説明可能なAI

2021年12月13日(月)

第13回 自動チューニング技術の現状と応用に関するシンポジウム（ATTA2021）

【招待講演】

ソニーグループ（株）R&Dセンター

シニアマシンラーニングリサーチャー

鈴木　健二

自己紹介

鈴木　健二

博士（工学），学士（法学）

シニアマシンラーニングリサーチャー

職歴

1999年　東京大学 生産技術研究所

2000年　フランス Institut d'Électronique et de Microélectronique du Nord (IEMN)

2001年　ソニー（株）

現在　　ソニーグループ（株） R&Dセンター

研究開発テーマ

説明可能なAI，機械学習における公平性，データ流通

Agenda

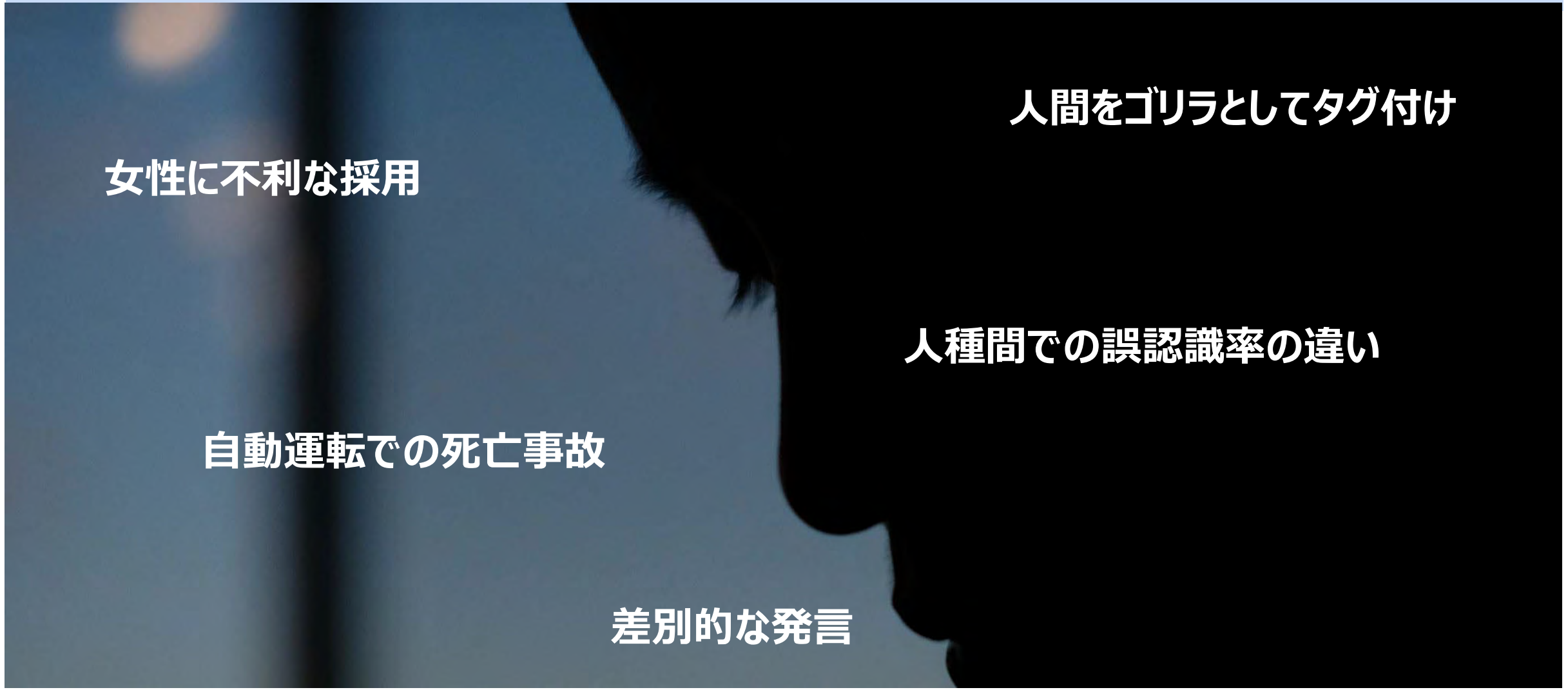
- AI倫理と技術
- 判断根拠の可視化手法
- モデルへのデータ影響度
  データクレンジング, ミスラベル検知
- データの不確実性
- 近年の動向や課題
- 説明可能なAIツール

