SONY

Deep Learningの最新動向と今後の予測

ソニーグループ株式会社 /ソニーネットワークコミュニケーションズ株式会社
小林 由幸

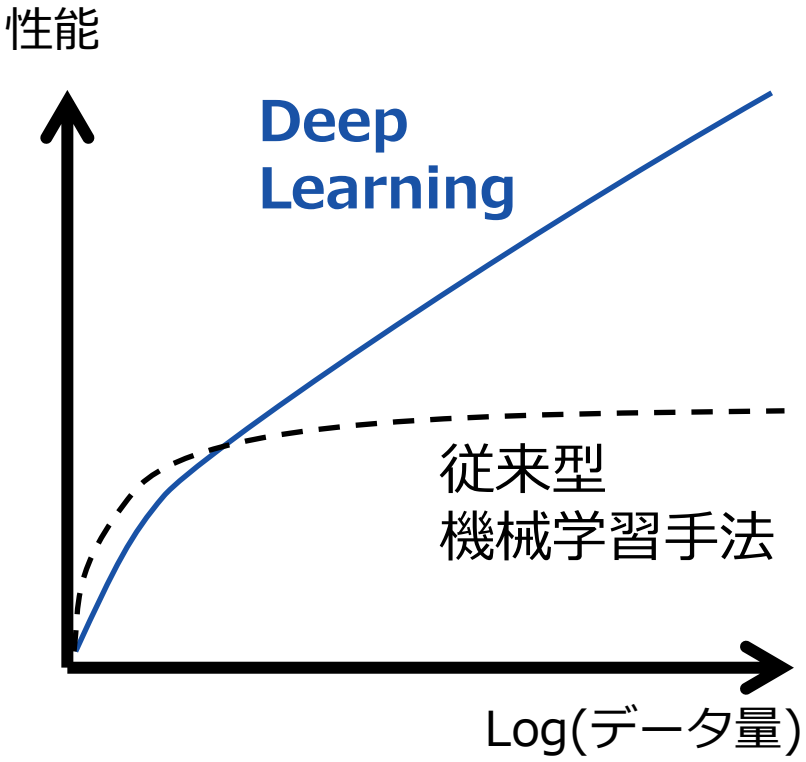
自己紹介

こばやし　よしゆき
小林 由幸

1999年にソニーに入社、2003年より機械学習技術の研究開発を始め、音楽解析技術「12音解析」のコアアルゴリズム、認識技術の自動生成技術「ELFE」などを開発。近年は「Neural Network Console」を中心にディープラーニング関連の技術・ソフトウェア開発を進める一方、機械学習普及促進や新しいアプリケーションの発掘にも注力。

産業にパラダイムシフトをもたらすDeep Learning

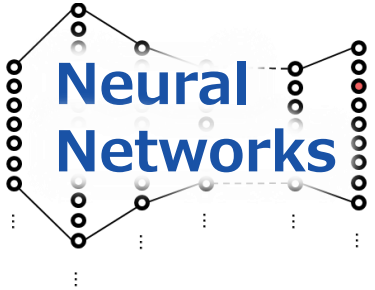
1. 人を超える高い精度



過去の技術の蓄積からなる
既存技術の多くを精度で圧倒

2. 扱いが容易

学習



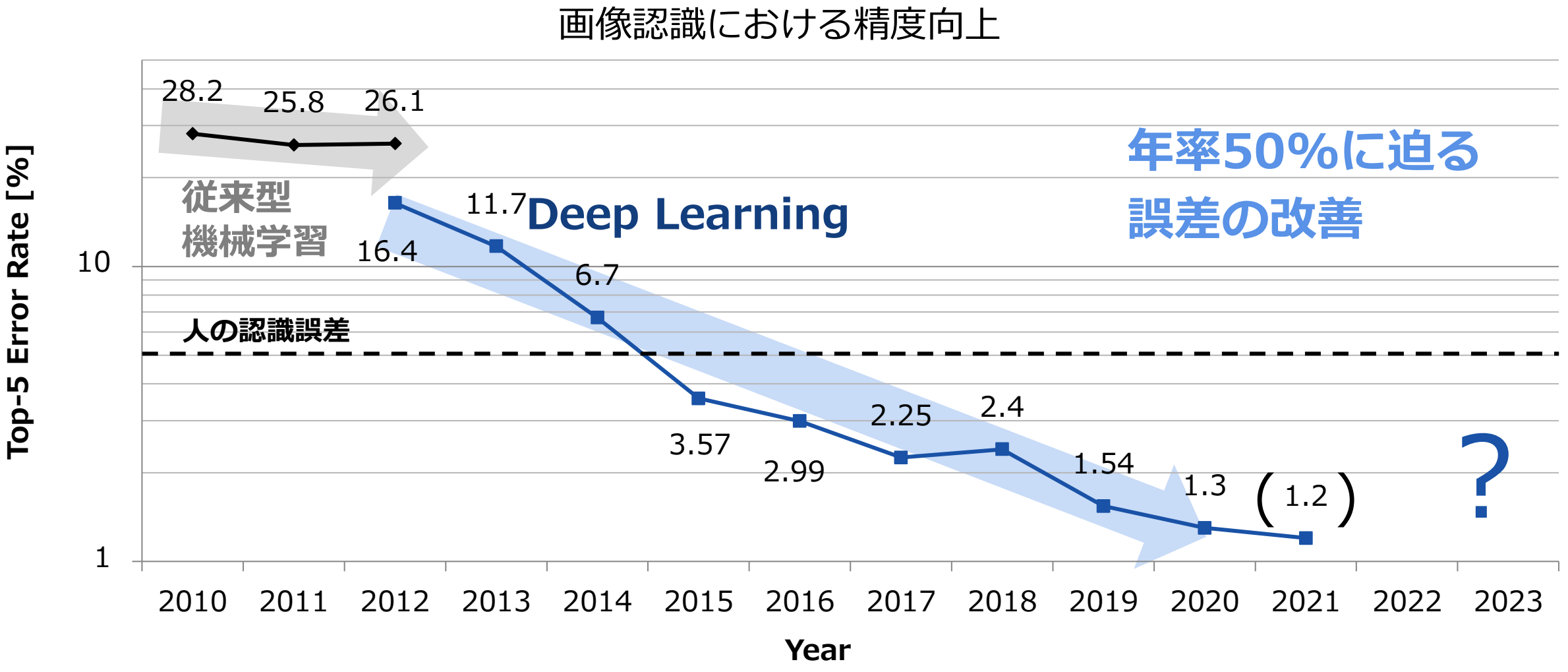
データからの学習のみで
機能を獲得

3. 用途を選ばない高い汎用性

実現する機能	入力	出力
画像認識	画像	カテゴリ
文章分類	文章	カテゴリ
音声認識	音声	文字列
機械翻訳	英単語列	日単語列
チャット	入力単語列	応答単語列
異常検知	観測信号	異常度
ロボット制御	ロボットのセンサ	アクチュエーター
…		

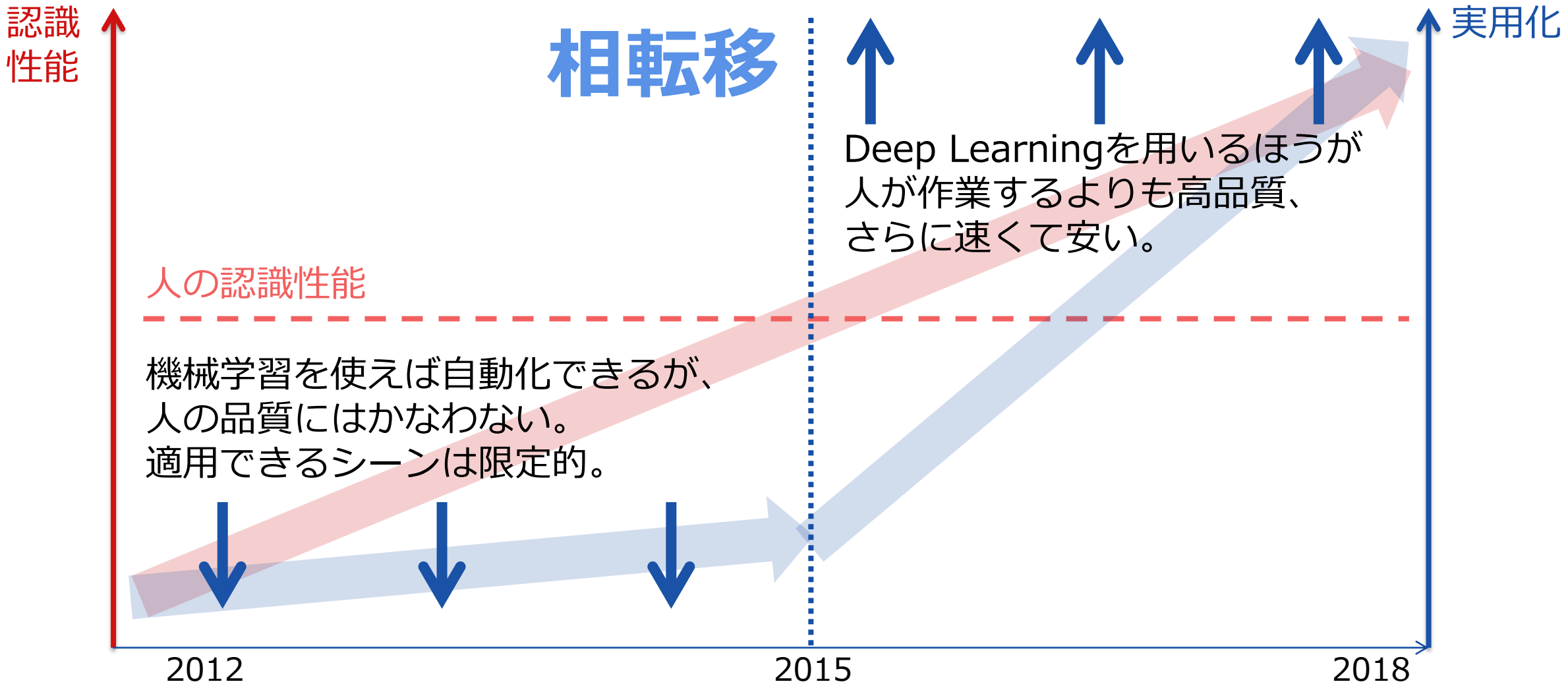
Deep Learningひとつで
様々な知的機能を実現

圧倒的な認識性能を示すDeep Learning



従来の性能限界を打ち破り、数々の課題で人を超える性能を達成しつつある

人の認識性能を超えたことで、機械学習の実用化が急加速



ビジネスの現場における人とAI

	人間	AI
Pros	• 生身の人間であること自体 • 高い柔軟性で様々なタスクに対応可能 • 簡単なタスクであれば初期コスト低	• 高速・高精度 • 運用コスト低 • コピー可能 • 柔軟にリソースを増減可能
Cons	• 低速・低精度 • 運用コスト高 • コピー不可 （高度なスキルほど教育にコストと時間を要する） • 人的リソースの増減には様々な制約がある	• 開発費高（特に物理の伴うロボット） • 社会的受容性の不足 • 説明性の不足

- 開発に要する初期投資がペイする領域では急速に機械への置き換えが進んでいる
- 人の役割は主に機械の教育（データの生成）と、初期投資のペイしないスキマ需要への対応となる
- **コンピューティングリソースを追加することで柔軟に知的作業を伴う労働力を増強できる世界へ**

AIは何らかの知的作業に依存している業界＝すべての業界に影響を及ぼす

Deep Learning最新動向

2021年現在のDeep Learningの社会実装状況

- **認識・予測タスク**にDeep Learningを用いることはもはや**当たり前**
  - 特に画像や音声の認識では2012年以降急速に応用研究が進んだ（ここまで紹介した事例は全て認識・予測）
  - 情報系以外の産業領域においても急速に普及が進んでいる

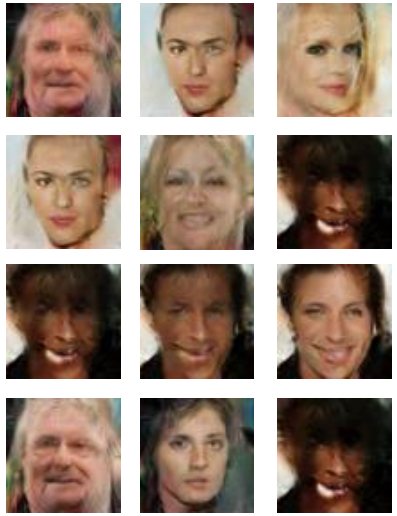
- **信号処理**領域でも**実用化フェーズ**へ
  - 2015年以降、信号処理利用域でも従来手法と比較して高い性能が得られることが次々示された
  - コンスーマ向けエレキ製品での実用化には**膨大な演算量**がボトルネックとなっている

- **今後数年で実用化が急速に進むと予測されるのは以下の3領域**

アート　**制御技術**　**自然言語処理**

アート領域におけるDeep Learningの発展 （画像生成における例）

DCGAN(2015)



