エンベデッド ディープラーニング フレームワーク

「KAIBER」（カイバー）

for IoT & Edge Computing

ご紹介資料 Ver3.6

ディープインサイト株式会社

2017年7月

会社紹介

ディープインサイト株式会社

DeepInsight Inc.

設立：2016年3月

代表取締役＆CEO：久保田 良則

取締役＆CTO：古川 智洋

業務内容：ディープラーニングエンジンの開発と応用システムの構築・販売

略歴：久保田 良則

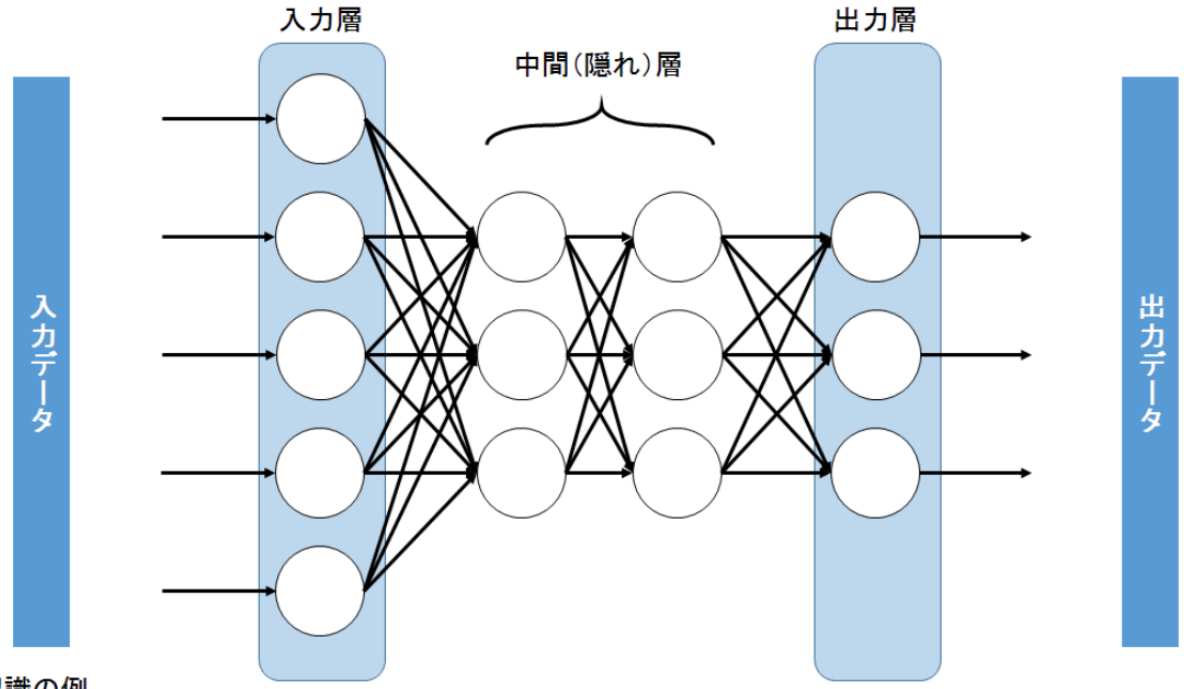
ベンチャー企業にて組込み開発支援ツールや人工知能言語のビジネス推進、米Sun Microsystems等でパートナー開拓、独SAPではIoTや組込み分野へのDB拡販を担当。その後、ビッグデータ分析エンジン開発のベンチャー設立に参画。ソフトバンクと連携しIBM Watsonの日本最初のアプリケーションパートナーとして開発プロジェクトを推進。

略歴：古川 智洋

ジャストシステムで一太郎のコアエンジンなど日本語処理技術の開発をはじめ、米CA Technologies社などでドキュメント処理システムや配信プラットフォームの開発に従事。前職では、ビッグデータ関連ベンチャーにて、IBM Watsonプロジェクトの日本語アプリケーション開発を推進。自然言語処理・統計・機械学習に精通。

ディープラーニングとは

ディープラーニングは、データをもとに、コンピュータが自ら特徴量をつくり出す。人間が特徴量を設計するのではなく、コンピュータが自ら高次の特徴量を獲得し、それをもとに画像や音声などを分類できるようになる。