AI倫理と技術

AI倫理の問題

（社会的課題）　AIの社会的受容性の向上

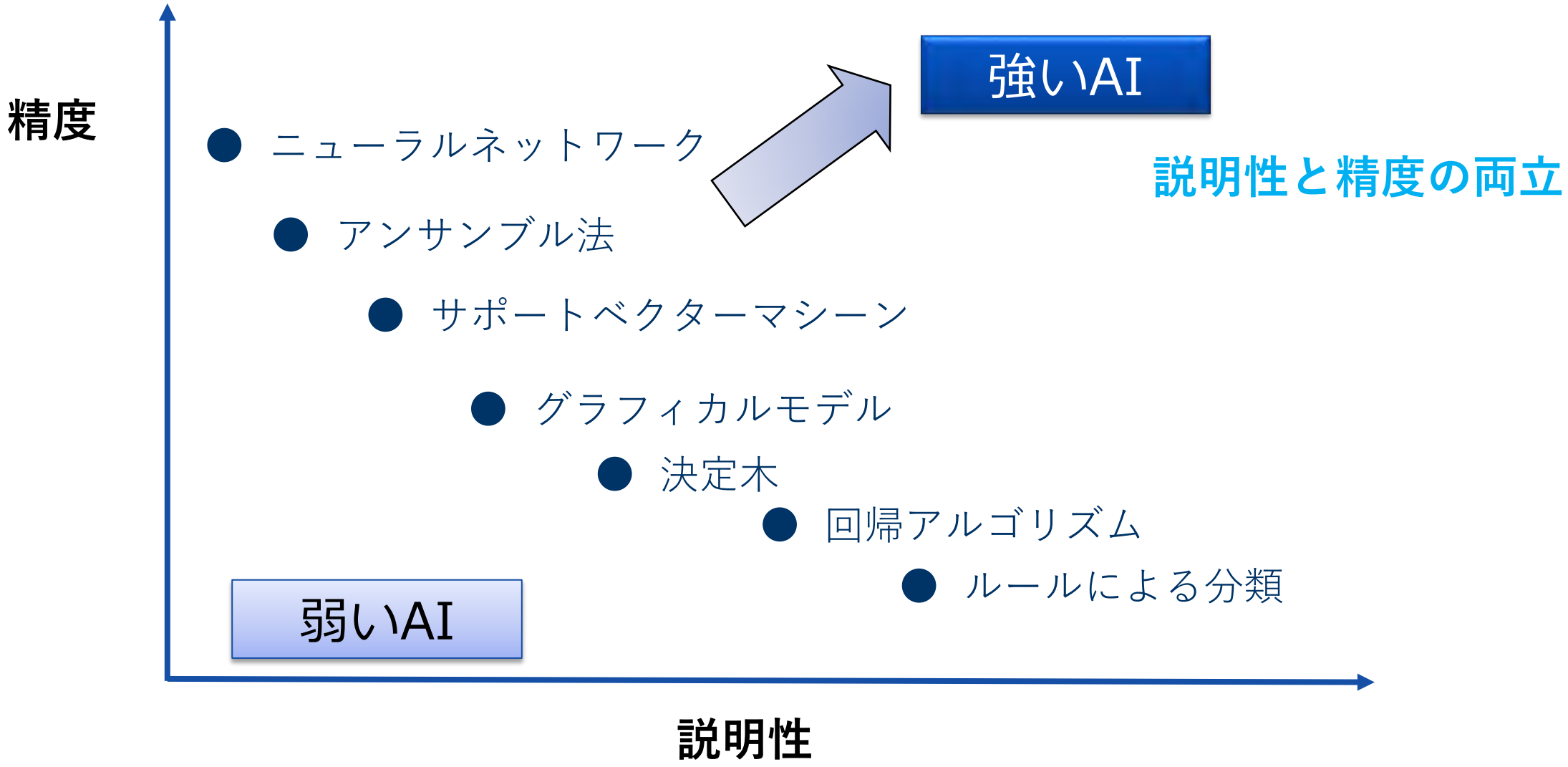
（技術的課題）　AIの説明性の解明と公平性



AIに関する法制度・ガイドラインの動向

GDPR 一般データ保護規則	欧州	2018	決定に含まれているロジックに関する意味のある情報等を提供しなければならない（13～15条）. 完全自動意思決定の原則禁止（22条）.
AI利活用ガイドライン	日本	2019	透明性の原則, アカウンタビリティの原則
欧州AI規制法案	欧州	2021	AIの包括的な規制法案. 許容できないAI, ハイリスクAI, 特定のAIシステム, 最小リスクAIにカテゴライズされ, その義務や要件を規定.

機械学習における説明性と精度



説明可能なAI

- AIはブラックボックス
- 判断根拠がわからない
- 説明できないのか?