PGGAN(2017) – HD resolution

BIGGAN(2018) – General Image

DALL・E (2021) – Text to Image

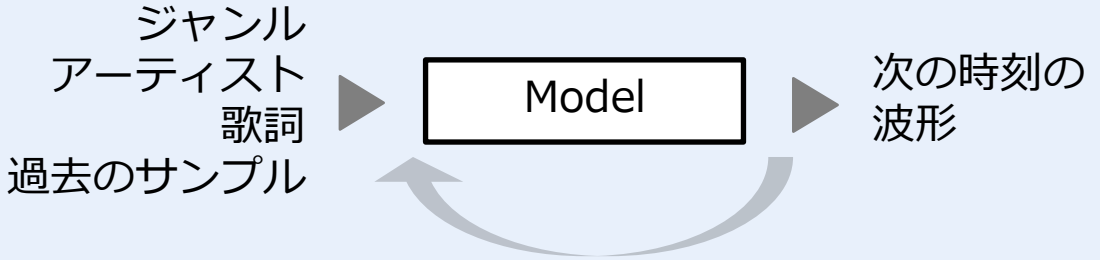
https://openai.com/blog/dall-e/

Time-Series Prediction：自動作曲

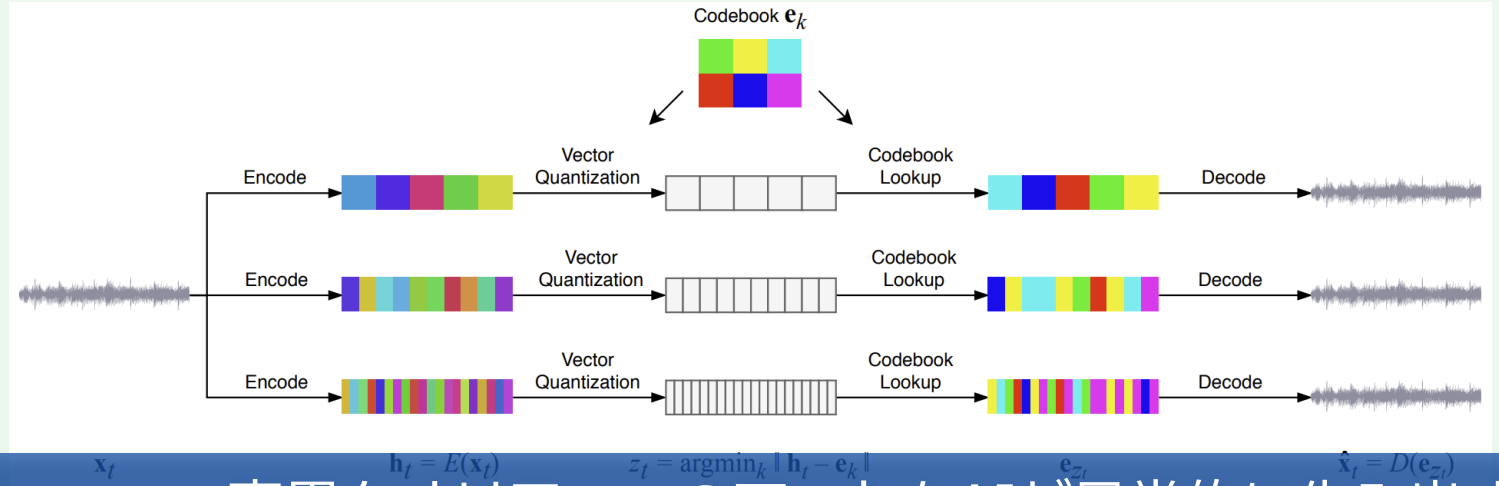
任意のジャンル、アーティスト、歌詞の楽曲を生成

OpenAI Jukebox
https://openai.com/blog/jukebox/

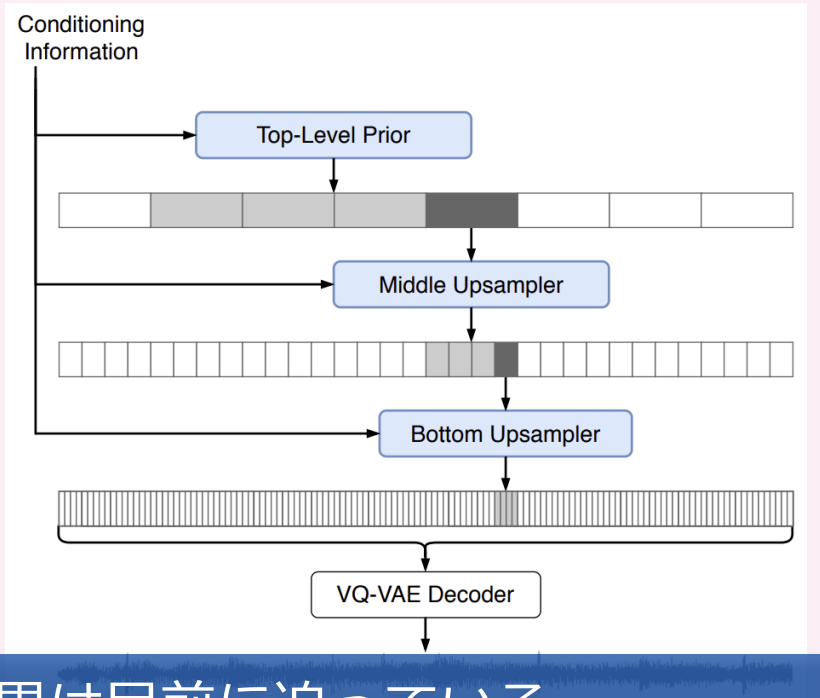
Jukebox: A Generative Model for Music
Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, Ilya Sutskever
https://arxiv.org/abs/2005.00341



Training Phase



Generation Phase



商用クオリティーのアートをAIが日常的に生み出す世界は目前に迫っている

強化学習技術動向

Deep Learningベースの強化学習が従来のAIの持っていた数々の課題を克服しつつある

囲碁

- 2015年10月　Google傘下のDeep Mindが開発したDeep Learningによる囲碁プログラム Alpha Goがプロ棋士に勝利
- 2016年3月　世界最強棋士の一人である李セドル九段に勝利
- 2017年5月　世界棋士レート一位の柯潔に三局全勝

https://ja.wikipedia.org/wiki/AlphaGo

DeepMindのAIがストラテジーゲーム『StarCraft II』のプロプレイヤーに圧勝
https://news.denfaminicogamer.jp/news/190128f

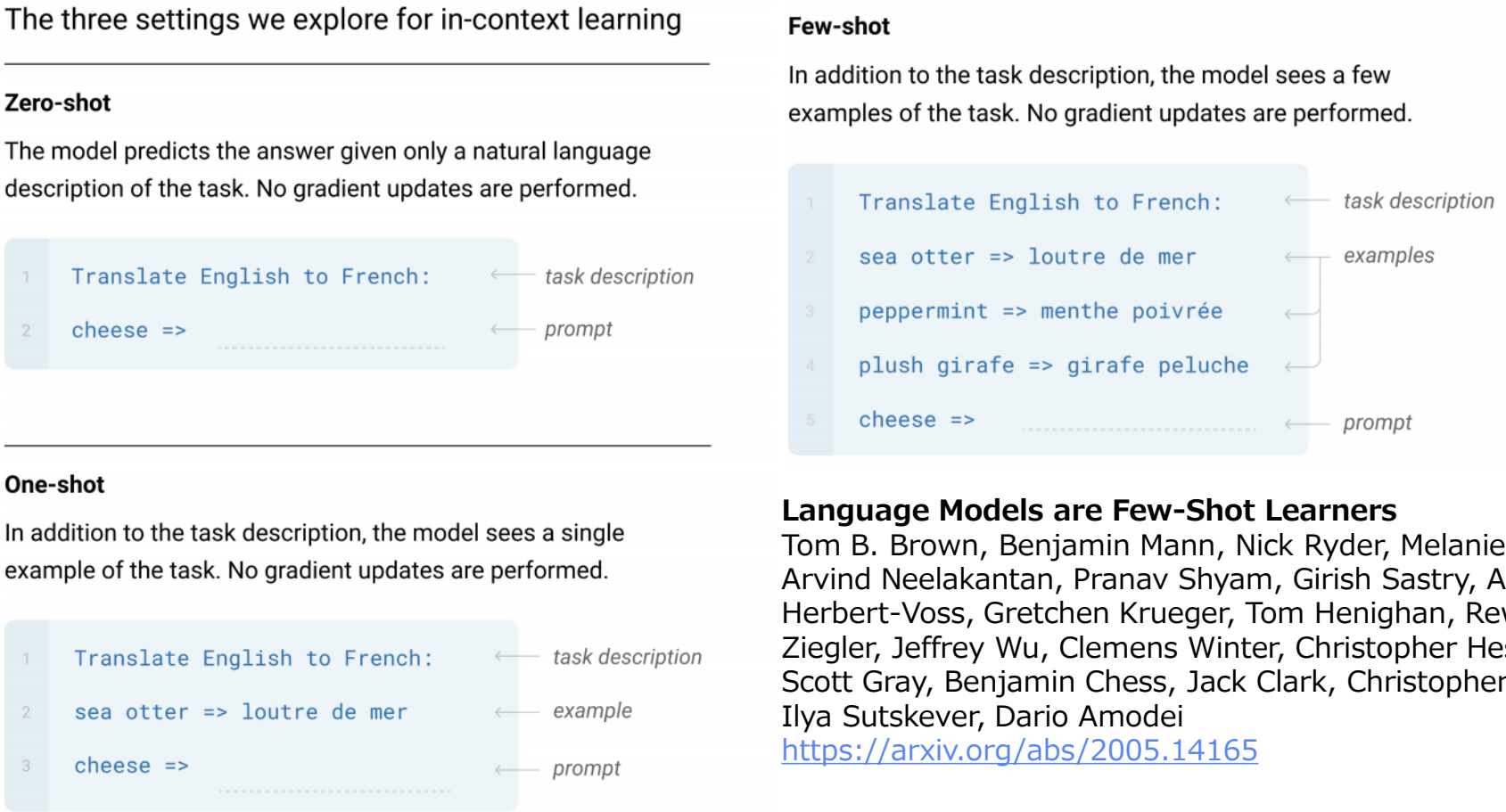
DeepMind’s AI can defeat human players in Quake III Arena’s Capture the Flag mode
https://venturebeat.com/2019/05/30/deepminds-ai-can-defeat-human-players-in-quake-iii-arenas-capture-the-flag-mode/

シミュレーションの世界での実証が進み、実世界への展開フェーズにある

自然言語処理における例

過去の単語列 ▶ Model ▶ 次の単語

様々な自然言語タスクを単一の巨大モデルで可能にした**GPT-3**の登場は2020年の最大のトピックの1つ



GPT-3がFew-shot学習で実現する応用技術の例
分類・翻訳
創作を含む文章生成
対話・計算
プログラミング・作曲…

Language Models are Few-Shot Learners
Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, Dario Amodei
https://arxiv.org/abs/2005.14165

2020/9/22　マイクロソフトはGPT-3の独占的ライセンスを取得
https://blogs.microsoft.com/blog/2020/09/22/microsoft-teams-up-with-openai-to-exclusively-license-gpt-3-language-model/

AIの発展やそれを支えるDeep Learningの限界はどこにあるのか？

- 2015年頃：「AIが認識性能で人を超えても、アートでは人に勝てない」
  ➡ 多くのアマチュアを上回る品質のアートが実現された
- 2017年頃：「AIがプロ棋士に勝利できるのは、人の棋譜が存在するからこそ。ひとたびルールが変わってしまえばAIは人間に勝てなくなる」
  ➡ 翌年2018年には棋譜ゼロからの自己学習でプロ棋士を上回るように
- 2018年頃まで：「画像や音声の認識領域では人を超えつつあるが、自然言語領域では人間との性能に大きな開き」
  ➡ 2019年には人と区別がつかない文章を生成するGPT-2が話題に
- 「Deep Learningは画像や音声など特定のデータを扱うのは得意だが、人のように複数の情報を統合的に扱うことはできない」
  ➡ 文章や画像・映像を統合的に扱うテクニックが多数登場
- 「Deep Learningで高性能を得るためには膨大なラベル付きデータが必要、人のように柔軟に様々なタスクに対応することはできない」
  ➡ 大量の生データからのみの学習する自己教師あり学習・自らの体験を通じて成長する強化学習の発展
  ➡ GPT-3は自然言語による指示により柔軟に様々なタスクに対応

2021年時点でDeep Learningの限界を予測するのは難しい
コンピュータ同様応用の可能性の広がり続ける技術になる可能性も

Deep Learning応用技術開発のワークフロー

従来のソフトウェア開発

- 仕様策定
- 設計（機能ブロックに分解）
- 実装
- デバッグ
- コンパイル
- パラメータ調整
- QA

Deep Learningベースの技術開発

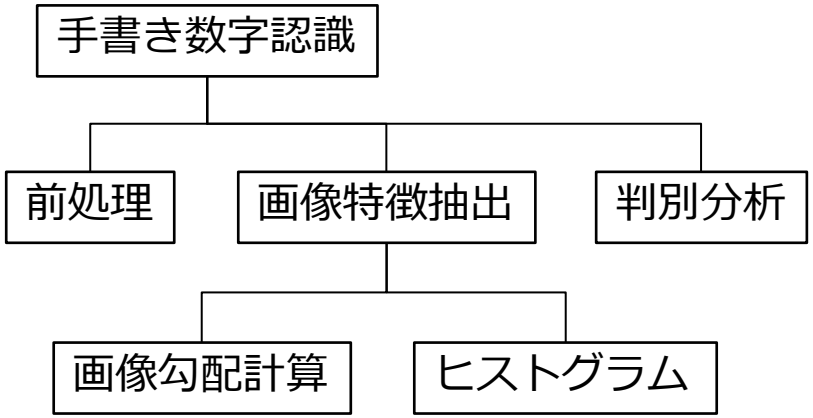
- 仕様策定
- 入出力・ネットワークアーキテクチャ設計
- **データ収集**
- **データ収集、ラベルミス修正**
- 学習
- **データ収集**
- **テストデータで評価**

Deep Learningベースの技術開発はデータドリブン
開発における大半の労力をデータワークに費やす世界

Deep Learningにより大きく変わる機能開発の概念

従来

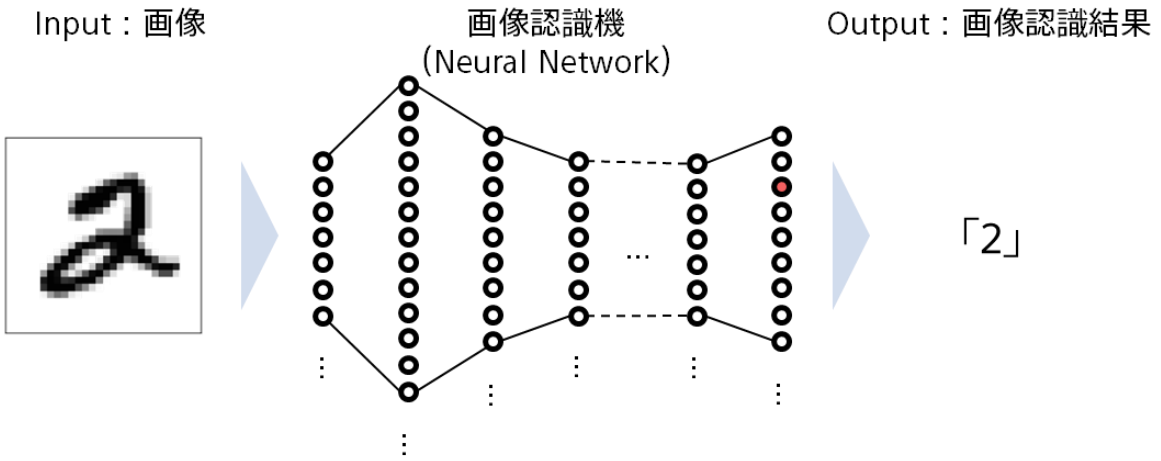
現象や機能をモジュールに分解しプログラミング



- 開発には多くの専門知識を要求
  ⇒開発難易度が高い
- 開発者の理解の範囲でのみが機能化可能
  ⇒実世界の課題では多くの場合低精度

Deep Learning時代

End-to-end学習：データのみを元に機能全体を獲得

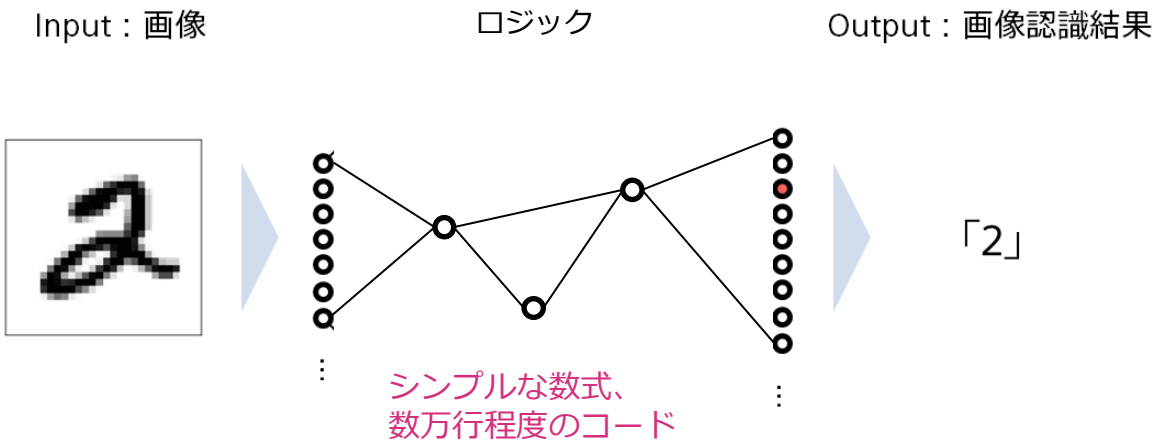


- データを用意するのみで、ほとんど専門知識なく機能の開発が可能⇒開発難易度が低い
- 開発者の知見を超え、データの持つ情報が余すところなく機能に反映される⇒高精度