画像認識の例

花

ディープラーニング 適用分野

レコメンデーション クラスタリング	分類、識別	市場予測	評判分析
情報抽出	文字認識	ロボット	画像解析
遺伝子分析	検索ランキング	金融	医療診断

ディープラーニング 応用例-1

- 株価予想システム

SNSに投稿される声と株価を比較し、株価予測

- 医療画像診断

CTスキャナーなどのデータを分析し、医師の診断を支援

- デジタル・フォレンジック

膨大なデジタル文書から不正の証拠や特許の調査のキーワードを自動抽出

- 似ているデザイン広告検索

ネイルやファッションなど、お気に入りデザイン画像に近い画像をネット検索

- ターゲティング広告

複数アプリの利用状況や口コミ情報を分析し広告を配信

- AI教師

学習ソフトの回答プロセスを分析し、問題の出し方を変更・分析

ディープラーニング 応用例-2

- 人材マッチング

企業・求職者のチェックリストや求人データサイトのヒートマップ分析で精度Up

- 不適切画像のフィルタリング

わいせつ画像などを自動排除

- サーバーのシステム異常の予測

ログやステイタスを分析し、システムの稼働率の向上を行う

- 自動Q&A支援

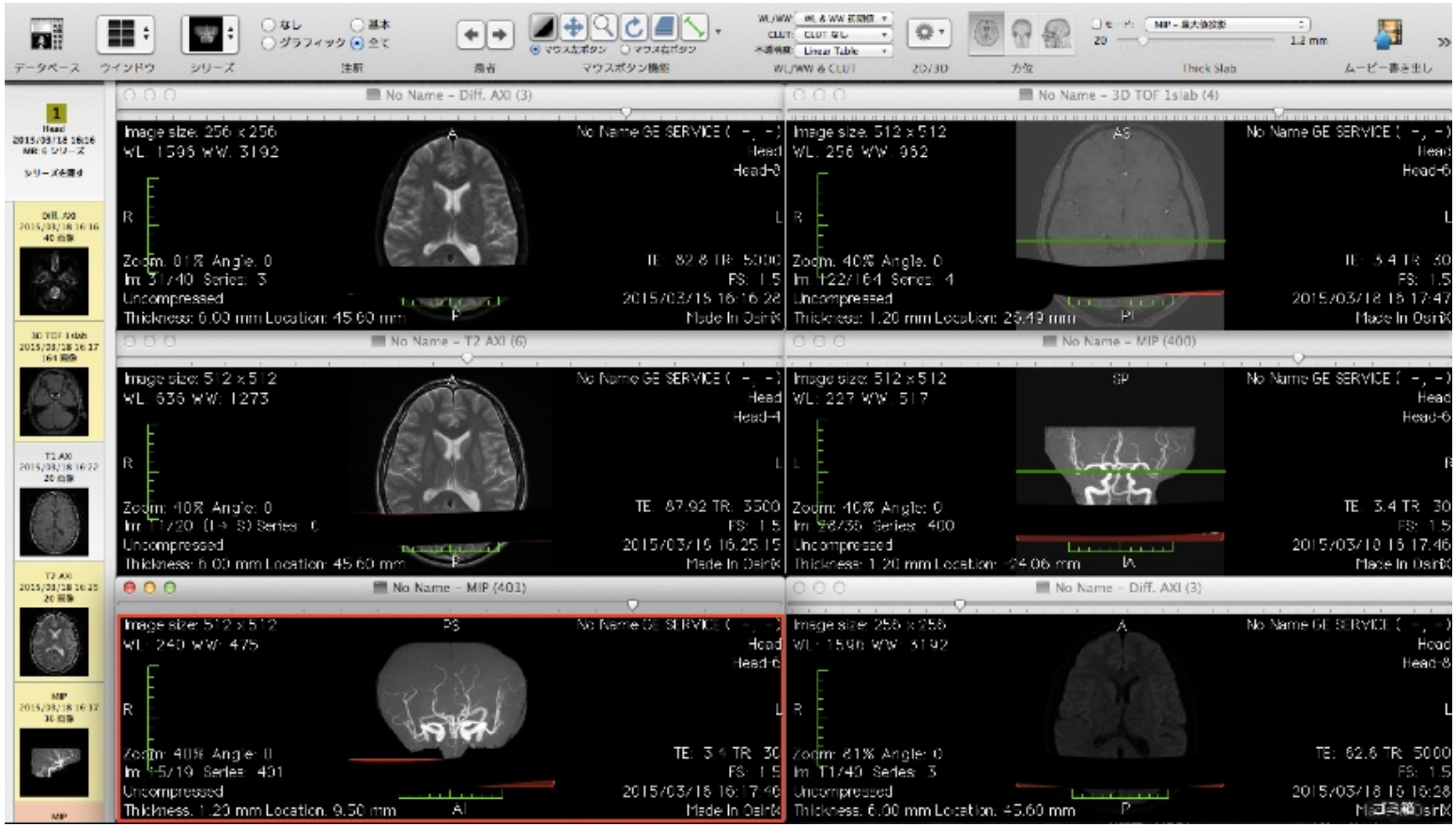
コールセンターやSNSの問い合わせを自動学習し、適切な回答例を支援

- 会話ロボット

自然言語認識、顔認識、音声認識に応用される

ディープラーニング 応用例-3

人工知能で正確な画像診断を行う

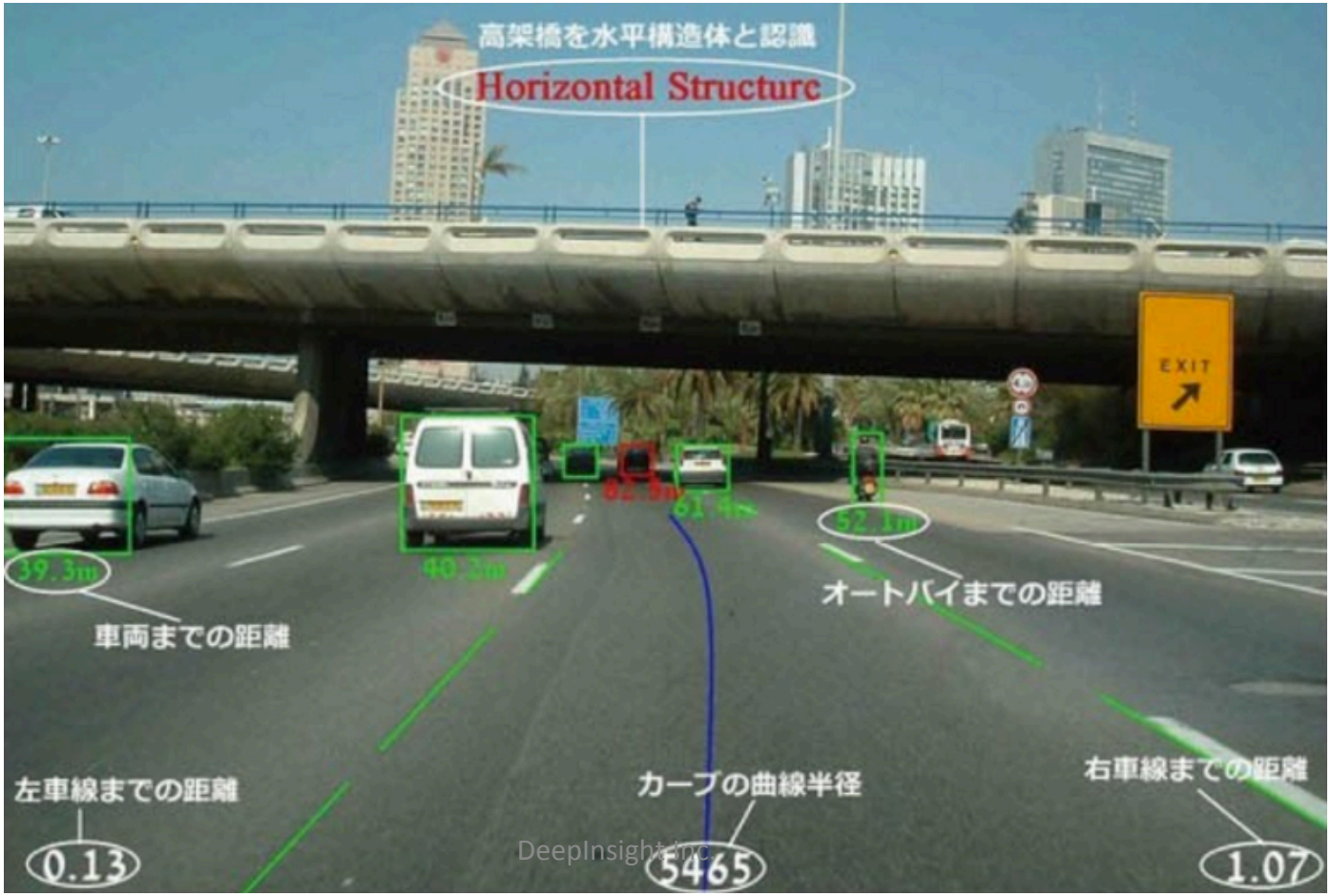


日本でも日本医療情報学会 医用知能情報学研究会と人工知能学会 医用人工知能研究会 (SIG-AIMED) が設置され、2015年9月29日(火)に合同研究会が開催された。