AIは人間を超える性能を持つ。その判断の根拠はどうなっているのだろうか？

AIの判断根拠が必要となる例

AI利活用

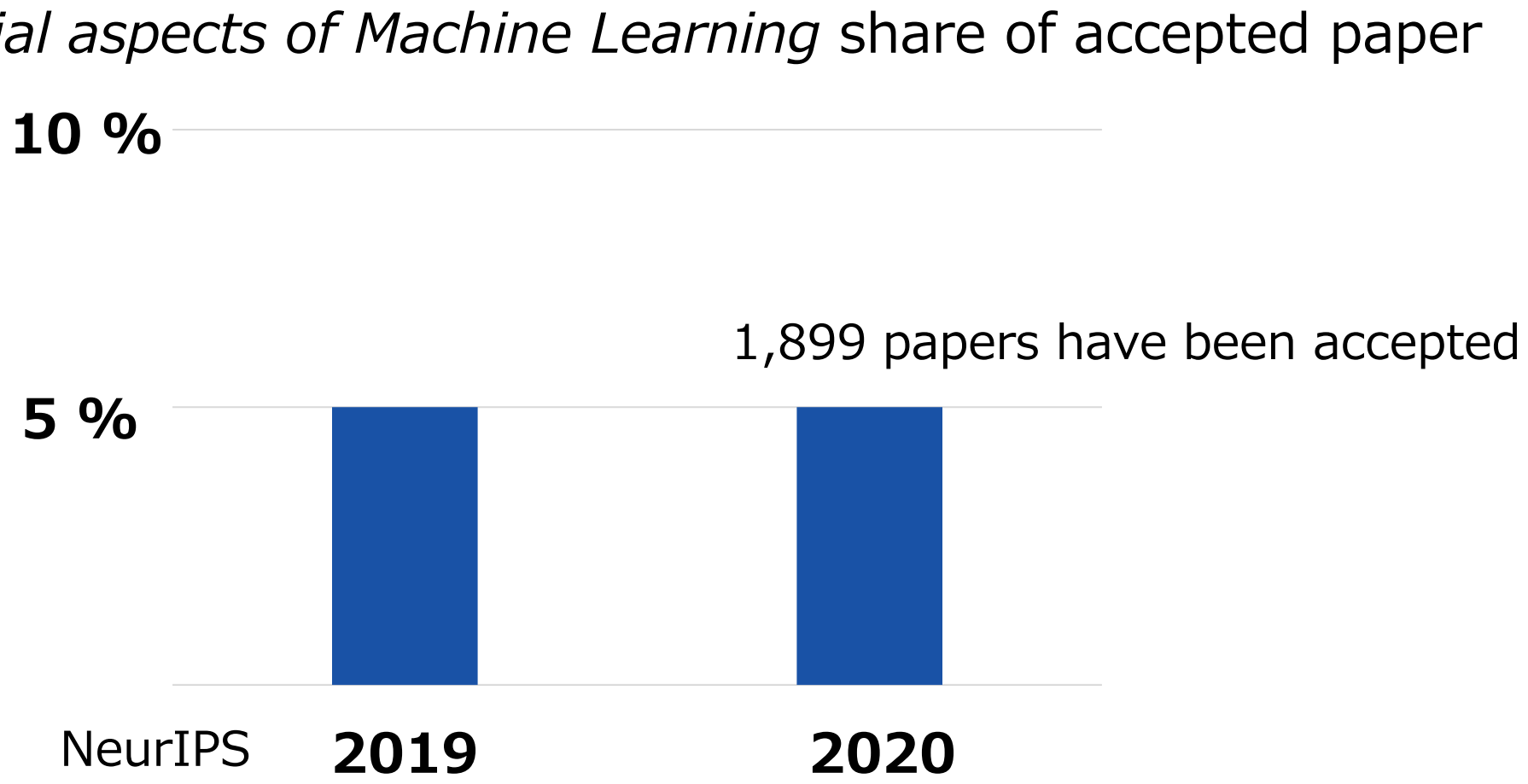
優れている点	課題
高い精度	判断理由の説明

医師による診断でのAI利活用のケース

AIの判断理由が分からないと, 原因の解明ができない.

AIトップカンファレンスNeurIPS2020におけるAI倫理技術関連論文

Fairness, Accountability, Transparency, Explainability, Privacy etc.



AIトップカンファレンスNeurIPS2020 投稿論文のAI倫理レビュー

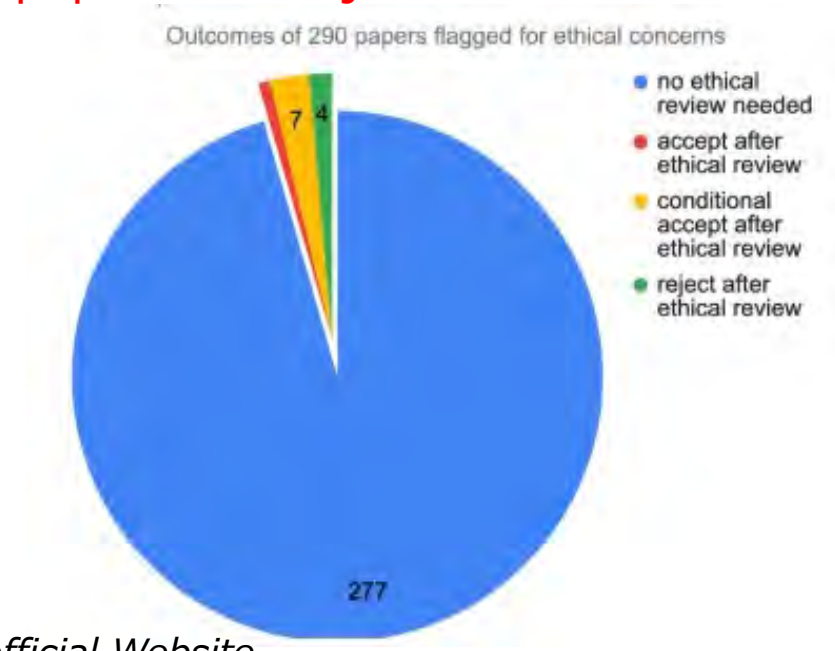
Machine learning has real-world impact

Ethical review

Statement added to the Call for Papers :
"Regardless of scientific quality or contribution, a submission may be rejected for ethical considerations, including methods, applications, or data that create or reinforce unfair bias or that have a primary purpose of harm or injury."

Results of ethical review

- Outcomes of 290 papers flagged for ethical concerns
- 13 papers were ethically reviewed.
- 2 papers were accepted.
- 7 were conditionally accepted.
- 4 paper were rejected.