Deep Learningは実世界の多くの課題において、開発効率と精度で従来型開発手法を圧倒

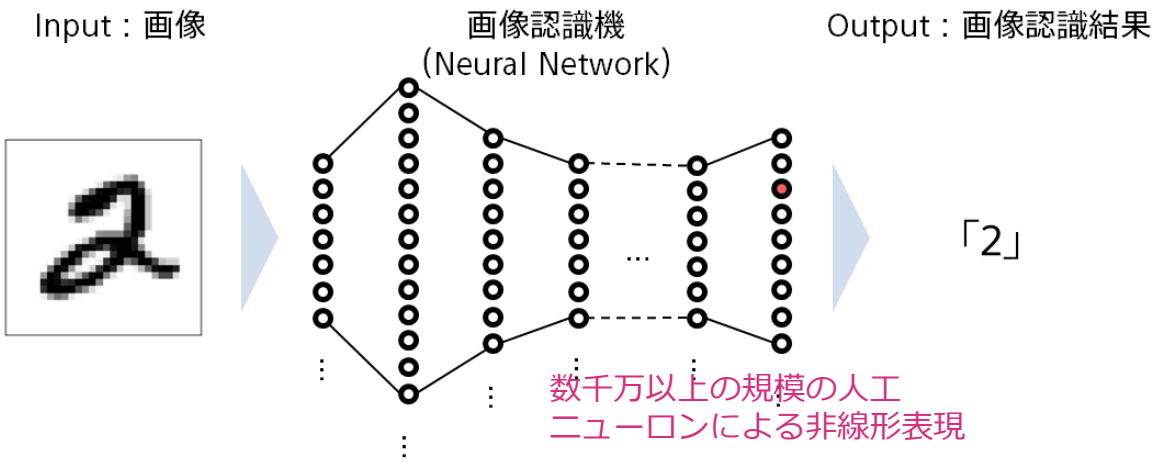
なぜEnd-to-end学習でより高い性能を実現することができるのか

従来
現象や機能をモジュールに分解しプログラミング



- 人間の理解できる範囲の**シンプルなロジック**（論文に書くことができる範囲）の組み合わせ
- ごく限られた条件での物理現象を除き、実世界の課題の多くは説明しきれなかった
- **Deep Learningの登場以後、従来法による過去の研究開発の多くが無用の長物と化しつつある**

Deep Learning時代
End-to-end学習：データのみを元に機能全体を獲得

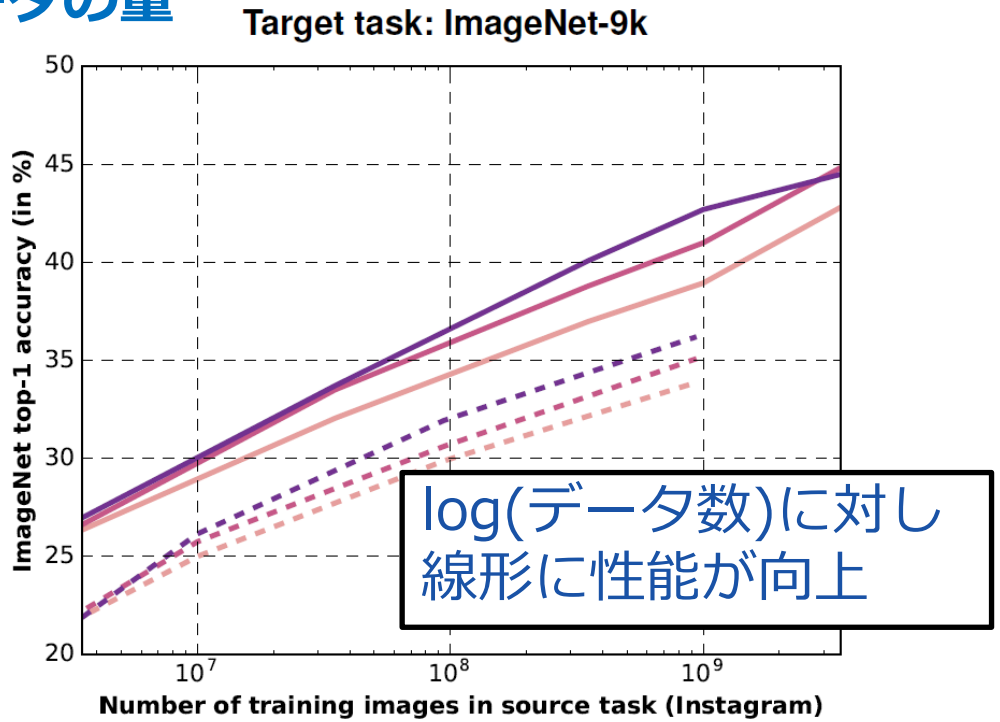


- 人の理解を超えた**複雑かつ無数のロジック**をデータから獲得させることに相当
- ニューラルネットワークのサイズを大きくし、多くのデータを与えれば与えるほど青天井に性能が向上
- **様々な分野で従来手法を急速に置き換えつつある**
  （ノイズ除去などの信号処理、物理シミュレーション、天気予報…）

Deep Learningは、原理的にシンプルな理論で記述できない問題の多くで高い精度を実現

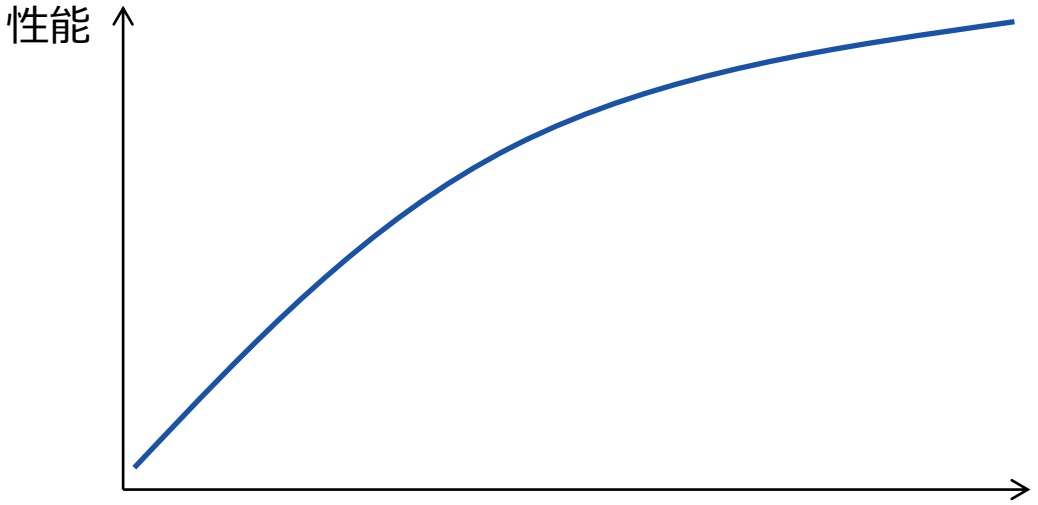
Deep Learningにおけるデータの重要性

データの量



Exploring the Limits of Weakly Supervised Pretraining
Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, Laurens van der Maaten
https://arxiv.org/abs/1805.00932

データの質



データの質

局所的データ
高ノイズデータ
低解像度データ
単一モーダル
統計値・匿名化/加工済みデータ
…

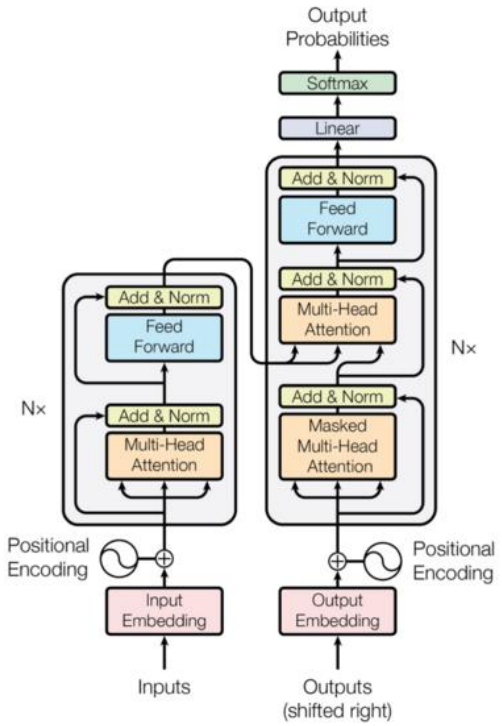
多様なデータ
低ノイズデータ
高解像度データ
マルチモーダル
生データ
Machine Readable

Deep Learningにおいて、データの量と質は性能に直結
実益を目的とした研究開発の多くでは、大量高品質なデータ無く競争力確保は困難な時代に

Attention機構とそれを用いたTransformerアーキテクチャの台頭

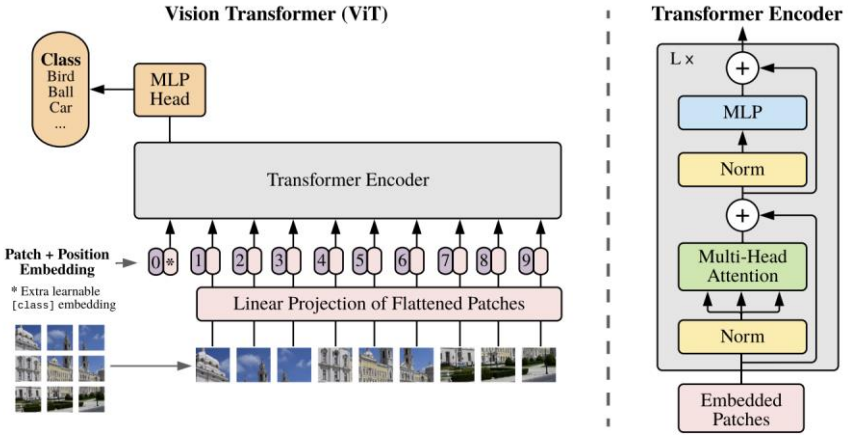
自然言語処理の領域で大きな性能向上を達成したAttention機構とそれを用いた
Transformer系技術が画像系タスクにおいてもこれまでのCNNに変わり用いられつつある

Transformer
Attention Is All You Need
https://arxiv.org/abs/1706.03762

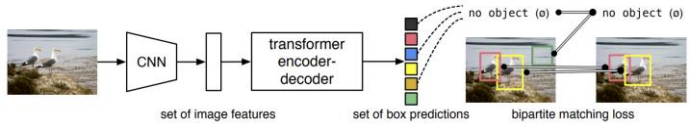


GPT-3
Language Models are Few-Shot Learners
https://arxiv.org/abs/2005.14165

ViT（Vision Transformer）
An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale
https://arxiv.org/abs/2010.11929



DETR
End-to-End Object Detection with Transformers
https://arxiv.org/abs/2005.12872