ディープラーニング 応用例-4

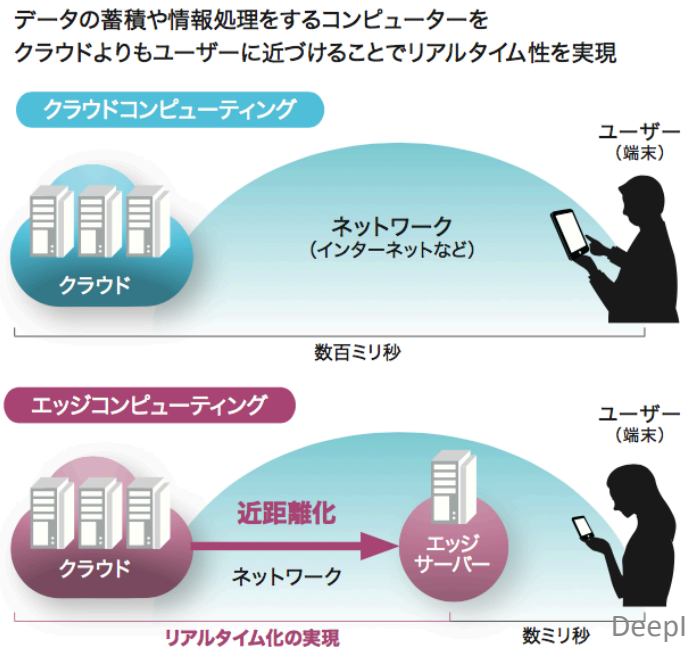
- 自動運転

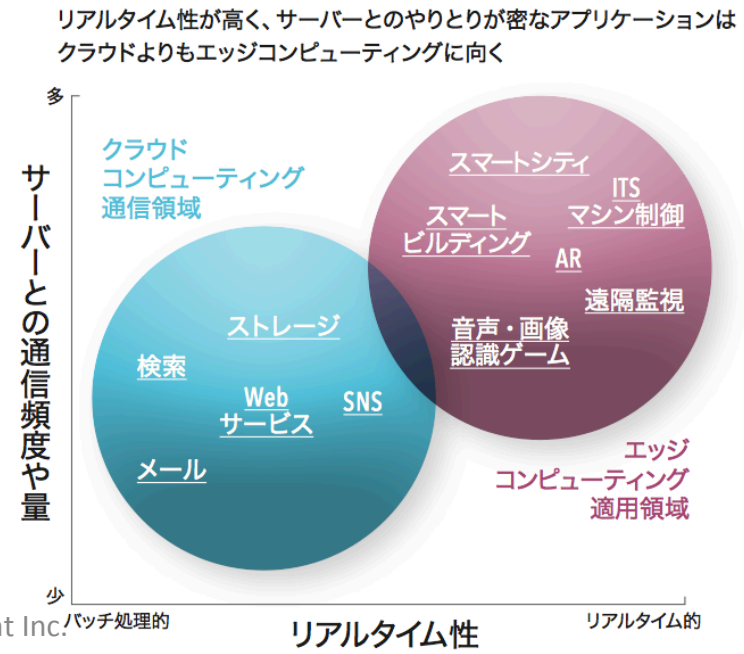
画像認識、位置情報など各種用途に応用が期待されている



エッジコンピューティングとは

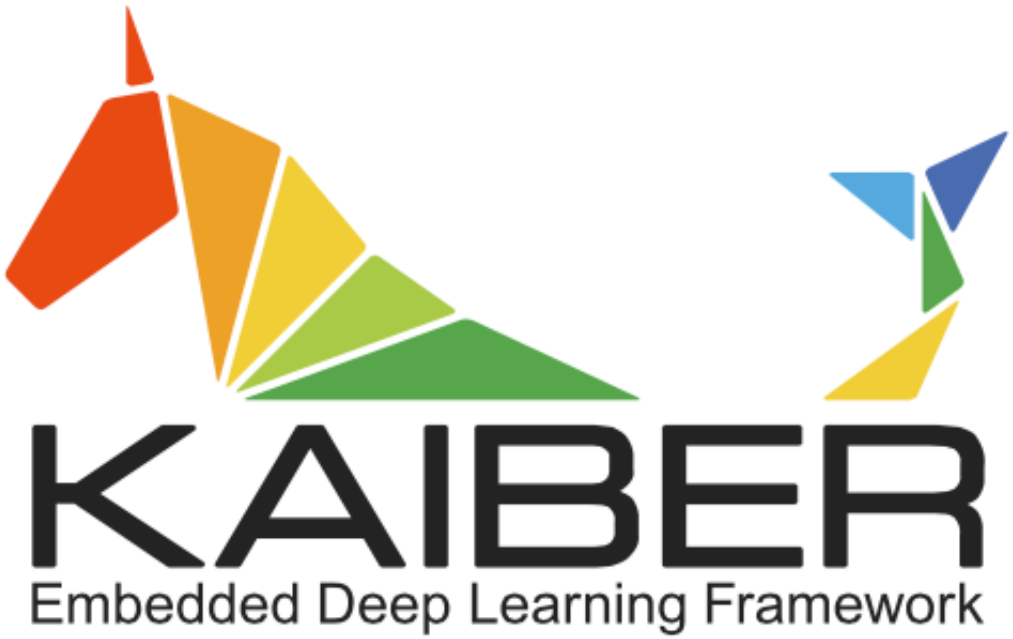
- ユーザと物理的近距離にあるエッジサーバ上で処理を実行し、遅延要求の厳しいリアルタイム・アプリケーションを実現。
- 地域性の高いM2M、ビッグデータの一次処理をエッジサーバで行い計算を効率化、データセンタに集約する為のネットワーク帯域を削減。
- 端末の高負荷処理をエッジサーバで分散処理することで、端末性能に依存せず、高速な処理が可能。





エンベデッド ディープラーニング フレームワーク
「KAIBER」

（カイバー）



エンベデッド ディープラーニング フレームワーク
2016年10月3日「KAIBER」製品発表

エンベデッド ディープラーニング フレームワーク

「KAIBER」とは？

（カイバー）

「KAIBER」は、IoT分野等で重要なトレンドになると予想される「エッジコンピューティング」向けに特化して開発されたエンベデッド ディープラーニング フレームワークです。

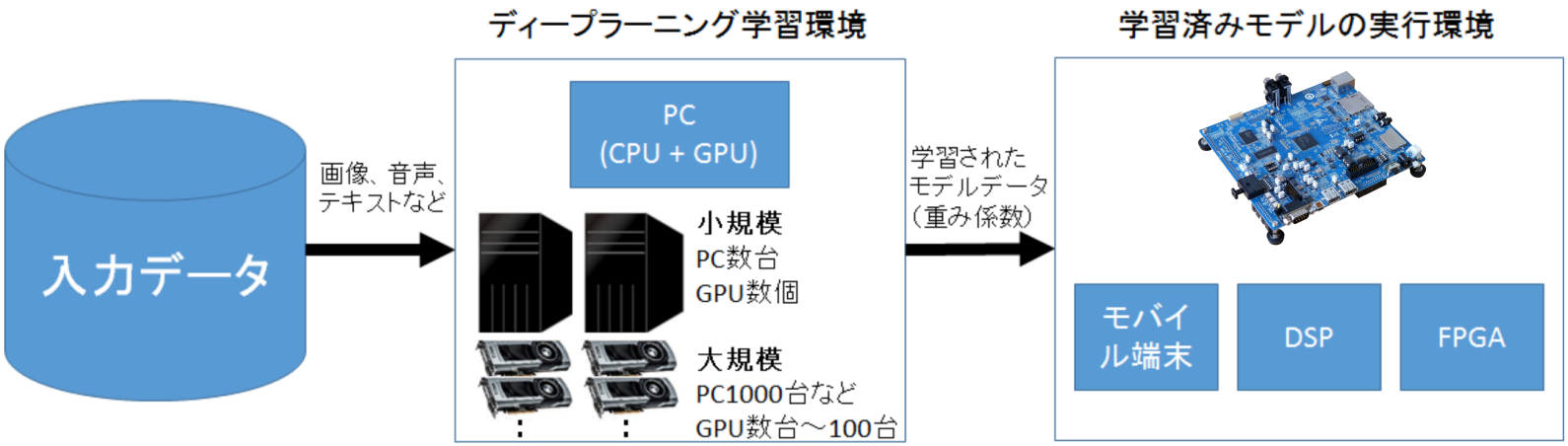
- 組込み用途に特化した国産初の汎用深層学習フレームワーク
- オープンソースでなく、商用サポートを提供
- 組込みに適した推論実行環境と学習環境のモジュール化構造
- 業界最小クラスの省スペース・省リソース設計