Refer from NeurIPS2020 official Website

AIトップカンファレンスNeurIPS2020からみるAI倫理技術のトレンド

Explainability beyond classification

More reliable and robustness explanation methods

Rigorous evaluation of explanation

Theoretical analysis of explanation

Bias mitigation with fairness

システム1からシステム2ディープラーニングへ

人間的な思考へ

現在のディプラーニング

- 直感的
- 高速
- 無意識
- 非言語的
- 習慣的

将来のディプラーニング

- 論理的
- 低速
- 意識
- 言語
- 計画
- 論拠

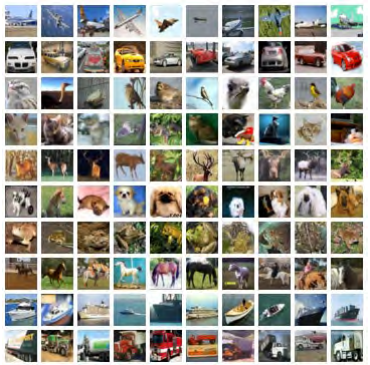
ハイレベルの概念を操作

Yoshua Bengio “From System1 Deep Learning to System2 Deep Learning” NeurIPS 2019 https://youtu.be/T3sxeTgT4qc

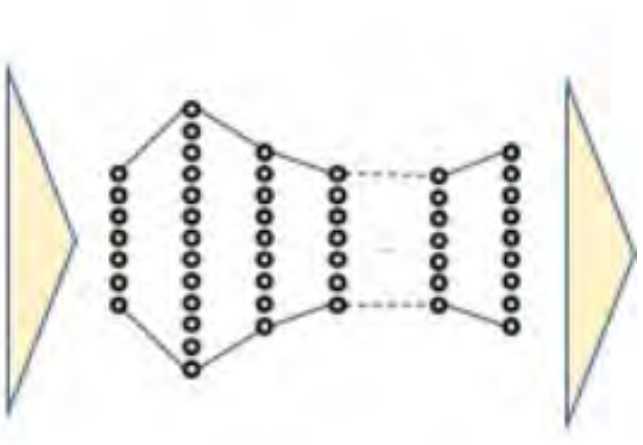
説明可能なAIとは

説明可能なAI

eXplainable AI (XAI)



データ
Data



ディープニューラルネットワーク
DNN

説明モデル

説明インターフェイス

CIFAR-10 dataset : Learning Multiple Layers of Features from Tiny Images, Alex Krizhevsky, 2009.

説明可能なAIは，AIの判断根拠を可視化することができる技術.

Responsible AIと説明可能なAI

Responsible AI

Fairness
Accountability
Transparency

Defining Sony's stance in relation to AI initiatives and promoting dialogue

Sony Group AI Ethics Guidelines

SDGs

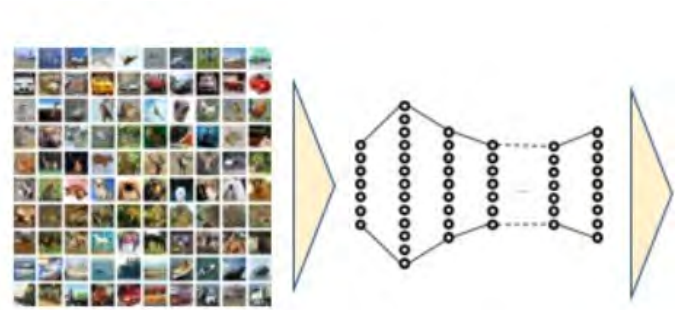
5. Gender Equality

10. Reduced Inequalities

12. Responsible Consumption and production

16. Peace, Justice, and Strong Institutions

eXplainable AI (XAI)