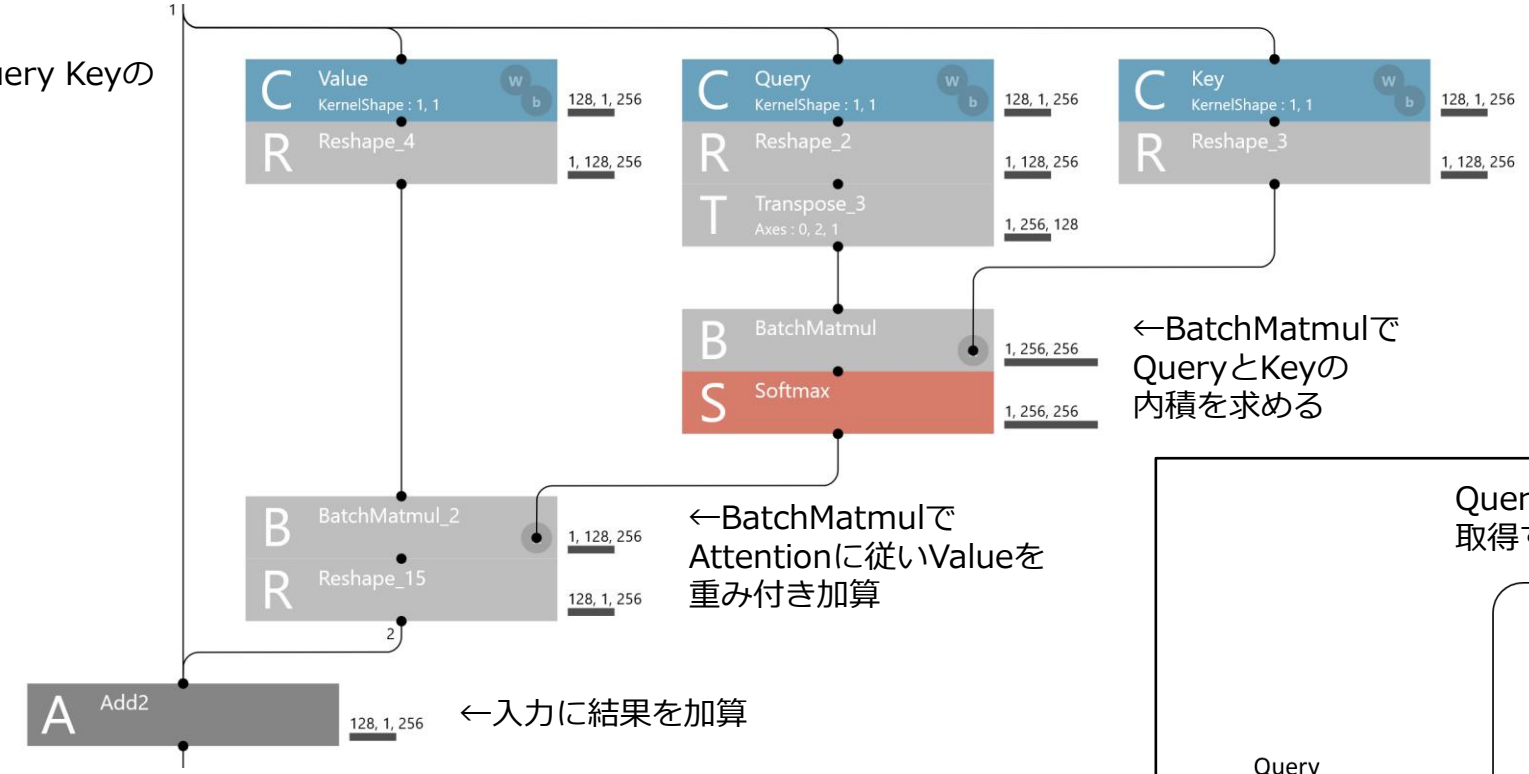


DALL・E (2021) – Text to Image
https://openai.com/blog/dall-e/

Self-AttentionのNeural Network Consoleにおける実装例

128次元の特徴× 256時刻の入力

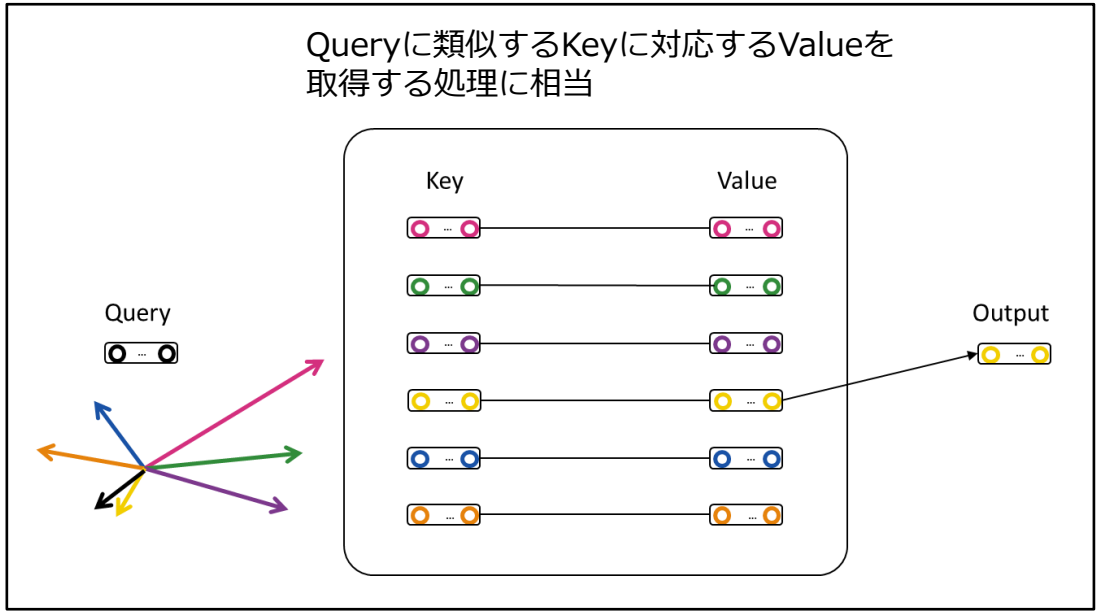
Value Query Keyの
作成→



←BatchMatmulで
QueryとKeyの
内積を求める

←BatchMatmulで
Attentionに従いValueを
重み付き加算

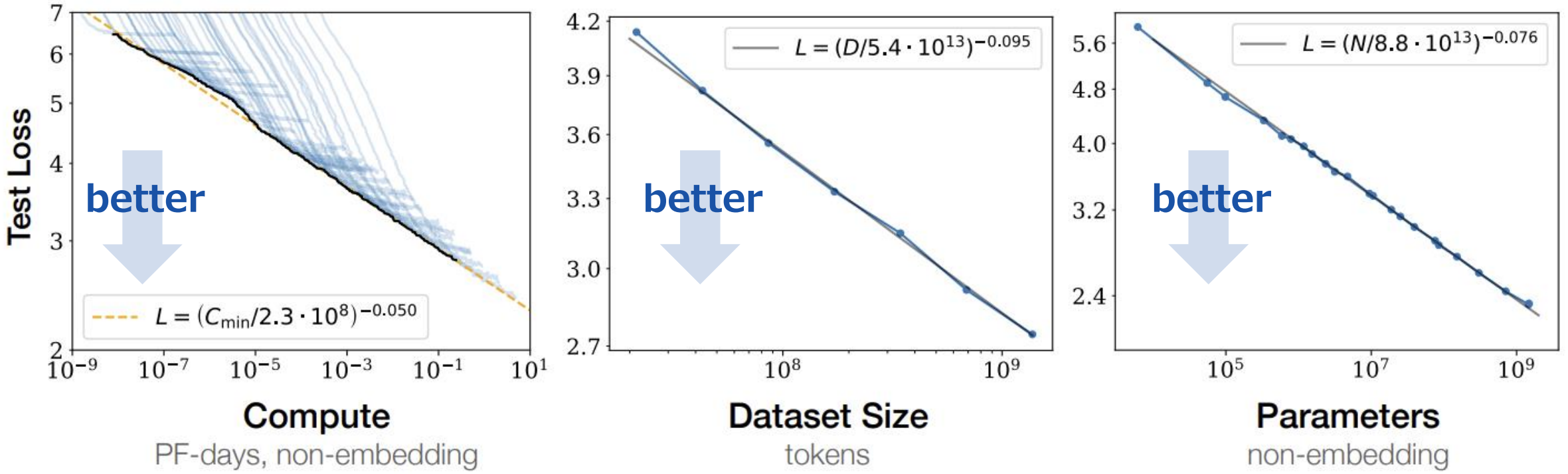
←入力に結果を加算



Attentionの詳細については以下の動画にて
Deep Learning入門：Attention（注意）
https://www.youtube.com/watch?v=g5DSLeJozdw

Scaling Rows

近年急速に発展を遂げたアーキテクチャTransformerにおいてもデータを増やし、ネットワークを大きくすることで精度を向上させられることが確認された
（性能は計算量・学習データサイズ・モデル規模のべき乗則に支配されている）



Scaling Laws for Neural Language Models
Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, Dario Amodei
https://arxiv.org/abs/2001.08361

ネットワーク設計の自動化とAutoML

従来
似たようなタスクでもデータセットごとにネットワークアーキテクチャなどを試行錯誤して探索

- 入力データサイズ
- 前処理
  - Data Augmentationなど
- ネットワークアーキテクチャ（マクロ）
  - 層数・ニューロン数
- ネットワークアーキテクチャ（マイクロ）
  - 各層の構成・ConvolutionのKernel Shapeなど
- ハイパーパラメータ
  - Learning Rateとそのスケジューリングなど
- …

- モデル構築に多くの人手による開発を要する

NAS / AutoML
典型的タスクであれば、データを与えるだけで自動的にモデルを構築することが可能に

ex) Neural Architecture Search
データに合わせたネットワーク構造を自動最適化

DARTS: Differentiable Architecture Search
Hanxiao Liu, Karen Simonyan, Yiming Yang
https://arxiv.org/abs/1806.09055

ProxylessNAS: Direct Neural Architecture Search on Target Task and Hardware
Han Cai, Ligeng Zhu, Song Han
https://arxiv.org/abs/1812.00332

- モデル構築のために必要なのは学習用データと計算環境のみ

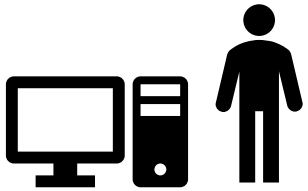
相対的にますますデータ・計算環境の重要性の比率が高まる傾向

データ量・計算資源の産むDeep Learningに対する企業の認識格差

企業A

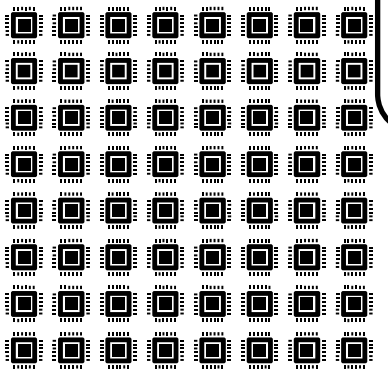
データ数：数百～数万規模
計算資源：通常1台のGPUで学習

Deep Learningを使ったところで
そう簡単に高い精度は得られない。
AIで人が行うような知的作業を
代替させるのはまだまだ困難。



企業B

データ数：数百万～数億規模
計算資源：定常的に数十台のGPUで学習



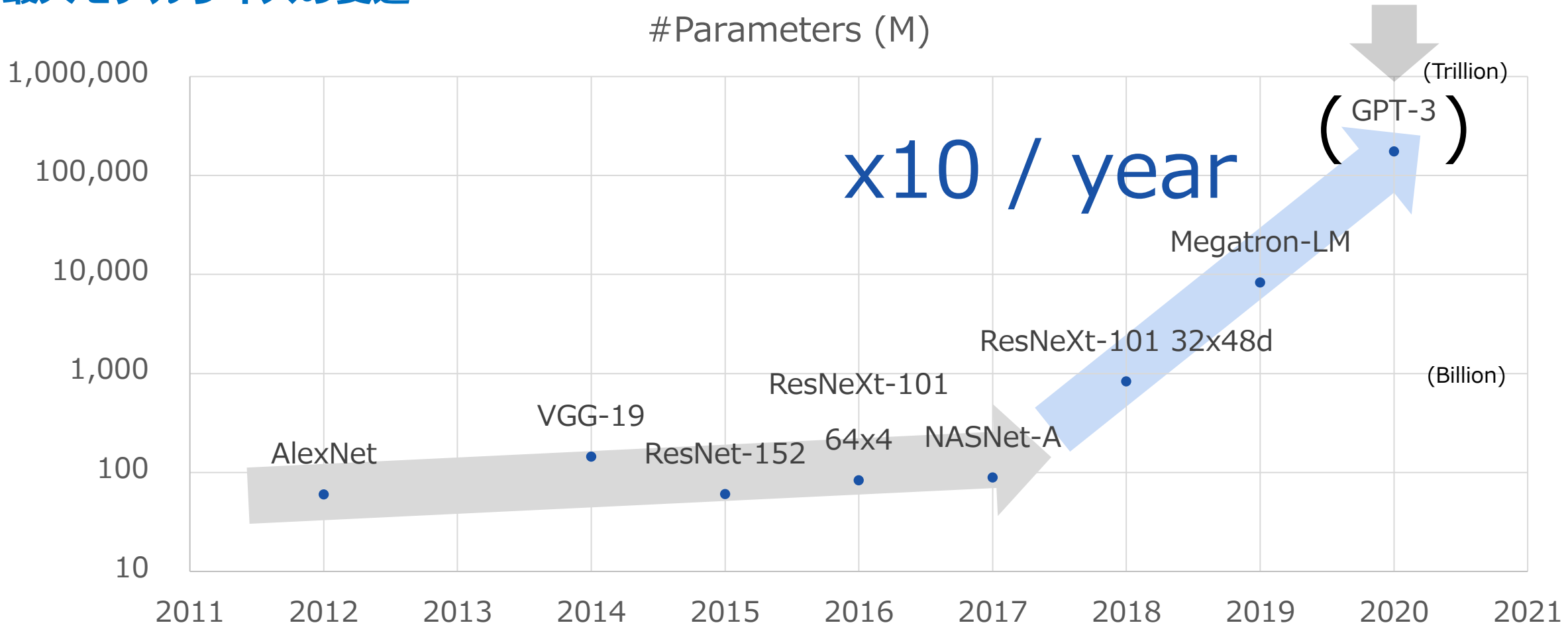
データ次第で人と遜色ない
高い性能が得られる。
より高度なタスクについても
応用が広がっている。

Deep Learningにより得られる性能は学習データ量や計算量により大きく異なる
これらへの取り組みのスケールが企業間に格差を生む

Deep Learningの扱うモデルサイズの肥大化傾向

1度の学習に要するコストは460万ドル！（355GPU/年）

最大モデルサイズの変遷



学習のために求められる計算環境も急速に大規模化

Deep Learningのもたらすゲームチェンジと新たな課題

従来型技術開発	Deep Learning時代の技術開発
• 機能分解、積み上げ型の開発・設計 • プログラミング、CAD等により開発 • 性能・精度はノウハウの積み上げで改善する • 対象領域に対する深い専門知識を持った人材が高い性能を実現	• データワークが主な業務 • コンピュータにデータから学ばせることで開発 • 性能・精度はデータ量と計算量で改善する • 膨大なデータと計算資源が高い性能を実現 • 新規技術開発の敷居が大幅に低下

Deep Learning時代の課題

データ | **ソフト環境** | **計算環境** | **応用開拓** | **人材育成** | **・・・**

従来型技術開発からDeep Learning時代の技術開発への本格移行には
多くの点でマインドセットの変革が求められる

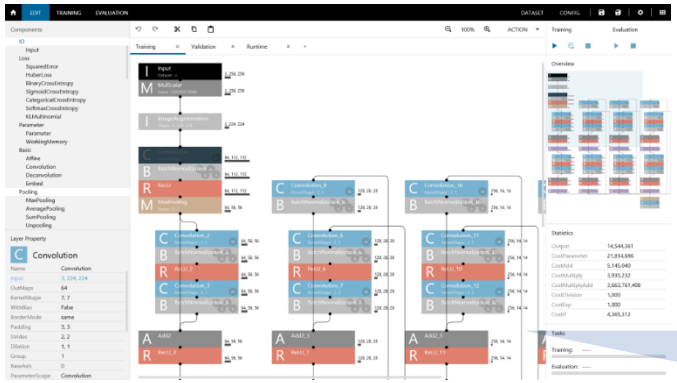
Neural Network Libraries・Console / Prediction One

