Versatile Video Coding）を超える圧縮性能を実現し現在も研究が続いています．

End2Endの素晴らしいところは，人間の目に近い画質評価指標やGAN（Generative Adversarial Network）と呼ばれる評価用ネットワークも自由に接続し，同時に学習できる点です．また，現在の符号化はあまりにもモード数が増えすぎた結果，驚くほどの演算量を符号化側で必要としているのですが，End2End画像圧縮は，符号側の演算時間を短くできる可能性があります．JPEGではJPEG-AIでまさにEnd2End画像圧縮の規格化を開始し，MPEGでもVideo Coding for Machinesという名称で，画像と認識のマルチタスク符号化の規格化を検用しています．

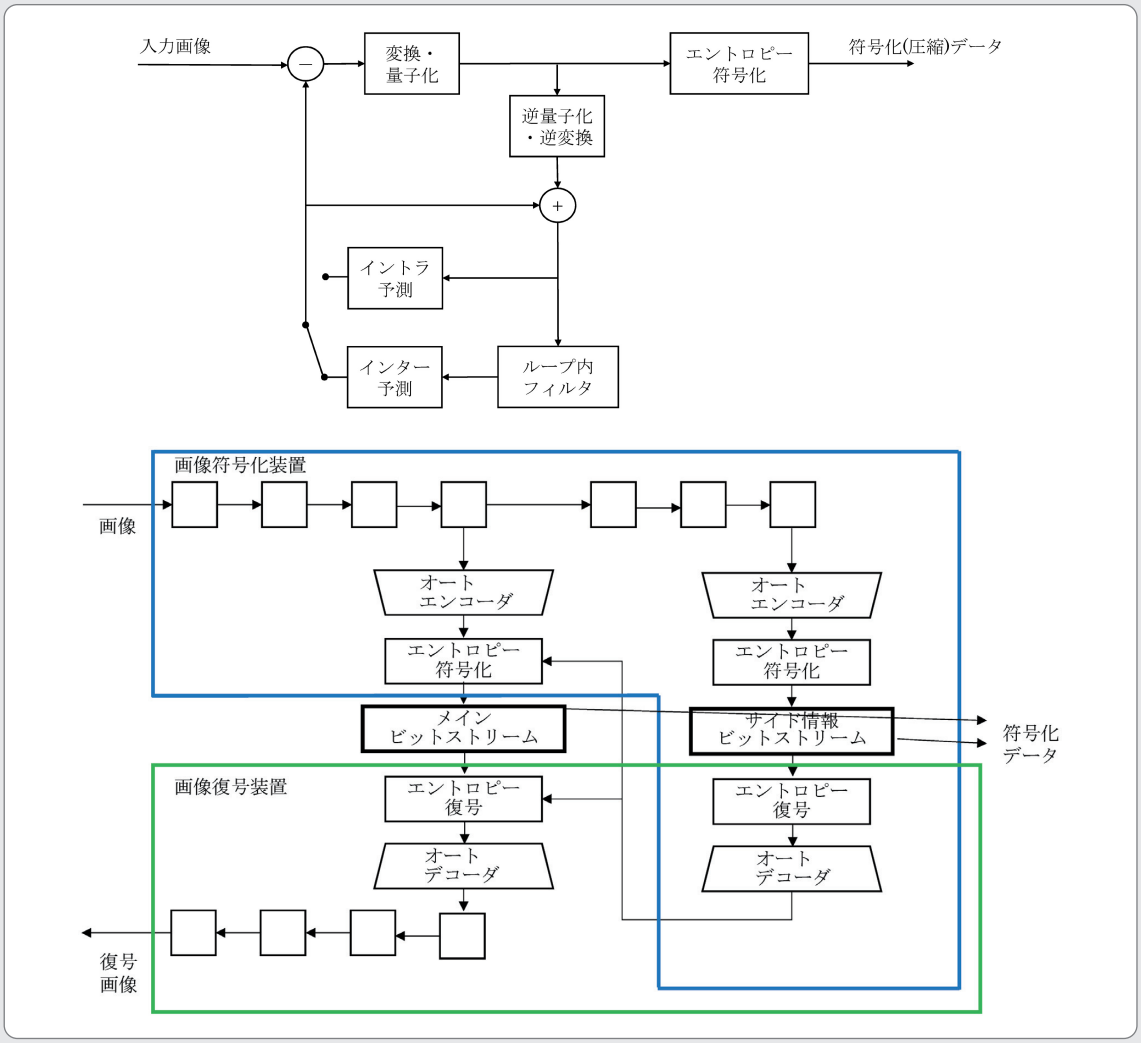

図4　従来技術の画像符号化装置（上）とEnd2End画像符号化・復号装置（下）

むすび

「部分部分に分解して処理をする」従来処理と「全体を一度に設計するEnd2End処理」は排他的なものではありません．一つには，音響特徴量や言語ベクトルなど前後処理における従来型処理の利用です．処理を分割することで設計が容易になりモジュールとして再利用されます．二つには，ニューラルネットワーク自体の設計においても，部分的な設計の概念が有効利用されています．つまり，中身がわからないまま学習できるはずのニューラルネットワークにおいても，完全にブラックボックスのまま，利用することはほぼなく，特定の構造を利用した効率的なネットワークが利用されています．例えば，画像の意味を利用するにはマルチ解像度の処理に適したU-Net構造を用いますし，言語などの時系列を扱う場合には，Encoder-Decoder構造により中間表現に一度変換します．近年では言語，画像，音声の処理においてTransformerによる系列間相関を利用するネットワークが利用されます．

このように，深層学習のネットワーク設計や，段階的な学習において，人間の手が加わっており，人間の知恵による進歩が続いています．

（2022年10月2日受付）

参　考　文　献

1）西田，井島，田良：“エンドツーエンド深層学習のフロンティア”，信学誌，101，9（Sep. 2018）
2）甲藤：“【映像符号化】映像符号化・配信における深層学習の広が映像符号化・配信における深層学習の広がり”，信学誌，105，5（May 2022）

猪飼　知宏（いかい　ともひろ）　1997年，東京大学工学部卒業．1999年，同大学大学院工学系研究科修士課程修了．同年，シャープ（株）入社．映像録再機器およびネット接続端末の開発後，HEVCおよび拡張規格の標準化活動に参加．現在，VVCの標準化活動に従事．正会員．