Data DNN

Explanation model

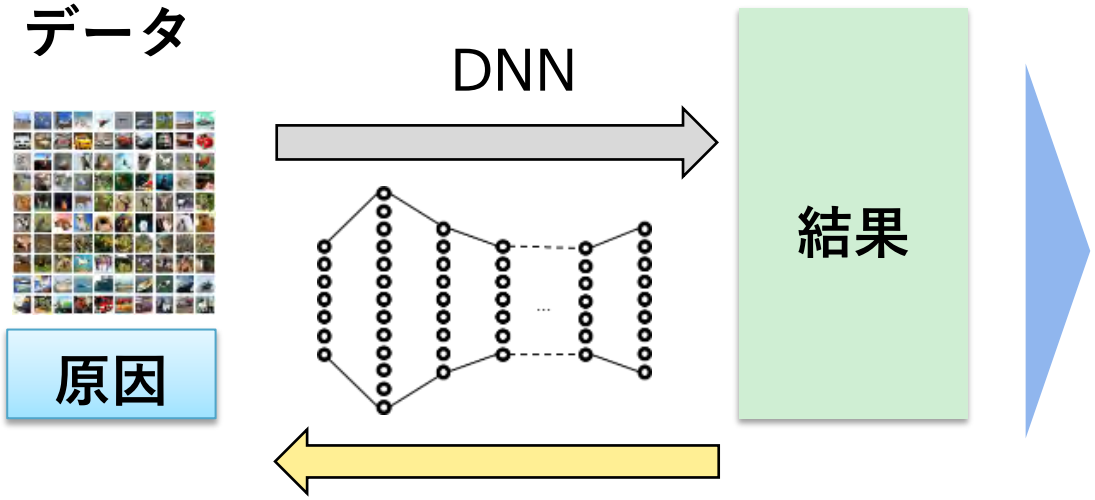
Explanation Interface

- AIの判断の理由を理解
- 人間の意図にあったAIの制御

AI倫理技術の実装


公平性
Fairness
説明可能性
eXplainability
透明性
Transparency
アカウンタビリティ
Accountability
Responsible AI

説明可能なAIの利用価値



CIFAR-10 dataset : Learning Multiple Layers of Features from Tiny Images, Alex Krizhevsky, 2009.

XAI

DNNの潜在的な予測精度を引き出す.

AI倫理、法令遵守 AI ethics, Compliance
画像分類，顔識別, 人物体検出など
画像生成など
金融，マーケティング
自然言語処理
製造，検査
マテリアル・インフォマティクス
医療
自動運転，行動学習
解析全般, 最適化問題
データの質向上

XAI技術は，AI倫理のみならず，多方面へ応用展開できる.

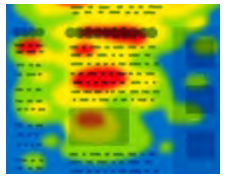
説明可能なAIのスコープ

Fairness Accountability Transparency (FAT)

Improving accuracy is an approach to the FAT issues

* Data cleansing
* Fairness check/definition/solution
  *Data bias mitigation
* Learning Mechanism
* Model debugging
* Adversarial defense
  *Robustness

Visualization
XAI shows the reason of judgement



Understanding
Humans understand the reason

Feedback
XAI corrects AI

説明可能なAIは，AI 倫理だけでなく，ディプラーニングの可能性を更に引き出す.

AIの判断根拠の可視化手法

判断根拠

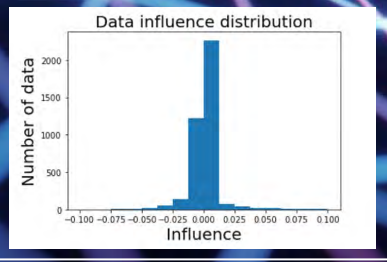
Basis for judge

- Grad-CAM
- LIME
- SHAP
- Smooth Grad
- Attention Branch Network

データの影響度

Data influence

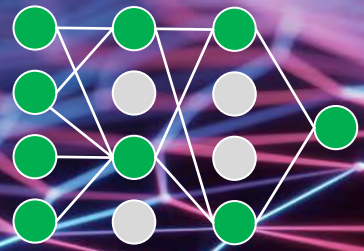
- SGD influence
- Influence functions
- TracIn