  （フットプリント：組込みモジュール最小20K～＋学習済データサイズはシステムに依存）
- Non OS環境でも動作可能(推論実行エンジン)
- 純国産によりマイコン・GPU・FPGA等の独自デバイスに最適化可能
- プラグインにより開発評価環境を独自拡張可能
- 学習中に推論精度をリアルタイムに可視化可能（On-The-Fly Learning機能）
- 組込み向け学習済みデータ圧縮機能（実装予定）

概要（システム構成）

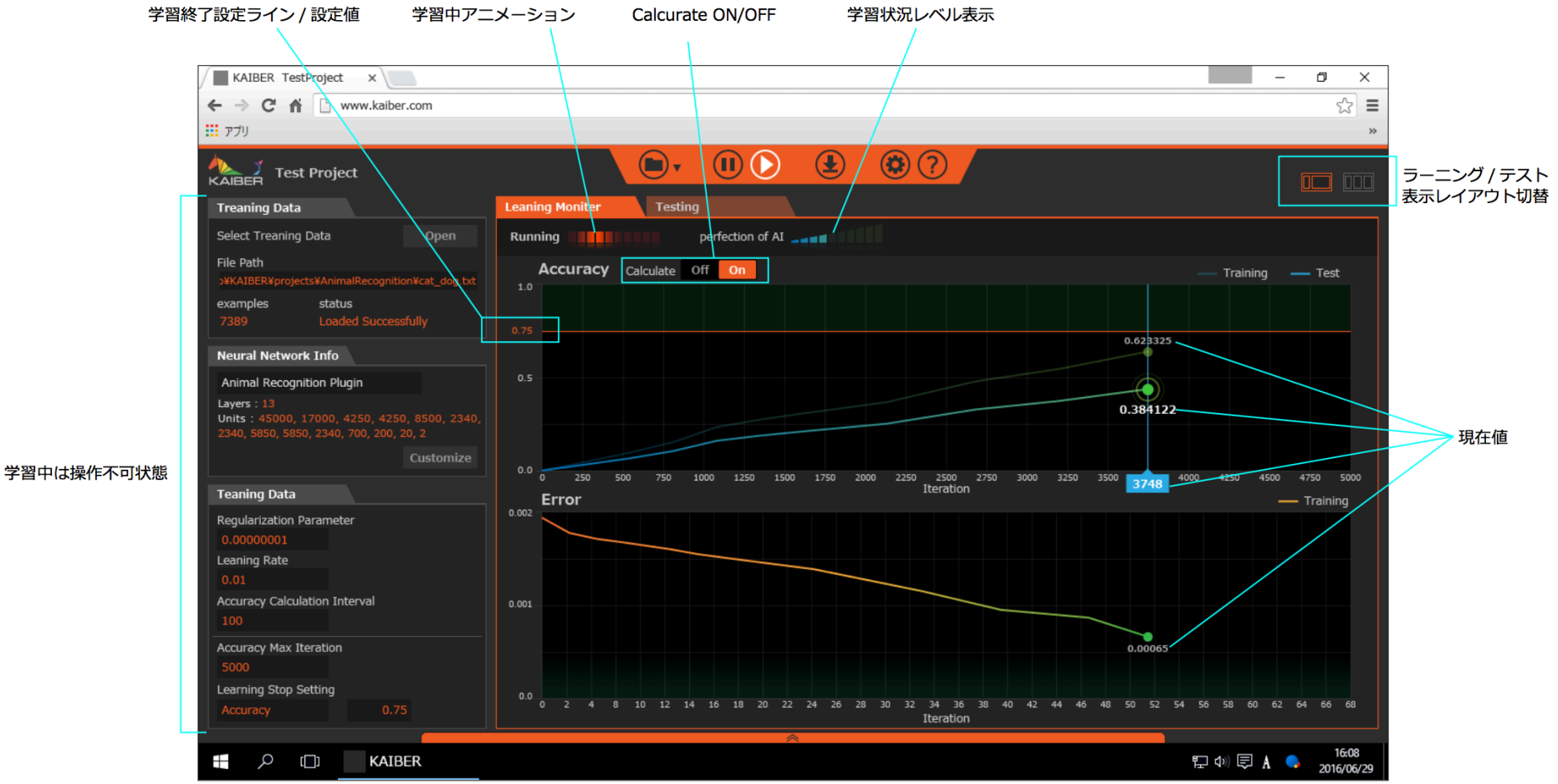
- 「KAIBER」推論実行エンジンはCライブラリーで提供
- 学習サーバ環境はオンプレミス利用（重要データのセキュリティ管理性が向上）
- 学習サーバ環境はNVIDIA GPUに対応
- RaspbianとWindows DLL MKL対応の推論実行エンジン標準搭載
- 学習環境サーバーの月額使用料課金方式（機能保守サービス含む）

  標準版以外の組込み向けは月額課金＋ロイヤリティモデル（別途契約）



概要

先進のGUIを搭載



（次期バージョンに搭載予定/開発中）

概要のポイント

- 機能優位性

組込み向けに特化した小さなフットプリント

完全分離型構造の採用によりセンサーやスマホなどスケーラブルにサポート可能

最新のディープラーニング技術の実装

進化する技術をコンスタントに追加し、組込み分野に新しい可能性を提供

優れた開発・運用効率

JavaとC言語の採用により、多様なデバイスへも迅速に組込み可能。企業の情報システムへの統合も容易

- ビジネス優位性

国産フレームワークの商用サポート

完全自社開発により最適な機能追加と迅速なメンテナンスを提供

推論実行モジュール(スマホAndroid/iOS版)は無料・無制限(カスタマイズ版は除く)

柔軟な契約・課金モデル(無償とロイヤリティの2方式)により、開発コスト計算が容易

機能拡張への柔軟な対応

変化の激しい技術トレンドに対応できるモジュール化構造とサポート体制

DL　比較資料

	KAIBER	Uncanny DL	Chainer	TensorFlow	Caffe
開発者	Deep Insight (JP)	Uncanny Vision(Ind)	Preferred Networks(JP)	Google(US)	BVLC(US) UC Berkeley
主要開発言語	Java C(推論エンジン)	C	Python	C++	C++
商用サポート	◎ 商用サポート提供	○ 海外サポート	△提携企業のみ オープンソース	X オープンソース	X オープンソース
最適化対応	◎ マイコン・FPGA	○ ARM, Intel中心	△ 限定提携企業のみ	X	X
サポートOS	Linux, Win	Caffe準拠 組込みライブラリーのみ開発	Ubuntu, Cent, Mac OS	Linux, Mac OS	Linux, Win Mac OS
デバイス対応	◎独自仕様チップ対応可	○ 個別対応	△ DIMo(ネット前提)	△ Android, iOS	X
組込みレベル	◎ 各種マイコンまで可	○ ARM, Intel専用	X	X	X