Neural Network Libraries
https://nnabla.org/

様々な特長を兼ね備えた最新世代のDeep Learningフレームワーク

Neural Network Console
https://dl.sony.com/

商用クオリティのDeep Learning応用技術開発を実現する統合開発環境



実現

Prediction One
https://predictionone.sony.biz/

非専門家でも簡単操作で活用することのできる予測分析ツール

- AI技術者の迅速な育成
- 効率的なAI応用技術の研究開発～実用化

優れたAIの開発環境を提供し、需要の急拡大するAI技術の普及・発展に貢献

開発コンセプト

AI活用に立ちはだかるあらゆる障害を取り除き、
目的達成のためのハイウェイを整備する

Photo by Oscar Nilsson on Unsplash

Photo by Philipp Sewing on Unsplash

課題：新しいDeep Learning応用技術の早期立ち上げ
新しいアプリケーションのラピッドなPoC・プロトタイピング

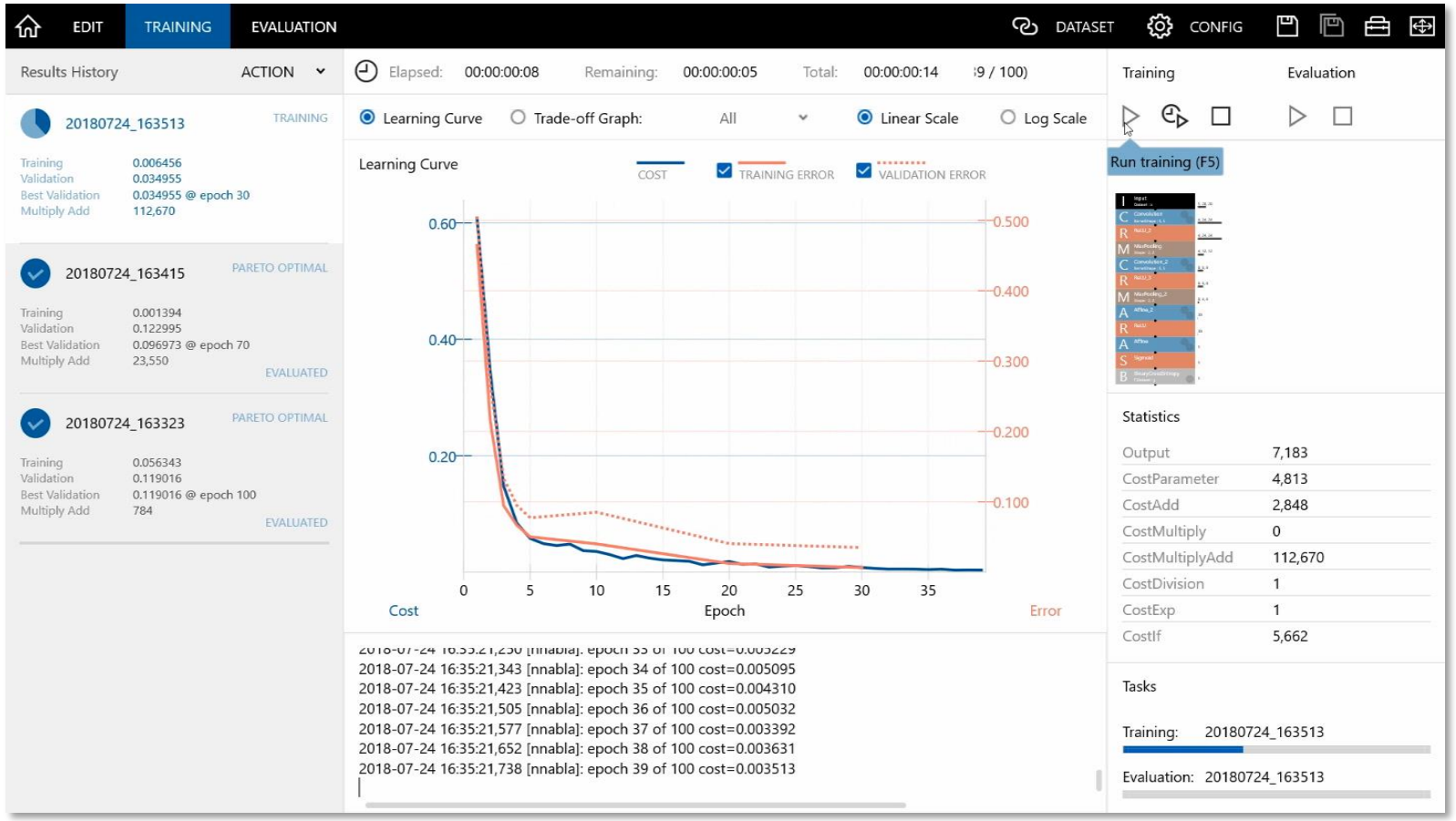
Deep Learning時代の課題

数クリックで簡単に予測分析

Prediction **One**

モデルの学習をほぼ全自動で行うAutoMLソリューション

Neural Network Consoleによる迅速な応用技術開発



- 構築したニューラルネットワークを視覚的にデバッグ
- GPUを用いた高速な学習
- 過去の履歴を管理しながら効率的に試行錯誤

DEMO

GUIの提供する様々な機能がDeep Learning応用技術のラピッドな開発を支援

EDIT | TRAINING | EVALUATION | DATASET | CONFIG

Results History ACTION

Training 0.009141
Validation 0.016989
Best Validation 0.016989 @ epoch 10
Multiply Add 7,581,894

20180725_121744

Training 0.010312
Validation 0.019297
Best Validation 0.019297 @ epoch 10
Multiply Add 4,621,506

20180725_121542

Training 0.010479
Validation 0.017505
Best Validation 0.017505 @ epoch 10
Multiply Add 4,550,082

20180725_121342

Training 0.011089
Validation 0.018268
Best Validation 0.018268 @ epoch 10
Multiply Add 4,110,658

20180725_121126

Training 0.010233
Validation 0.014832
Best Validation 0.014832 @ epoch 10
Multiply Add 9,582,594

20180725_120908

Training 0.008341
Validation 0.015817
Best Validation 0.015817 @ epoch 10
Multiply Add 11,848,002

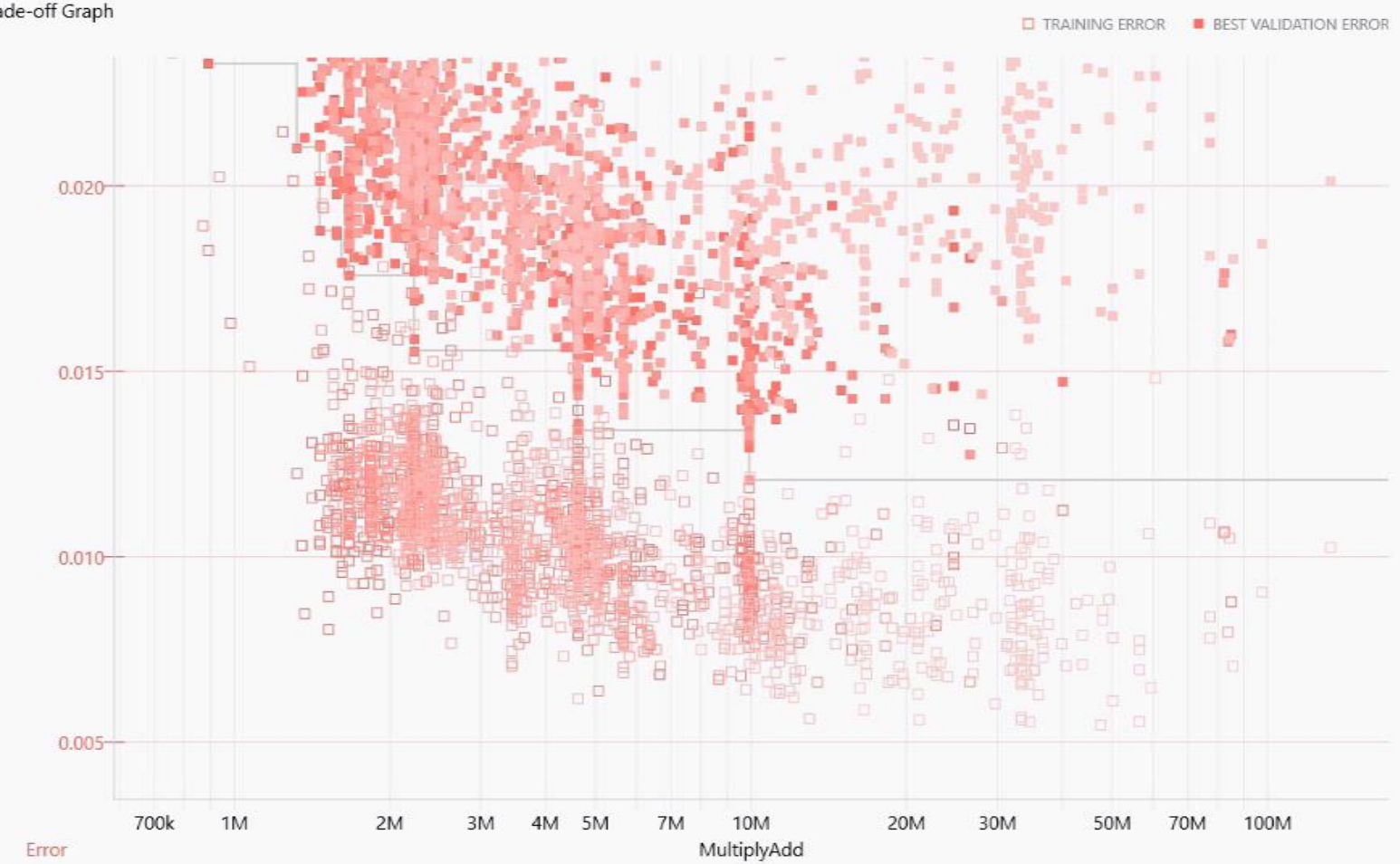
20180725_120658

Training 0.011355

Elapsed: 00:00:01:48 Remaining: 00:00:00:00 Total: 00:00:01:48 EPOCH: (10 / 10)

Learning Curve | Trade-off Graph: Previous | Linear Scale | Log Scale | 18101%

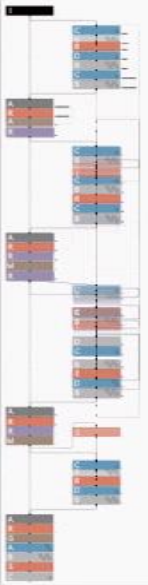
Trade-off Graph



2018-07-25 12:35:08,384 [nnabla]: epoch 5 of 10 cost=0.053760 {train_error=0.035983, valid_error=0.037724}
2018-07-25 12:35:15,307 [nnabla]: epoch 6 of 10 cost=0.046060
2018-07-25 12:35:22,370 [nnabla]: epoch 7 of 10 cost=0.040196
2018-07-25 12:35:29,496 [nnabla]: epoch 8 of 10 cost=0.035009
2018-07-25 12:35:36,559 [nnabla]: epoch 9 of 10 cost=0.031548
2018-07-25 12:35:48,232 [nnabla]: epoch 10 of 10 cost=0.028260 {train_error=0.012010, valid_error=0.016933}
2018-07-25 12:35:48,232 [nnabla]: Training Completed.

Training Evaluation

Overview



Statistics

Output	361,574
CostParameter	61,962
CostAdd	208,171
CostMultiply	100,951
CostMultiplyAdd	2,234,826
CostDivision	10
CostExp	4,618
CostIf	94,541

Tasks

構造自動探索機能

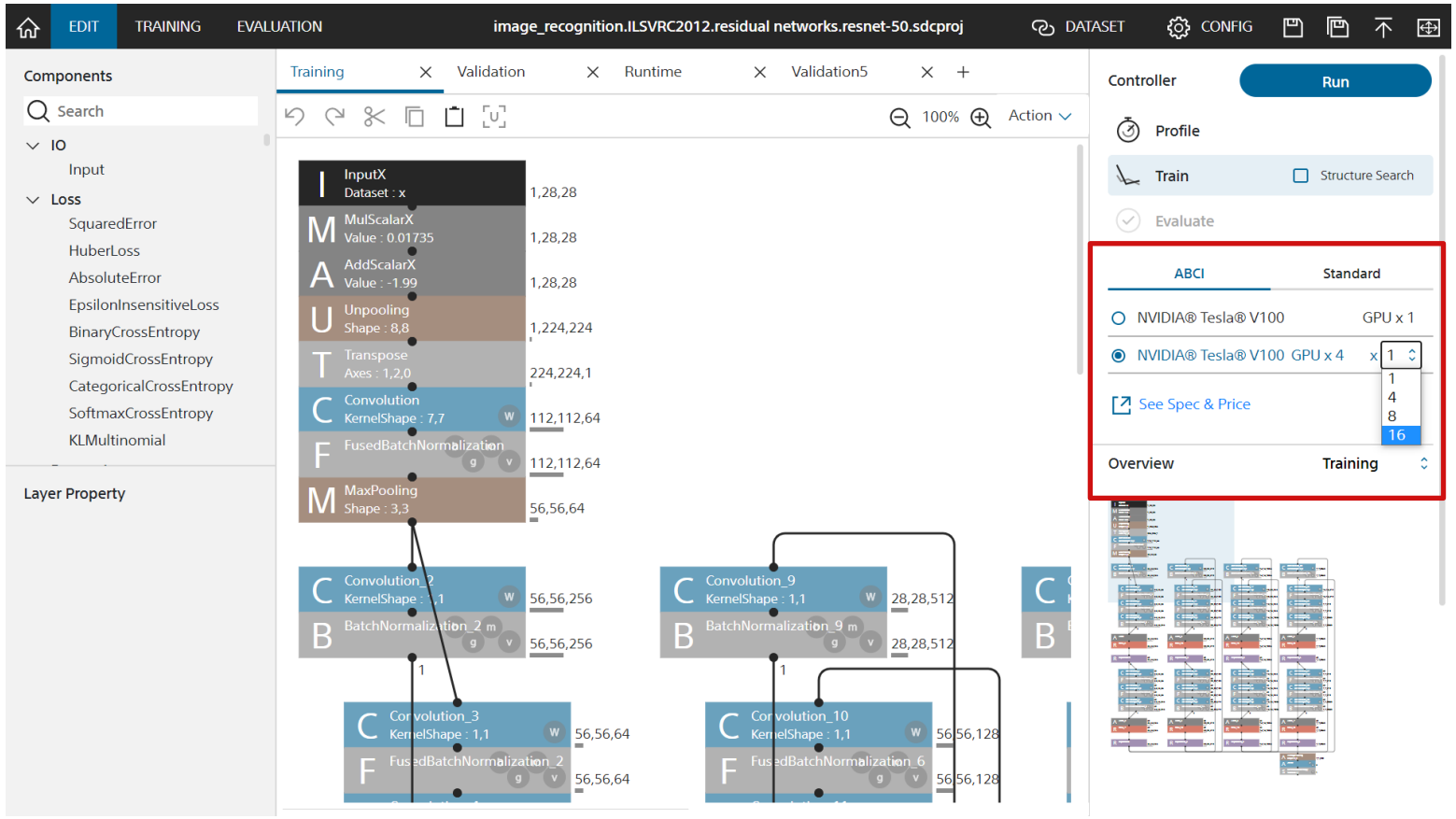
Evaluation: 20180725_123358

課題：複数かつ大規模なモデルの高速な学習を実現する豊富な計算環境

Deep Learning時代の課題

クラウド版：最大64GPUを用いたマルチノード分散学習機能 (計算環境)

国内最速クラスのスーパーコンピュータABCIの16ノードを用いた分散学習に対応



←メニューからノード数を
選択するだけで最大
4GPU/Node×16node=64GPU
を用いた学習が可能

学習速度を数十倍に高速化

これまでごく一部の研究者のみが利用してきたような計算環境を簡便なGUIを通じて利用可能に

課題：Deep Learning人材の垂直立ち上げ

Deep Learning時代の課題

Neural Network ConsoleによるAI人材の垂直立ち上げ

敷居の低いGUIベースのDeep Learning開発環境は人材の早期立ち上げに最適。ソニーでは20年度末時点でグループ内4,000人以上の社員がNeural Network Consoleを活用。
その後も急速にユーザが増えつつある。