データの不確実性

Data uncertainty

- MC dropout

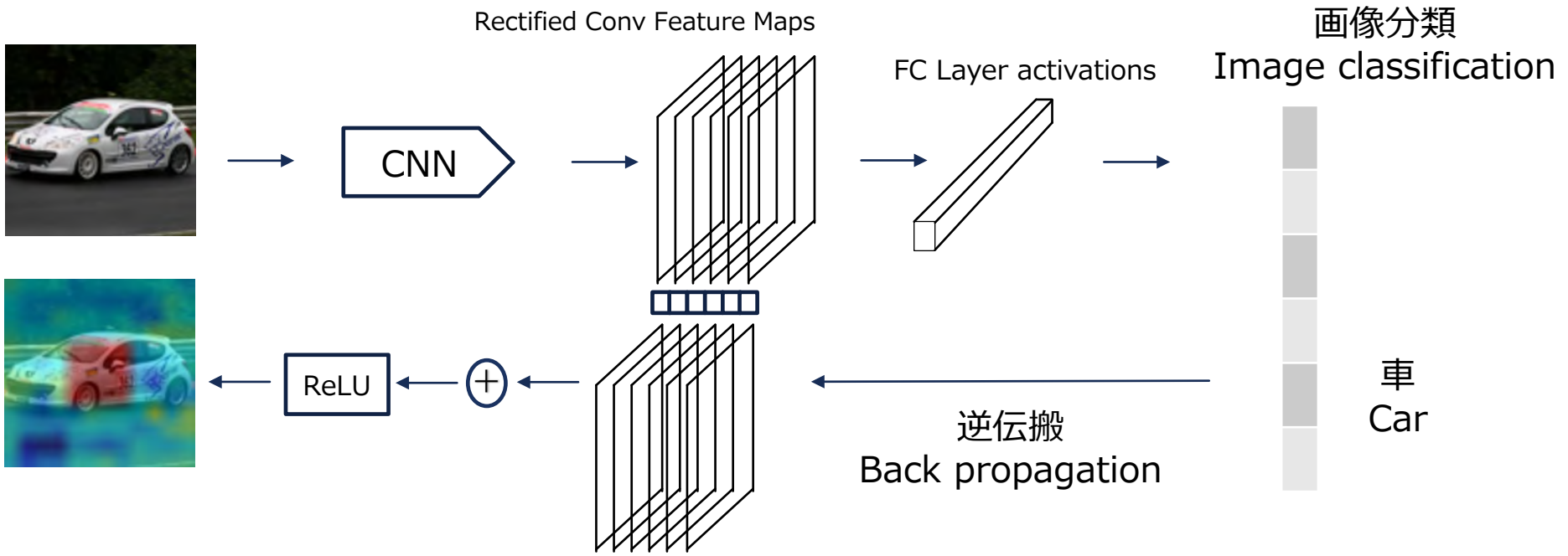


代表的な説明可能なAI

XAI技術	特徴	画像	表データ	テキスト
Grad-CAM	CNNでの畳み込み層の勾配を利用し，画像中の判断根拠となる箇所をヒートマップで表示する技術である.	○	×	×
LIME	モデルを局所的に線形モデルで近似することによって判断根拠を示す方法である.	○	○	○
SHAP	ゲーム理論のShapley 値を求める手法を使って各特徴量の寄与度を計算する手法である.	○	○	○
Smooth Grad	入力画像にガウシアンノイズを載せ，複数回の勾配計算をした後に平均を取ることによって，判断根拠となる箇所の可視化画像を 生成する手法である.	○	×	×

Grad-CAM

Deep learning はどこをみているのだろうか？



STL-10 dataset : Adam Coates, Honglak Lee, Andrew Y. Ng An Analysis of Single Layer Networks in Unsupervised Feature Learning AISTATS, 2011.

CNNでの畳み込み層の勾配を利用し，画像中の判断根拠となる箇所をヒートマップで表示する技術である．

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International conference on computer vision, pages 618 - 626, 2017.