組込みDL 比較資料

- Uncanny DL (Uncanny Vision)

学習機能はオープンソースCaffeなどを使用

将来エッジ側の学習機能配備に対応不可、オープンソースで商用サポートなし

画像認識（CNN)のみのライブラリー

CNN強化版やRNN系への対応がない

ARM NEON & Intel Atomにのみ対応

個別ユーザーのFPGAなどは国内対応できない

- Chainer (Preferred Networks)

オープンソースのフレームワーク

基本的に提携企業のみにプロジェクト単位での限定サポート提供

Pythonベースの一体型構造

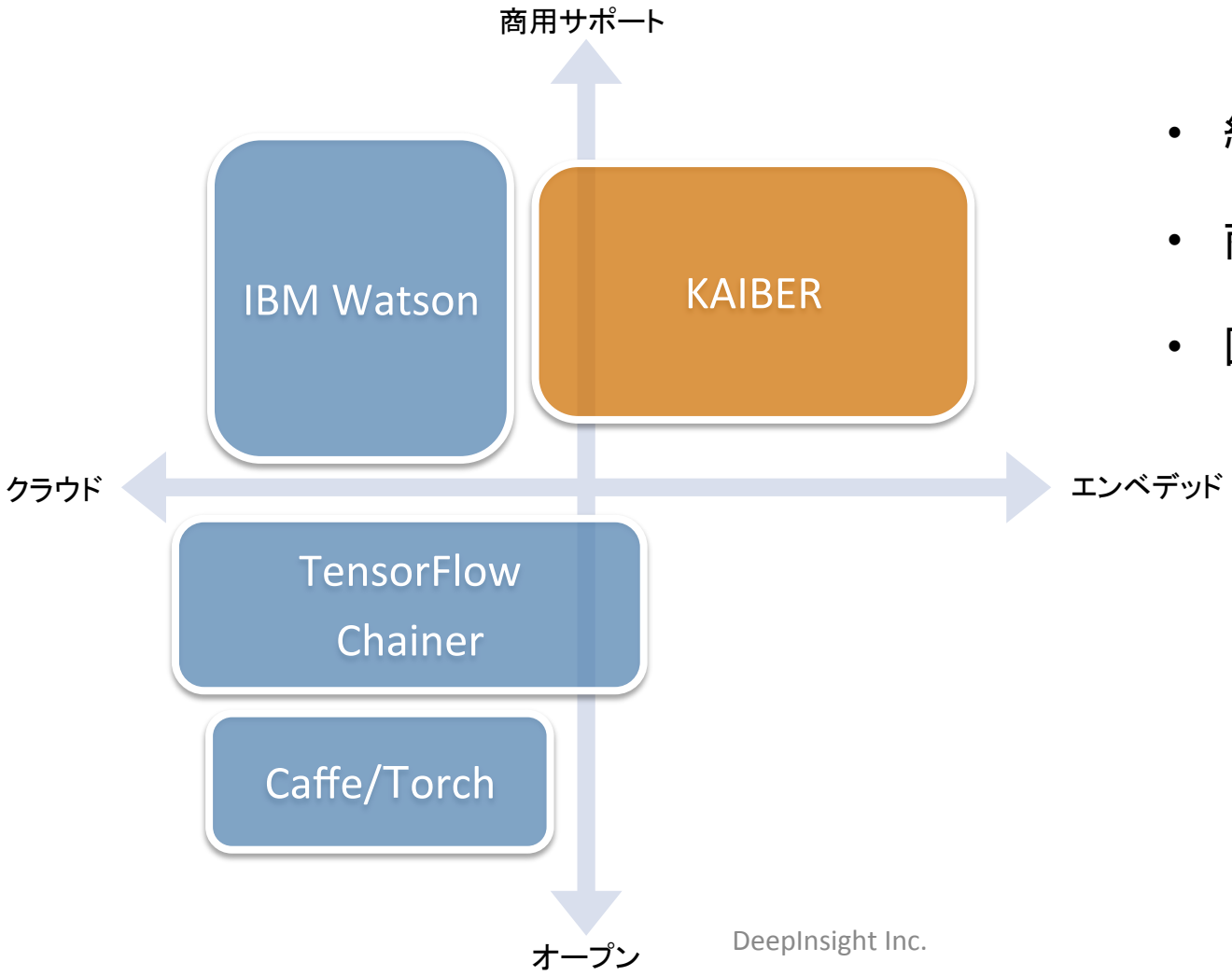
稼働環境はラズベリーパイ程度までを想定し、Linuxのみサポート

DIMoはネットワーク前提のIoTソリューション

推論はネット経由の上位層のChainerが処理し、スタンドアロンの組込みDLではない

協業モデル例とポジショニング

- 搭載チップに最適化したディープラーニングモジュールの開発



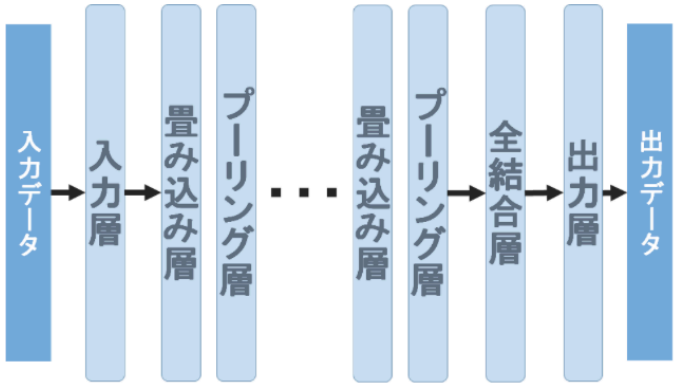
- 組込み環境に特化
- 商用サポートを提供
- 国産製品で協業が容易

実装テクノロジー

正解率を向上させるための手法

- 畳み込みニューラルネットワークConvolutional Neural Network（CNN）

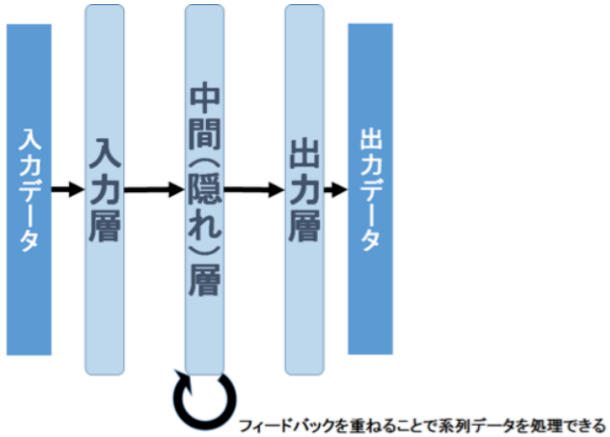
主に画像や自然言語処理に使われており、特に画像認識分野での実用化は急速に進んでいる。中間層は主に畳み込み（convolution）層とプーリング（サブサンプリング）層を交互に繰り返すことでデータの特徴を抽出し、最後に全結合層で認識を行う。