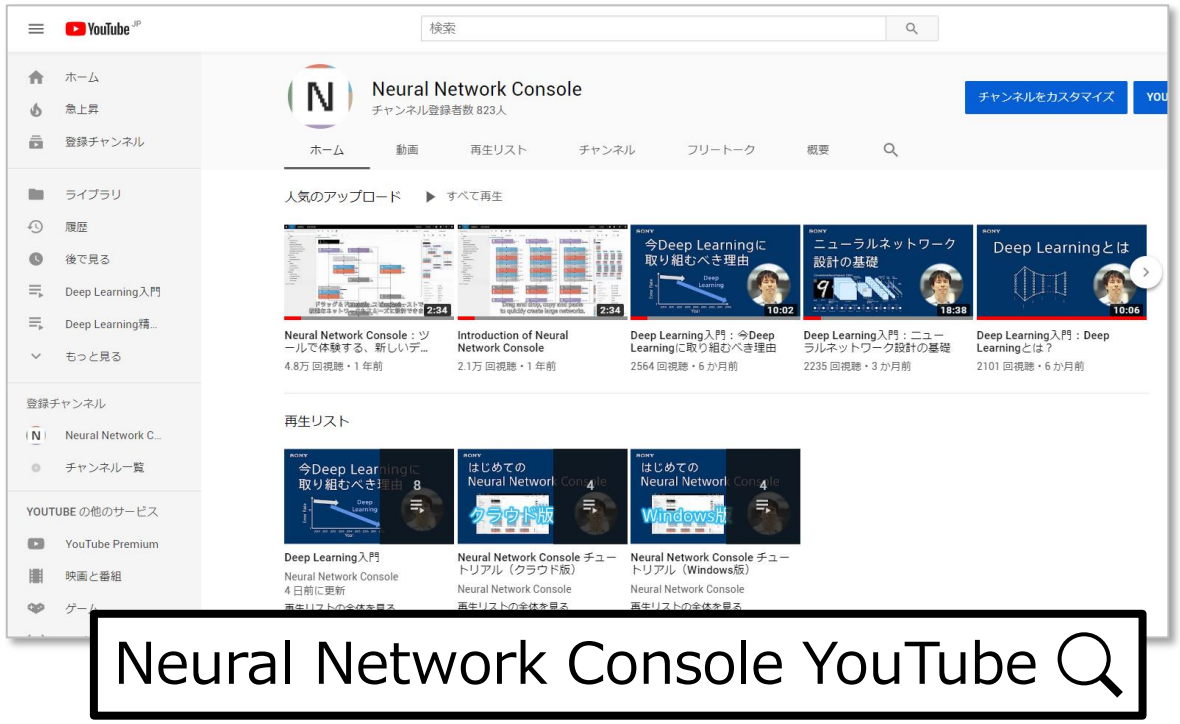
※ソニー社内Deep Learning講習会の様子（2016年頃）

Deep Learningは「習うより慣れろ」。直観的理解が活用促進につながる

動画コンテンツによる効率的な先端技術の習得

Neural Network Console

https://www.youtube.com/c/NeuralNetworkConsole



Neural Network Console YouTube

Deep Learningの入門動画、Neural Network Consoleの
チュートリアル動画を公開中

nnabla ディープラーニングチャンネル *New!*

https://www.youtube.com/c/nnabla

NNabla YouTube

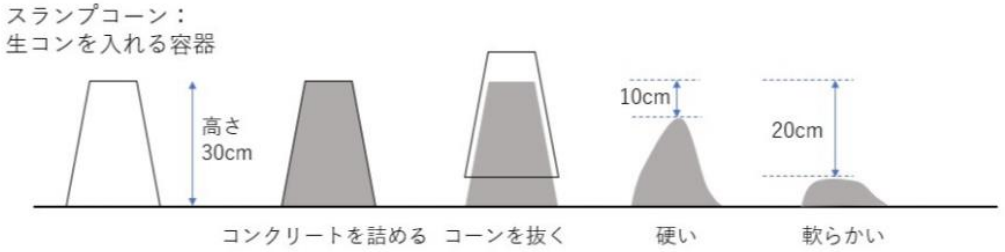
よりアドバンスドな論文紹介や
Neural Network Librariesのチュートリアル動画を公開中

必要な時、好きな時、知りたいところを集中的に、分からない所を何度も学ぶ

Neural Network Console
お客様応用事例

會澤高圧コンクリート株式会社様事例：生コンの品質判定

業界	生コンクリート
ユースケース	AIによる生コンの品質判定システムの開発
概要	画像と音響データを利用して生コンのスランプ値を判定 ※音響データを利用したモデルは別ツールで作成
効果	現場へ提供するコンクリートの品質向上
入出力データ	入力：①生コンの製造工程におけるミキサ内の練り混ぜ画像データ ②練り混ぜ後にコンクリートを一時的に貯留するホッパ内の画像データ ③コンクリート練り混ぜ中のミキサの音響データ 出力：スランプ値
NNC選定理由	✓ GUIで操作が簡単であること ✓ 概念化されすぎず自由度が高いこと



旭化成株式会社様 事例：製品検査自動化で、製造現場の課題を解決！ AIによる人間よりも高精度な製品検査システムの構築（生産技術本部）

業界	化学
ユースケース	AIによる外観検査
概要	目視で行われていた製品の外観検査を自動で判別するAIモデルの作成
効果	・従来手法では達成困難であった検査自動化の実現 ・人件費のコスト削減 ・製品の品質向上
入出力データ	入力：製品の画像 出力：OK or NG の判定結果
NNC選定理由	・GUIで簡単に開発ができる ・モデルの細かい設計が可能なため精度向上させやすい ・環境のセットアップが不要

AsahiKASEI

お客様応用事例:順天堂大学 様

「画像の鮮明化」

医療画像を用いた画像診断の最先端研究に応用。画質の荒い医療画像を鮮明化するモデル構築を実施。学会などでの受賞歴もあり。

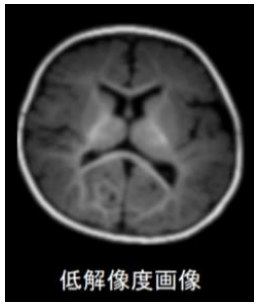
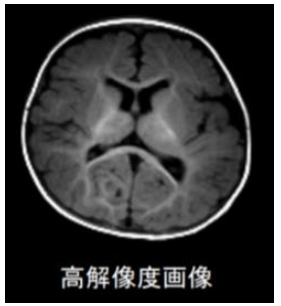
導入先	画像診断の研究
目的	医療画像の鮮明化

概要

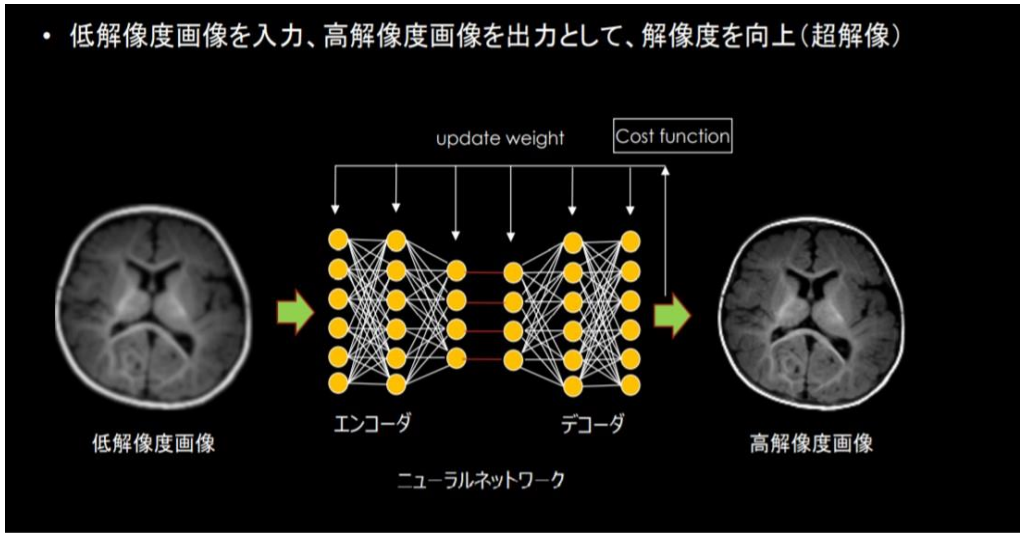
画像診断の研究

MRI、CTなどの医療画像を鮮明化のモデルを作成。古い装置で撮影された画像の高解像度化などへの応用を検討。

Deep Learning エンジン

モデル	医療画像を用いた画像の鮮明化
データ	入力 **画像**（医用画像）  低解像度画像 出力 **画像**（鮮明化画像）  高解像度画像

サービスイメージ

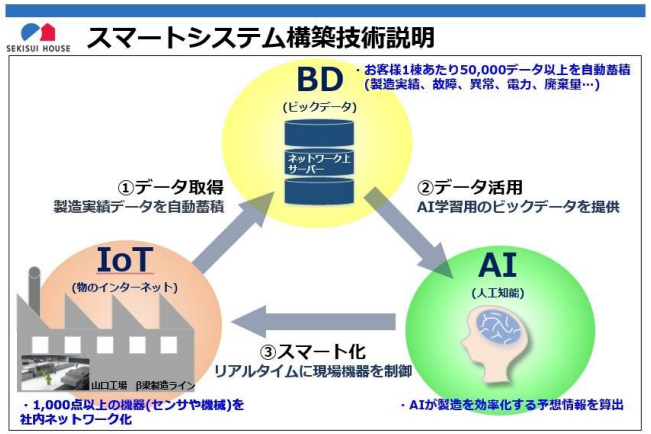


順天堂大学様 事例：画像診断の最先端研究にNNCを活用！
医用画像を用いた学習モデルの構築 （医学部放射線科）

業界	医療
ユースケース	医用画像の診断補助を行う学習モデルの構築
概要	・頭部MRIや脳核医学検査（脳血流）でのアルツハイマー病と他の認知症疾患の鑑別 ・医用画像の画質向上（超解像、ノイズ低減） ・脳MRIでの急性期脳梗塞病変検出 ・胸部のCTでの新型コロナウイルス肺炎の病変検出 ・MRIの医用画像から脂肪信号を除去して、病変を検出 ・乳幼児の頭部MRIから、年齢推定し、発達遅延の早期発見
効果	・画像診断時の補助 ・医療で実際に使うことができないため学会発表のみ →日本磁気共鳴医学会の展示ポスターが大会長賞受賞（2019年） →日本磁気共鳴医学会英文誌（MRMS)の掲載論文が最優秀論文賞受賞（2019年）
入出力データ	医用画像：DICOMデータをTiffあるいはPNGに変換して（8ビットグレイスケール）使用
NNC選定理由	・プログラミングの知識が不要 ・応用が簡単 ・開発者とユーザの距離が近い

積水ハウス株式会社様 事例：製造ラインの31%生産性向上！
AI・IT技術を駆使したスマートシステム構築

業界	住宅
ユースケース	スマートシステム構築 ①工場設備の自動制御 ②AIを利用した製造管理
概要	✓ 製造ライン上の機器を社内ネットワークに繋ぎ、データを自動蓄積 ✓ 生産状況のデータを基にAIが効率的なラインの制御を行うシステムの構築 ✓ 過去のライン状況のデータを基に省電力状態を切り替えるAI判断モデルの作成 ✓ AIで作成した製造計画を基に作業員の勤務表や生産指示数を自動調整する機能を実装
効果	✓ 製造ラインの生産性が31%向上 ✓ エコ運転の自動切り替え制御によって12%の電力削減 ✓ 無駄な残業や労働時間の9%削減
入出力データ①	入力：過去から現在の生産状況（加工情報や人の場所、生産時間等） 出力：生産時間の予測、エコ停止が可能な時間の予測
入出力データ②	入力：各支店情報、住宅情報 出力：日次で必要な生産量、出荷日の予測
NNC選定理由	✓ UIが分かりやすい ✓ 動作環境の構築が簡単 ✓ コマンドラインでのシステム連携が可能 ✓ AIに詳しくない人でもパラメータを修正して学習が簡単



お客様応用事例:日立造船株式会社様

「超音波探傷検査システム」

熱交換器溶接部の欠陥有無を超音波の波形画像から自動で判定するAIシステムを作成。

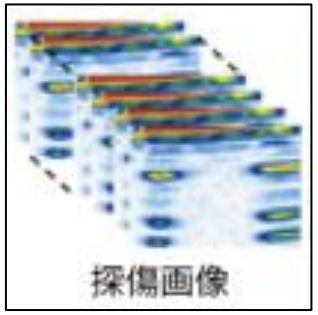
導入先	欠陥有無の検査工程
目的	検査時間の短縮

概要

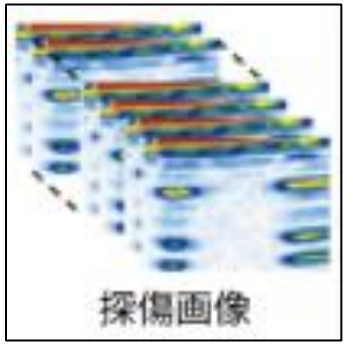
欠陥有無の検査

超音波の波形画像から熱交換器の欠陥有無を判定するモデルを作成。これにより、検査時間を75%削減、検査員の認識性能を超えた判定制度を実現。

Deep Learning エンジン

モデル	熱交換器の欠陥検査
データ	入力：**画像**（超音波の波形画像）  探傷画像 出力：**カテゴリー**（OK、NGの検査結果） OK　NG

サービスイメージ



溶接欠陥有りと判定された画像のみ目視検査を実施

群馬県蚕糸技術センター様・群馬産業技術センター様 事例：「蚕種(カイコの卵)の不良卵分類（一般財団法人大日本蚕糸会貞明皇后助成金事業）」

産卵台紙※1上の複数の蚕種の中から不良卵を一度に分類できるエンジンを開発し、蚕品種育成環境へフィードバックし孵化率の向上を目指す

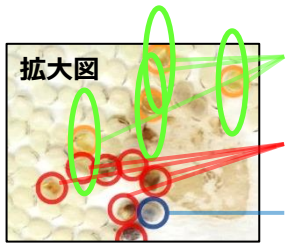
導入先	蚕種製造業者
目的	蚕種の孵化率向上

概要

蚕種の中から不良卵を分類

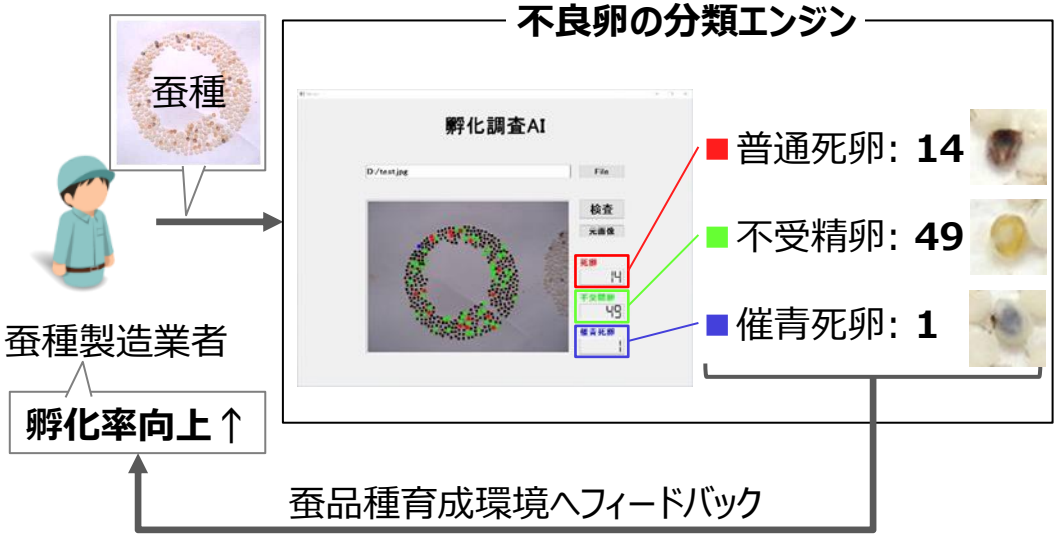
複数の蚕種の中から一度に不良卵を抽出・分類するエンジンを開発。蚕品種育成環境へフィードバックすることで孵化率の向上に寄与。