CNN

- Recurrent Neural Network（RNN）

時系列キーフレームを複数セットで解析する動画分類や、自然言語処理・音声認識での言語モデル、ロボットの行動制御などに使われる。このモデルの特徴は、唯一中間層への自己フィードバックができる点にある。例えば、前時刻の層の出力を考慮して現中間層の出力を計算したり、次時刻の層の出力を考慮して現中間層へと両方向に情報をフィードバックが可能。他のネットワークとは違い、系列データへの対応と応用範囲が広い。



RNN

実装テクノロジー(2)

正解率を向上させるための手法

- 正則化(regularization)

過学習を避けるための手法。重み減衰(weight decay)、または重み上限(max norm) のどちらかを使用する。

- ドロップアウト(dropout)

ランダムに入力データの一部を隠蔽して学習させることにより、1 個のニューラルネットワークの中に複数個のニューラルネットワークが入っているように学習を行うことができる。

(一般的に、複数個のニューラルネットワークを別個に学習させ、それらのアウトプットの平均を取ると正解率が向上することが知られている。ドロップアウトを使用すると、1 個のニューラルネットワークで複数個のニューラルネットワークの効果を得ることができる。)

- 自己符号化器(autoencoder)、スパース自己符号化器(sparse autoencoder)

ニューラルネットワークの各層を教師無しデータで学習させることによって、トレーニングデータの冗長な特徴から少数の特徴をうまく選び出すことができる。(自己符号化器を使用して各層の重みを初期化することにより、学習が効率的に行われることが知られている。)

- 白色化(whitening)

トレーニングデータを変換して偏りを除去する手法。PCA(Principal Component Analysis) と ZCA(Zero-phase Component Analysis) という手法がある。(トレーニングデータに偏り(特徴同士に相関関係がある場合)があると学習がうまく行われない場合があるため)

実装テクノロジー(3)

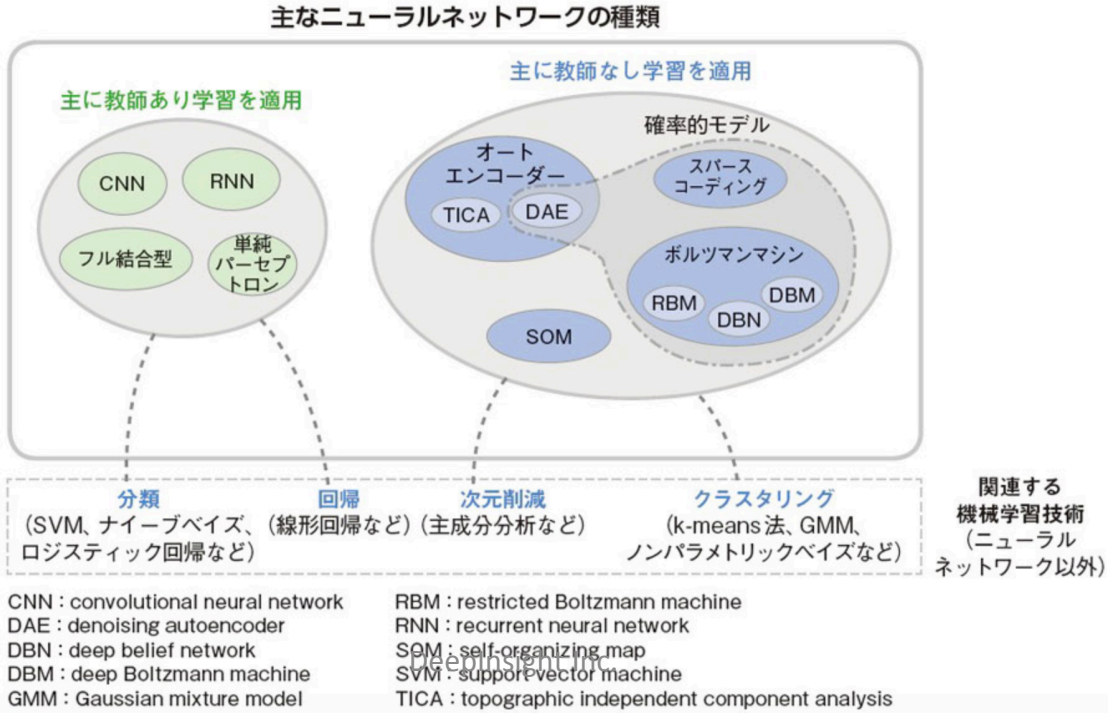
効率的に学習(トレーニング)を行うための手法

- ミニバッチ確率的勾配降下法(minibatch stochastic gradient descent)

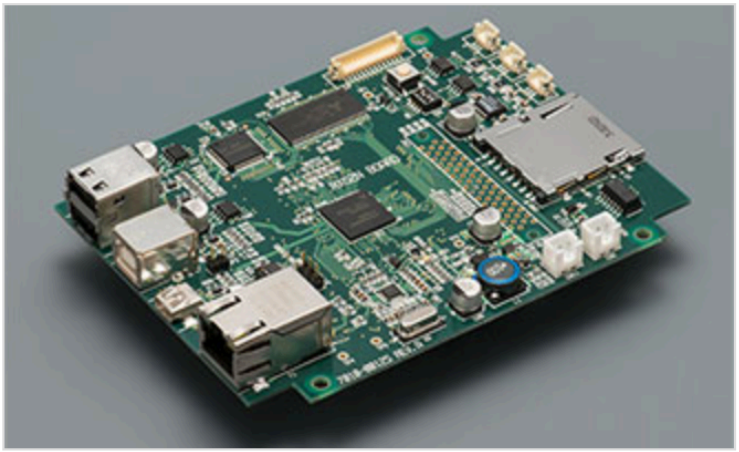
トレーニングデータの中からランダムに少数のデータを選択し、それらのデータに対して勾配を計算し重みを更新する手法。バッチ学習(トレーニングデータ全体を使用する手法)に比べて、誤差の大きい極小解に捕らわれるリスクを軽減できる。

- モメンタム(momentum)