Deep Learning エンジン

モデル	不良卵の分類
データ	入力：**画像**（産卵台紙※1上の蚕種）  出力：**カテゴリー**（不良卵の種類）**数値**（位置座標）  不受精卵 (10,13),(12,15) ・・・ 普通死卵 (8,9),(110,11) ・・・ 催青死卵 (13,5),(15,7)

※1 産卵台紙:カイコに卵を産みつけさせる厚手の紙
※2 催青死卵:「催青」とは、蚕種を適切に管理された環境に置き孵化させることを意味し、「催青死卵」は催青中に死んだ蚕種を指す。

サービスイメージ



三恵技研工業株式会社様・群馬産業技術センター様 事例：ロボットアームへの応用

画像認識による位置座標検出と分類を組み合わせることでロボットビジョンを構築し、ピッキング作業などを自動化するロボットアームへの応用をご検討。

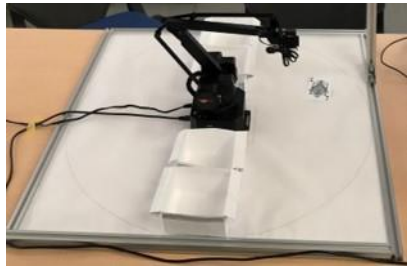
導入先	製造業 など
目的	ピッキング作業などの自動化

概要

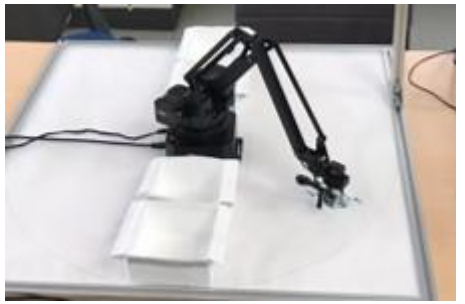
画像認識によるロボットビジョンの実現

画像をインプットに対象物の認識・分類と位置座標の検出を組み合わせることで、ロボットビジョンを実現。

Deep Learning エンジン

モデル	画像認識
データ	入力：**画像**（生産ラインなど）  出力：**数値**（位置座標） **カテゴリ**（対象物の分類） (10,30), (50,90) カード(スペード,A)

サービスイメージ



▲位置座標を元に対象物をピックアップ

▲対象物の分類ごとに仕分け

まとめ

まとめ

最先端のAIは圧倒的に高い性能を実現するだけではなく、簡単で汎用。

既にソフトウェア環境は整いつつあり、活用普及は急速に進んでいる。

AIの開発はデータトリブンなEnd-to-end学習へ。

高い性能を実現するためには膨大なデータと高速な計算環境が求められる。

AIを取り巻く状況は急速に変化しつつある。

最新動向を注視し、最先端技術やツールの積極導入を。

効率的なAI人材の育成と応用技術の開発、実用化のために

Neural Network Console / Prediction Oneも是非ご活用ください。

参考資料

Neural Network Console
https://dl.sony.com/ja/

Neural Network Libraries
https://nnabla.org/ja/

Neural Network Libraries/ Console Twitter
@NNC_NNL_jpn

Prediction One
https://predictionone.sony.biz/

モジュール分割型からEnd to endへ

従来

複数のモジュールを個別に開発し、それらを統合することで所望の機能を実現

音声認識(Audio Speech Recognition)の例

- 各モジュールの処理を経るたびに入力データに含まれる情報が失われ精度が低下
- 各モジュール個別の開発が必要で、開発効率・メンテナンス性が低い
- 各モジュール用の学習データが用意しやすい
- 各モジュールの機能が明確で小さく学習が容易
- モジュールの再利用性が高い

End to end

入力データから、最終的に得たい答えを直接的に導くモデルを学習する

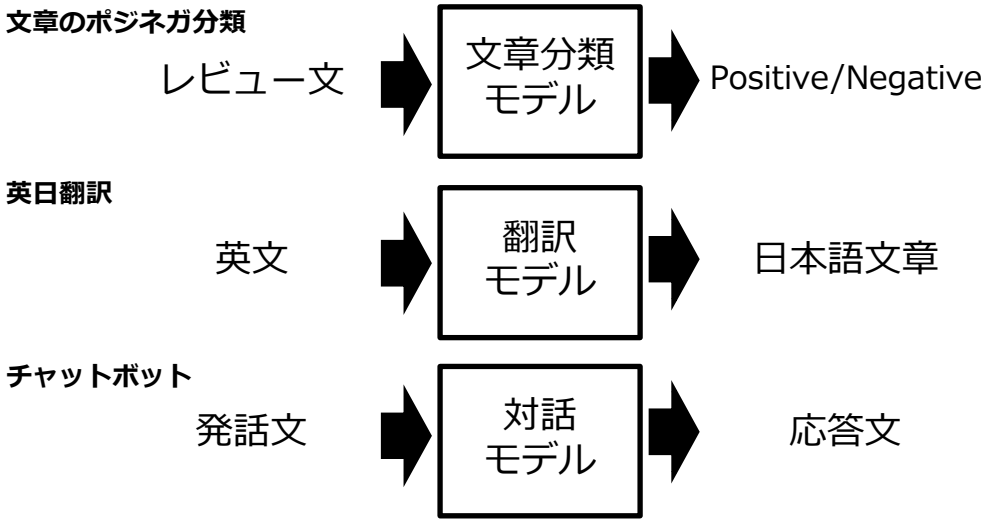
- 入力データに含まれる情報を最終結果に余すところなく反映可能なアーキテクチャであり、最終的な高性能が期待できる
- １つのモデルだけを開発すればよく、開発効率・メンテナンス性が高い
- End to end学習用のデータの用意が困難なケースも
- 学習が大規模になる傾向

現実的にはEnd to end学習用のデータを用意できないケースも多いが、
End to end化が可能なケースでは採用を検討する

タスクごとのモデル学習から、大規模モデルによる複数タスク対応へ

従来

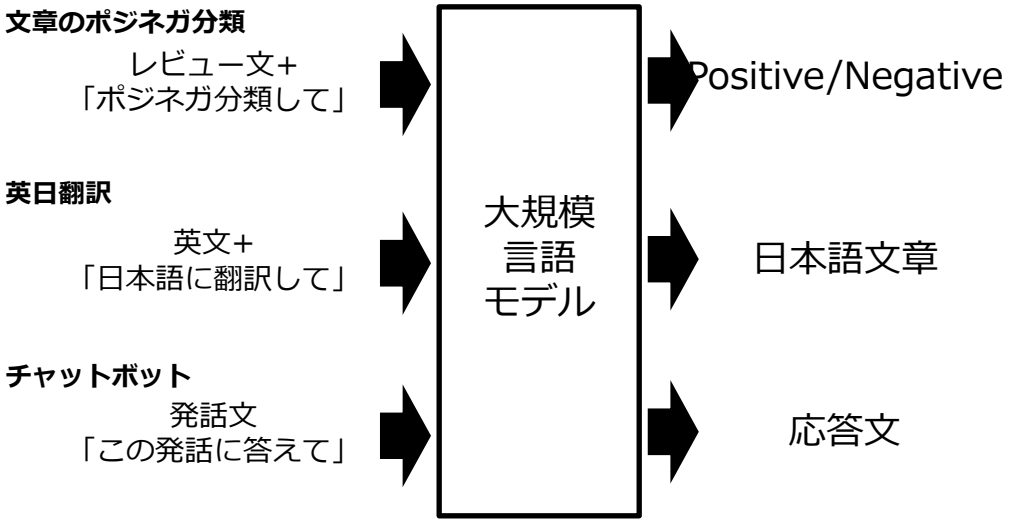
解きたいタスクごとにそれぞれモデルを学習



- 少数データしか用意できないタスクでは精度が出ない
- タスクごとに大量のデータが必要
- 十分なデータが集められる場合はタスク特化型モデルの方がメモリ量・演算量を小さくできる

今後

1つの大規模モデルで複数のタスクを解く



- 少数データから学習した小さなモデルと比較してより高い精度を実現できることが多い
- わずかな追加データで新しいタスクに適応可能
- 1つのタスクにだけ着目した場合、使用メモリや演算量が多くなる

個々のモデル開発の効率向上と高精度を両立

Deep Learningに取り組むにあたっての様々な障害

AIに仕事を奪われることに対する危機感

100%の精度が保証できなければ導入できない…

Deep Learningのコンセプトの習得が困難

膨大なデータが必要

一過性の技術ではないか…？

ブラックボックス技術、説明が困難であるため導入できない…

新たなソフトウェア環境の習得が必要

膨大な計算時間が必要

一部の企業にのみ関連する技術なのではないか…？

取り組みによる具体的効果が想像しづらい

デバッグ、試行錯誤に時間を要する

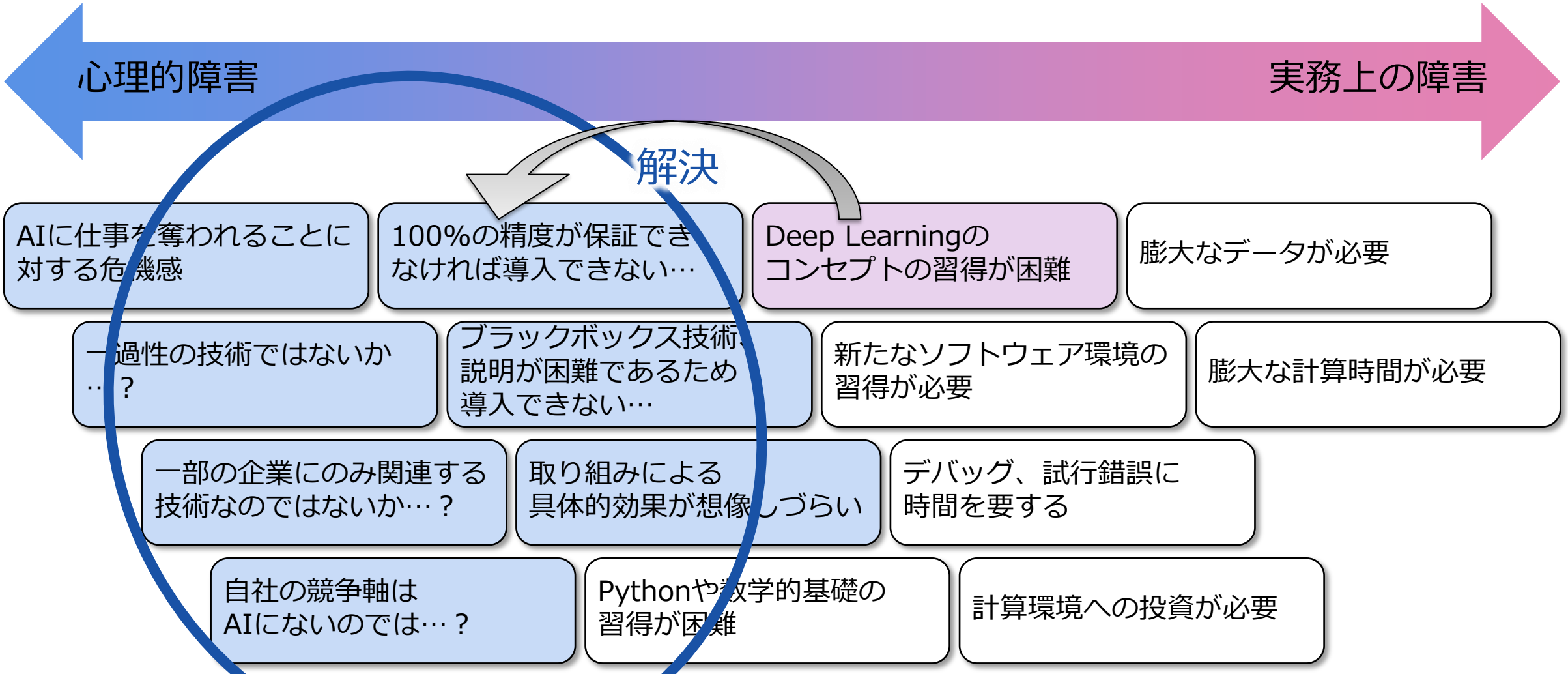
自社の競争軸はAIにないのでは…？

Pythonや数学的基礎の習得が困難

計算環境への投資が必要

迅速な習得、人材育成にはこれらを効率的に克服できる手順が有効

Deep Learningに取り組むにあたっての様々な障害



実際にDeep Learningに触れることで、心理的障害のほとんどは払拭することができる

Deep Learningに取り組むにあたっての様々な障害

心理的障害

実務上の障害

解決

解決

AIに仕事を奪われることに対する危機感

100%の精度が保証できなければ導入できない…

Deep Learningのコンセプトの習得が困難

膨大なデ

一過性の技術ではないか…？

ブラックボックス技術、説明が困難であるため導入できない…

新たなソフトウェア環境の習得が必要

膨大

一部の企業にのみ関連する技術なのではないか…？

取り組みによる具体的効果が想像しづらい

デバッグ、試行錯誤に時間を要する

自社の競争軸はAIにないのでは…？

Pythonや数学的基礎の習得が困難

計算環境への投資が必要

Deep Learningの**体験**

導入の敷居の低い環境を用いることで、効率よくDeep Learningを体験できる

心理的障害：FAQ

Q. AIによる性能は100%ではない。100%の性能が実現できなければクリティカルな現場には導入できない
A. 人による性能も100%ではない
人の性能を超えることでクリティカルな現場においてもこれよりも高い品質や安全性を実現できる。
完全でない人間が協力し合い相互にミスを補完する仕組みを構築しているのと同様、システム全体で品質や安全性を確保していく