重みの更新量を調整することにより、効率的に極小解にたどり着くことができる。



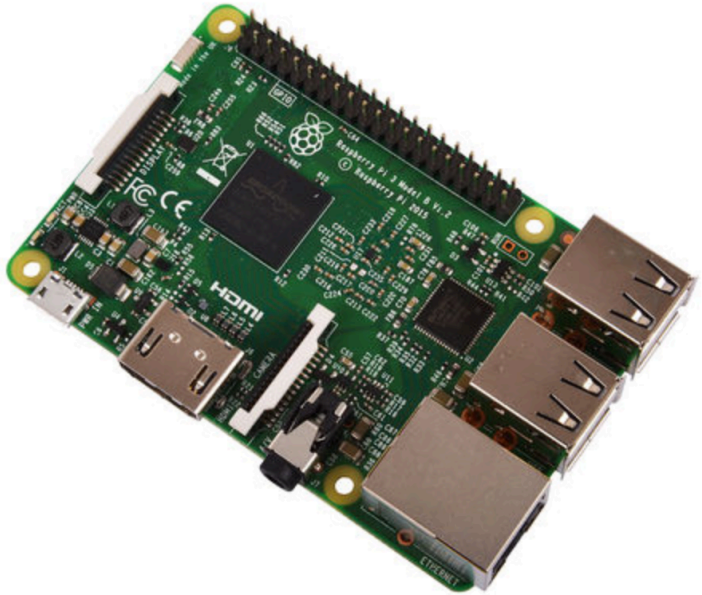
デモ環境 性能評価



データテクノロジー製ESPT-RX
CPU：ルネサス社製 RX63N　　100MHz

監視カメラによる数字の画像認識デモ(CNN)

DNNの処理時間：50ms
DNNの学習済みデータ：1.8MB（テキスト）
DNNの層：7
KAIBERコードサイズ：5KB



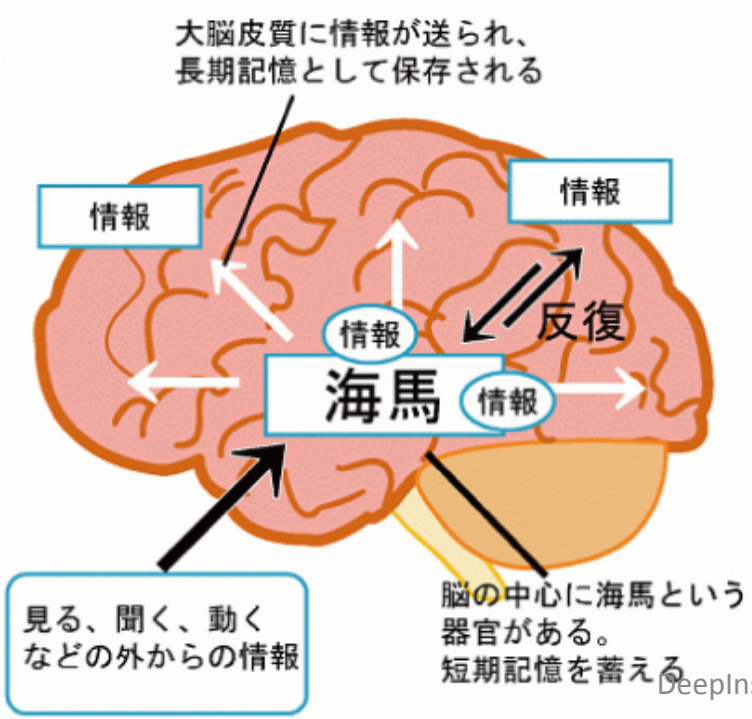
Raspberry Pi 3 Model B
CPU:Broadcom BCM2837（ARM Cortex-A53）　1.2GHz

ラズパイ接続カメラによるモデルカー
ジェスチャーによる画像認識制御のデモ（CNN）

DNNの処理時間：60ms
DNNの学習済みデータ：2MB(テキスト)
DNNの層：10
KAIBERコードサイズ：29KB

海馬とは

- 「海馬」（かいば・hippocampus）とは、特徴的な層構造を持つ記憶や空間学習能力に関わる脳の器官です。タツノオトシゴのような形をしており、外部からの情報を整理する司令塔です。重要な情報は神経細胞を形成し保存、不要な情報は消去します。必要な古い情報のみ大脳皮質に転写・保存します。
  ギリシャ神話の海神ポセイドンがまたがる架空の動物「海馬」の尾が似ている事から由来しています。



- 連絡先 -
ディープインサイト株式会社
久保田　良則
yoshinori.kubota@deepinsight.co.jp

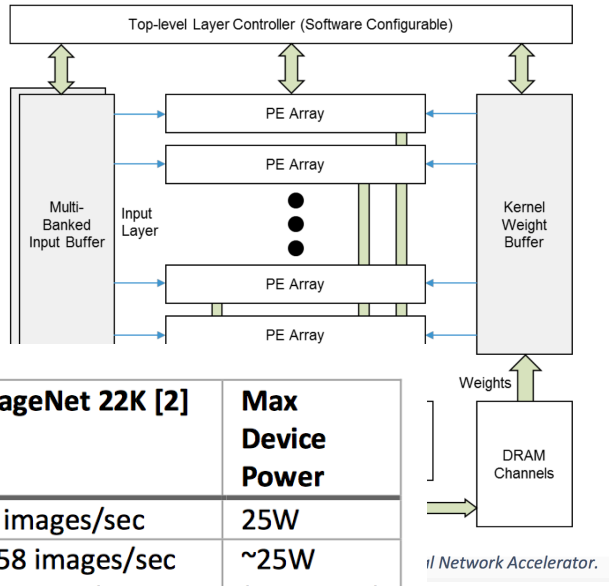
自動運転とディープラーニング

- **GPUからFPGAへ**

現在、ディープラーニングで主流となっている学習データの計算リソースはGPUが主流となっているが、今後は電力効率と最適化にすぐれたFPGAに置き換わっていく可能性がある。既に事例が出てきており、自動運転の車載デバイスとしては有望。

Accelerating Deep Convolutional Neural Networks Using Specialized Hardware

Kalin Ovtcharov, Olatunji Ruwase, Joo-Young Kim, Jeremy Fowers, Karin Strauss, Eric S. Chung
Microsoft Research
2/22/2015