Q. ブラックボックスな技術は、何か問題が起きた時の対応も性能改善も出来ず将来につながらない
A. 機械学習ベースのAIはデータにより教育できる。マクロなアーキテクチャの工夫でも性能向上できる
多くのケースで従来型機械学習技術による説明可能なモデルの性能をDeep Learningベースの根本的な説明は不能なモデルの性能を上回る。**説明可能なモデルの問題にいくら対応し、改良を続けてたとしても今後Deep Learningベースの技術の性能に追いつく日が来るとは考えづらい。**
人もまたブラックボックス、人を教育するようにAIを教育することができる。

人より安全・高性能であってもAIであると受容されづらい社会状況はAI普及に向けての大きな課題
・人より事故率の低い自動運転車
・医療（診断・創薬など）

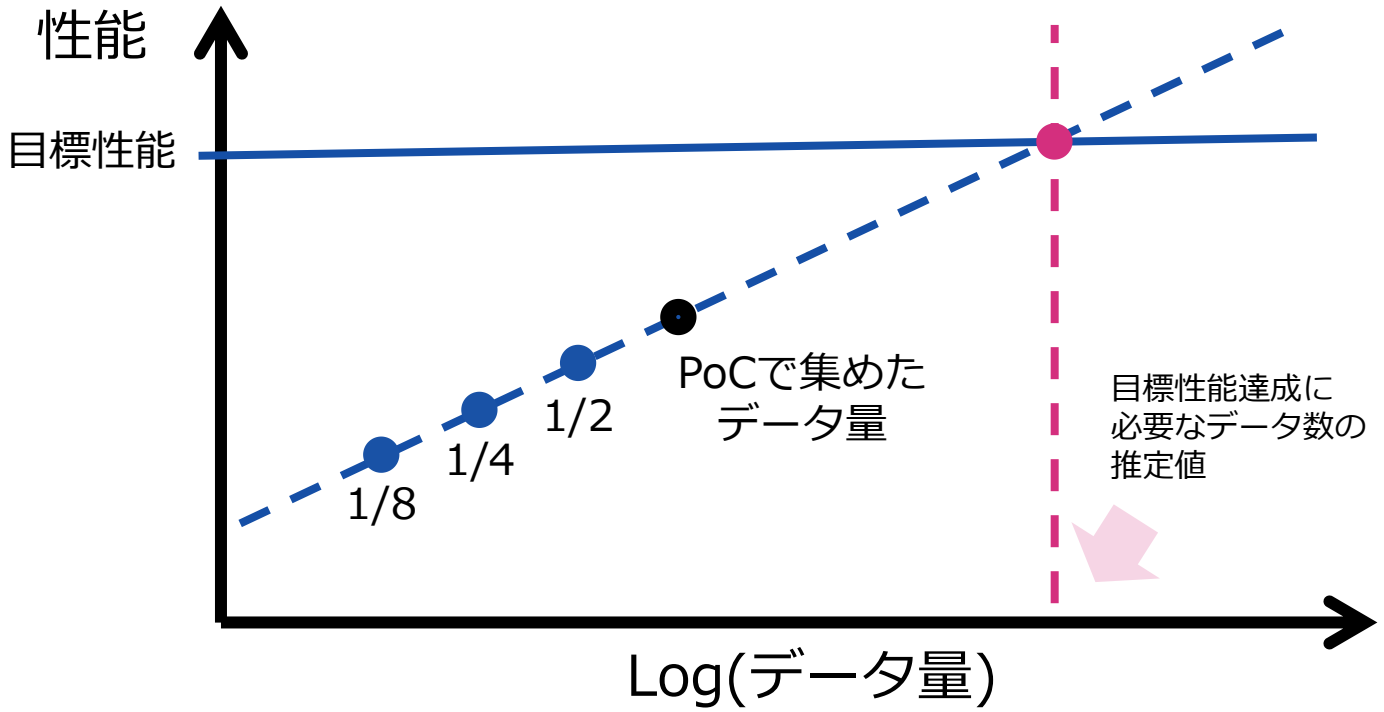
FAQ

Q. 十分な性能を出すためにデータはどのくらい必要ですか？

A. 問題の難易度とアプリケーションに求められる性能によりますが、性能はデータが増えるほど向上します

簡単な問題では数百のデータでも99%以上の性能が得られることもあれば、難しい問題では億単位のデータでも50%の性能に満たないこともある。
1%以下の精度で十分実用になるアプリケーション（アート系など）もあれば、99.999～%が求められるアプリケーション（個人認証など）もある。

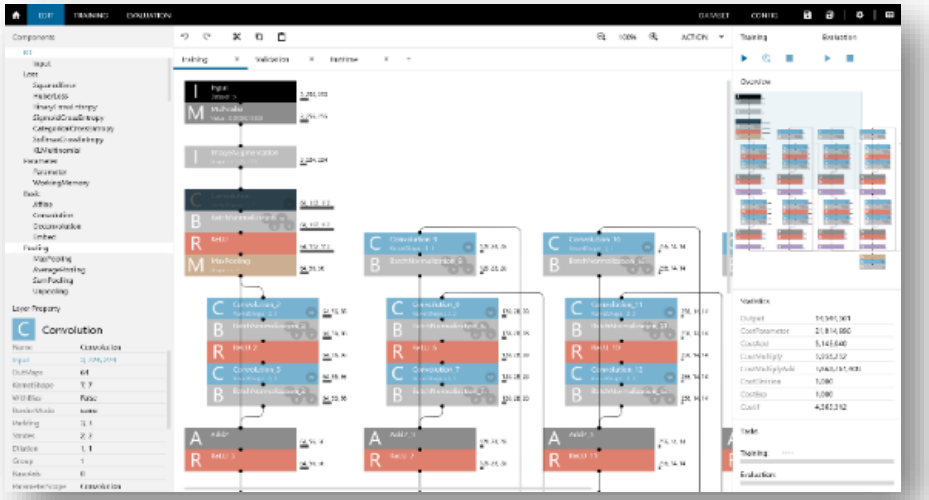
実務において必要なデータ量を見積もるには、
例えばPoC段階で集めたデータを1/2、1/4…と
減らしながら学習を行って性能を計測し、
目標性能達成のために必要なデータ量を
推定するという方法がある。



Neural Network Console クラウド版利用までの流れ

利用の流れ

Neural Network Console Cloud



Webブラウザでアクセス

Neural Network Console利用者

：利用量に応じた課金の対象

利用開始手続

次ページで詳細を説明

- NNCアカウント登録
- クレジットカード登録 or 法人申込み　※有償メニューを利用する際必須
- ABCI利用者申請　※ABCIメニューを利用する際必須

NNC利用

- 学習用データをアップロード　← ワークスペース容量
- ニューラルネットワークを編集
- 学習、評価結果　← 実行時間

学習済みモデル利用

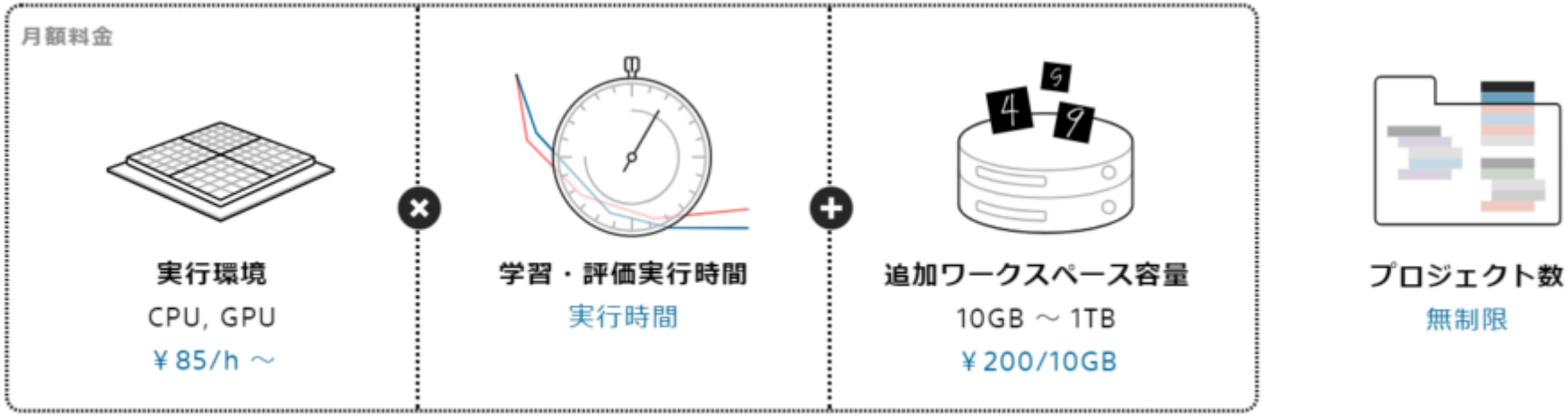
- ダウンロード
- ユーザーが開発するソースコードにマージ
- ユーザーのWebやデバイスで推論を実行

利用開始までの流れ

登録内容	手順	利用可能メニュー
STEP 1 NNCアカウント登録	Chromeブラウザで「https://dl.sony.com/ja/cloud/」にアクセスし「今すぐはじめる」をクリック	無料枠 (CPU10h、ワークスペース10GB)
	メールアドレス、パスワード、国/言語、生年月日を登録	
STEP 2 クレカ登録 or 法人申込み	クレカ登録：NNCクラウドの画面上から登録 法人申込み：申込書による登録(問合せフォームから連絡) 問合せフォーム：https://support.sonynetwork.co.jp/IoT/web/form113.html	ABCI以外のGPUメニュー、 ワークスペース追加
STEP 3 ABCI利用者申請	「https://dl.sony.com/ja/cloud/abci/apply.html」からABCI利用者申請フォームにアクセス	ABCIのGPUメニュー
	フォームに必要事項を入力し申請	
	審査結果通知、審査OKならNNCにてABCI利用可能	

Neural Network Console クラウド版の利用料金

■利用料金の考え方



■利用料金

リソース				料金
学習・評価	CPU	CPU×1		85円/時間
	GPU	Standard	NVIDIA® TESLA® K80 GPUx1	210円/時間
			NVIDIA® TESLA® V100 GPUx1	560円/時間
			NVIDIA® TESLA® V100 GPUx4	2,900円/時間
			NVIDIA® TESLA® V100 GPUx8	5,800円/時間
		ABCI	NVIDIA® TESLA® V100 GPUx1	300円/時間
			NVIDIA® TESLA® V100 GPUx４	1,650円/時間
追加ワークスペース(月額)				200円/10GB

NNCで作成したモデルを利用する様々な方法

利用方法	実行環境	言語	GPUの利用	メリット	デメリット
1. Web API	クラウド	環境により様々	Yes	最も簡単	
2. NNabla Python CLI	Neural Network Libraries	Python (CLI)	Yes	最も簡単	低速
3. NNabla Python API		Python	Yes	比較的容易	
4. NNabla C++ Runtime		C++	Yes	推論時にPython不要	
5. NNabla C Runtime		C	No	非常にコンパクトに組み込み可能	環境に合わせた最適化が必要
6. ONNX、TensorFlowフォーマット対応ソフトウェア、ハードウェア	各社の提供するONNX、TFフォーマット対応Runtime	環境により様々	環境により様々	環境により様々	現状は互換性の問題が生じることも

※ NNabla C++ Runtimeからの実行方法　https://github.com/sony/nnabla/tree/master/examples/cpp/mnist_runtime
※ NNabla C Runtimeからの実行方法　https://github.com/sony/nnabla-c-runtime
※ ONNX、TFへのコンバート方法　https://nnabla.readthedocs.io/en/latest/python/file_format_converter/file_format_converter.html

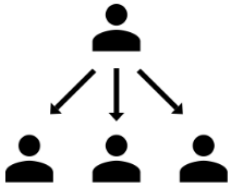
作成したDeep Learningモデルを迅速にアプリ・サービスに組み込み

法人プラン

■特徴

グループ機能により共同作業が便利に

グループ内でプロジェクトとデータセットをメンバー間で共有できます。メンバー間でのシームレスな共有により、複数人での共同作業が効率的に行えます。



グループの管理

グループ管理者は、メンバーの追加・削除やワークスペースを増減し、グループを適切に管理できます。メンバーの利用状況を一覧で簡単に把握できます。

請求書払いとおまとめ請求

GPU利用やワークスペース追加した場合の従量課金分も請求書払いが可能です。またグループに対して1つの請求書におまとめします。

■料金

基本料金：月額10,000円

グループ共有ワークスペース追加：月額200円/10GB

※ GPU等の従量サービス利用分は別途費用が掛かります。従量サービス料金表はこちら。

<月々のお支払いイメージ>

基本料金		CPU/GPU利用料		追加ワークスペース		
10,000円 グループ機能 +ワークスペース100GB	+	22,400円 NVIDIA®TESLA®v100 GPU ×40時間	+	200円 10GB追加	=	ご請求金額 月額 32,600円

技術支援

AIモデル開発における以下のようなお悩みはございませんでしょうか。
弊社ではこれらを解決する技術支援もご提案させていただきます。

・ツールのレクチャーをしてほしい
・モデル開発を委託したい、または伴走型で行ってほしい
・モデル開発後のアプリ開発を委託したい
・自社の課題をどのようにAIで解決すべきか整理ができない
・PoCの技術支援をおこなってほしい

研修プラン

#	研修アジェンダ	概要
1	人工知能の概要	• 基礎知識としての人工知能 (AI) のビジネス概況を整理 • Deep Learningにフォーカスし、その概要と特徴を解説
2	Deep Learning活用事例紹介	• ビジネスにおけるDeep Learningの活用事例を紹介
3	Neural Network Libraries/Console概要	• ソニーにおけるAIの取り組みをご紹介すると共に、 Neural Network Libraries/Consoleの特徴を解説
4	Deep Learning開発のポイント	• 実際のDeep Learningモデル開発における重要ポイントを、 データとモデルの両面から解説
5	Neural Network Consoleを使ったDeep Learning モデル開発	• Neural Network Consoleを利用したDeep Learning開発の 流れ(データ準備~設計~学習~評価~推論実行)と操作方法を解説
6	Deep Neural Network を用いたハンズオン	• Deep Neural Networkを解説し、サンプルデータをもとに、 一般的なDLによるモデル開発をハンズオン形式で実習
7	Convolution Neural Network を用いたハンズオン	• Convolution Neural Networkを解説し、 画像データを例にしてモデル開発をハンズオン形式で実習
8	Recurrent Neural Network を用いたハンズオン	• Recurrent Neural Networkを解説し、 時系列データを例にしてモデル開発をハンズオン形式で実習
9	社内課題へのDeep Learningを使った解決	• 貴社内の課題やアイディアをベースに、 Deep Learningによってどのような解決が望めるかをディスカッション