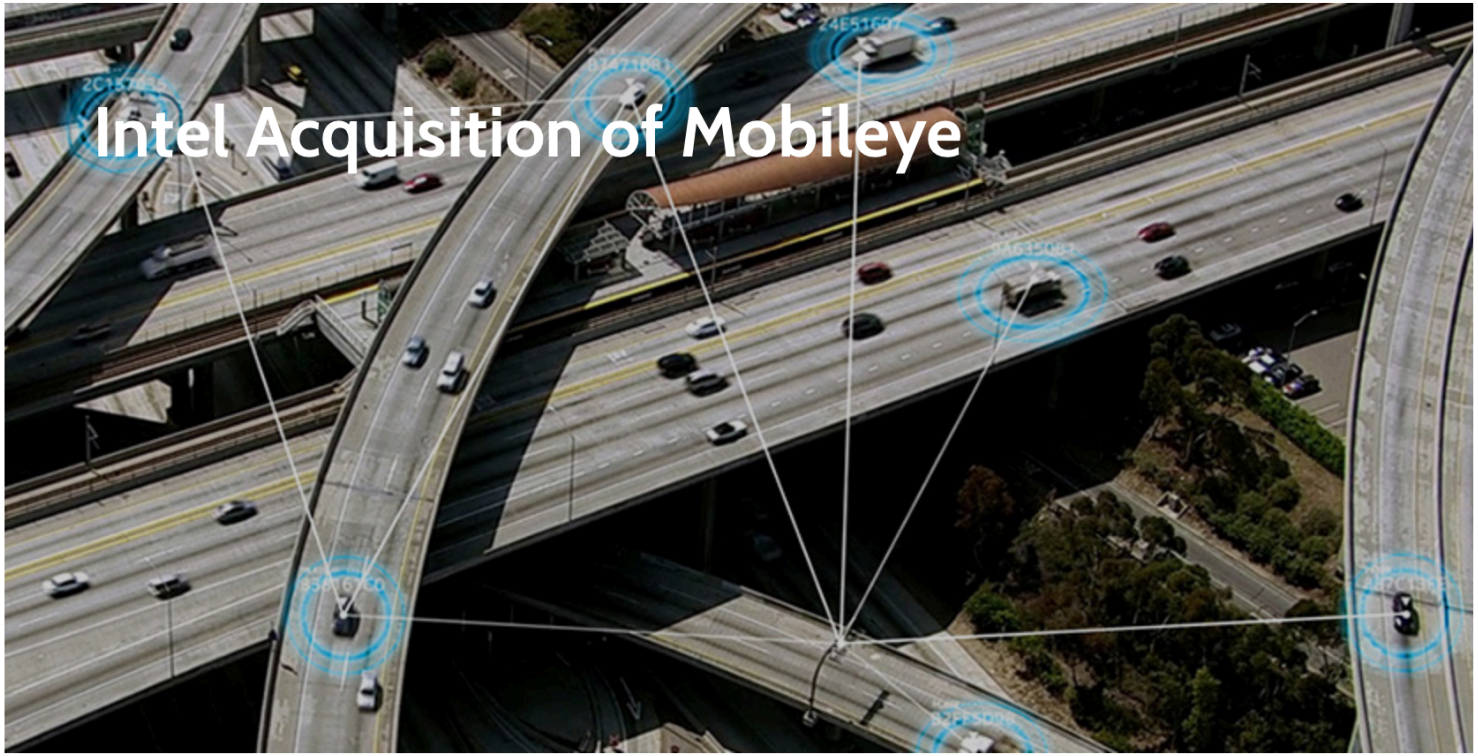
l Network Accelerator.

	CIFAR-10 [4]	ImageNet 1K [1]	ImageNet 22K [2]	Max Device Power
Catapult Server + Stratix V D5 **[3]**	2318 images/s	134 images/sec	91 images/sec	25W
Catapult Server + Arria 10 GX1150 **[8]**	-	~233 images/sec (projected)	~158 images/sec (projected)	~25W (projected)
Best prior CNN on Virtex 7 485T **[5]**	-	46 images/sec[3]	-	-
Caffe+cuDNN on Tesla K20 **[6]**	-	376 images/sec	-	235W
Caffe+cuDNN on Tesla K40 **[6]**	-	500-824 images/sec[4]	-	235W

Table 1: Comparison of Image Classification Throughput and Power.

Appendix

インテルが自動運転技術のモービルアイを買収

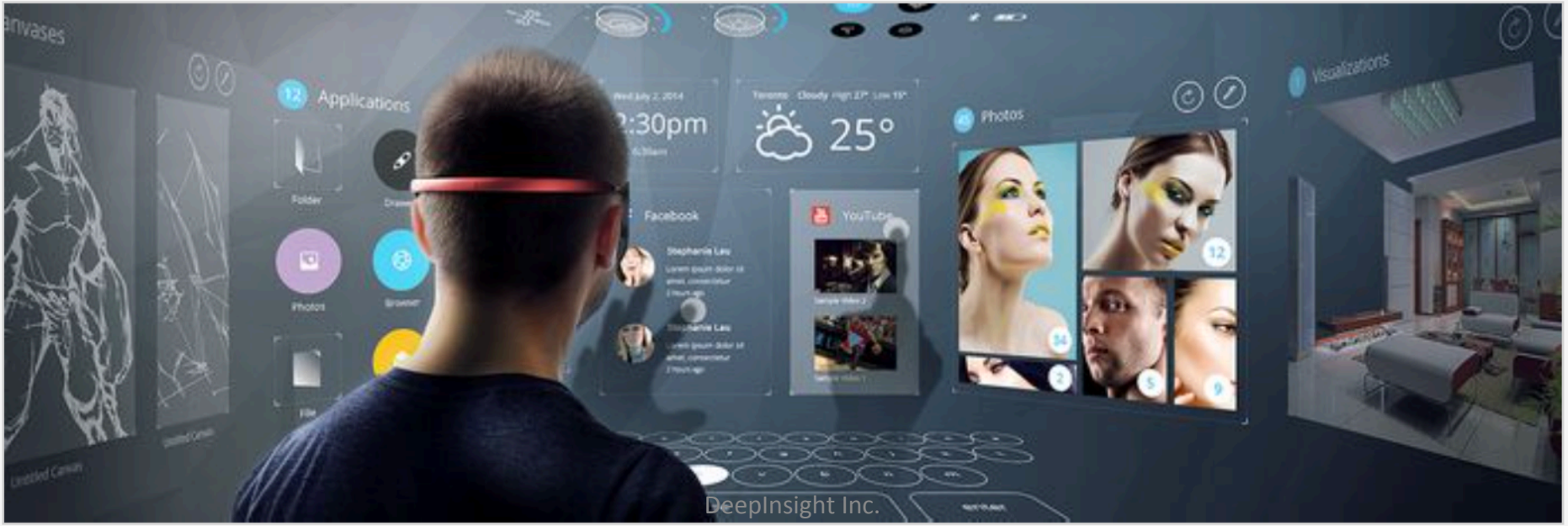


Intel and Mobileye announced on March 13 that they have entered into a definitive agreement pursuant to which Intel will acquire Mobileye. Under the terms of the agreement, a subsidiary of Intel will commence a tender offer to acquire all of the issued and outstanding ordinary shares of Mobileye for $63.54 per share in cash, representing a fully-diluted equity value of approximately $15.3 billion and an enterprise value of $14.7 billion.

Appendix

VRへのディープラーニングの応用

VRヘッドマウント前面のカメラによりジェスチャーを認識
組み込みDLモジュールによるリアルタイム操作の実現
安価なカメラによる画像認識が可能
360度カメラによる広範囲な操作性も実現



提案例：ソーシャルメディア画像検索分析

ディープラーニングによる画像認識技術を使うことによって、ソーシャルメディア上に投稿された写真から貴社のブランドロゴや製品を自動的に認識し、誰が、いつ、どこで、どのように貴社のブランドと触れあったのかという情報を抽出、これからのマーケティング分析に活用するコア技術となります。

- 隠れたファンを見つける事が出来る
- 競合他社のファンを見つける事が出来る
- オピニオンリーダーを見つける事が出来る
- どの商品が市場を牽引しているのかを確認できる。
- 位置情報も確認できる。
- どんな製品が人気になり、現代のトレンドとなっているのか把握出来る
- 誰が、いつ、どこでブランド写真の画像をソーシャルメディア上に投稿したかを把握。
- 貴社のブランドの一番のファンの年齢層や性別も把握出来る。
- 貴社のオフライン上のファンをソーシャルメディア上のファンに転換することが出来る。
- ユーザー自身が作成したコンテンツを使ったキャンペーンにより貴社とファンのエンゲージメントを強化出来る。