LIME

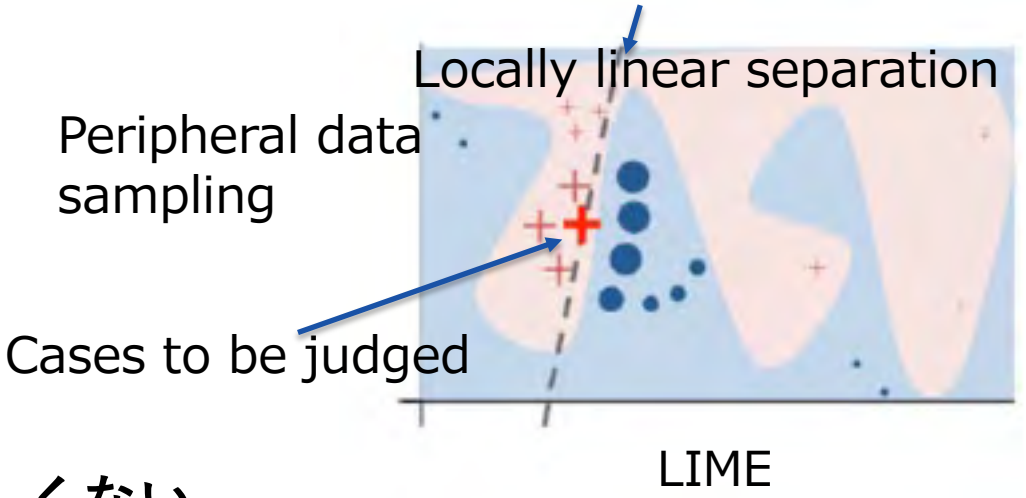
LIME (Local Interpretable Model-agnostic Explanation)

局所的な説明の代表的な方法

LIME

ハスキー犬が狼と判断

背景の雪が判断根拠

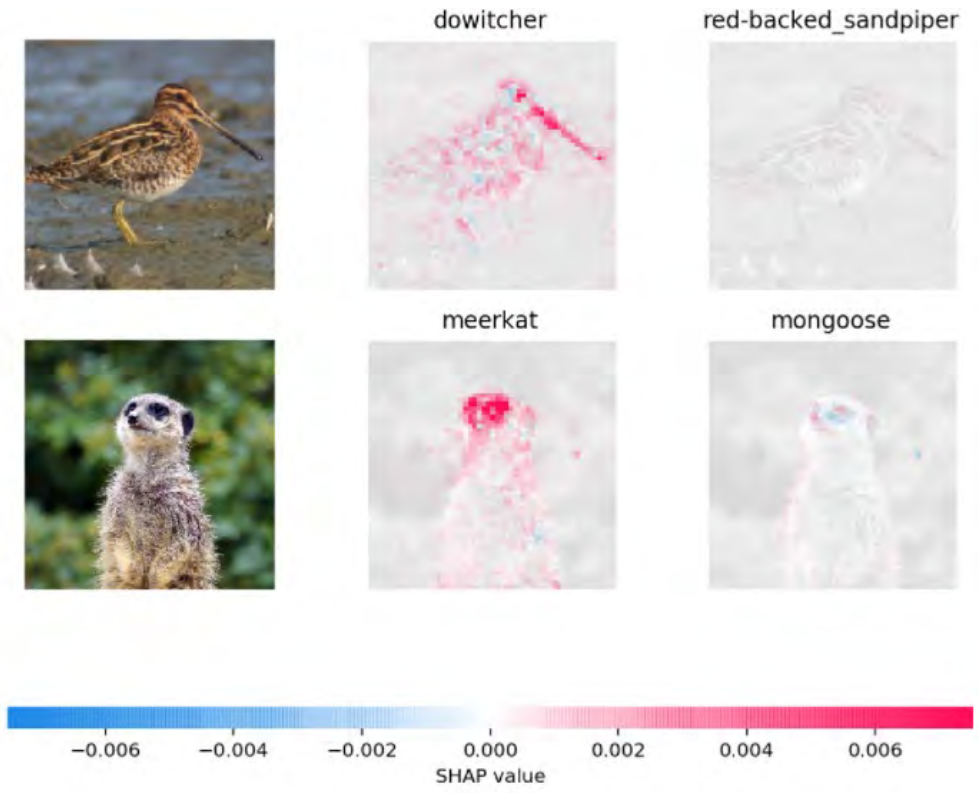


このモデルは, 判断根拠が正しくない

Marco Tulio Riberio, Sameer Singh, and Carlos Guestrin. "Why shoud I trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International conference on knowledge discovery and data mining, pages 1135 - 1144, 2016.

LIMEは、モデルを局所的に線形モデルで近似することによって判断根拠を示す方法である.

SHAP



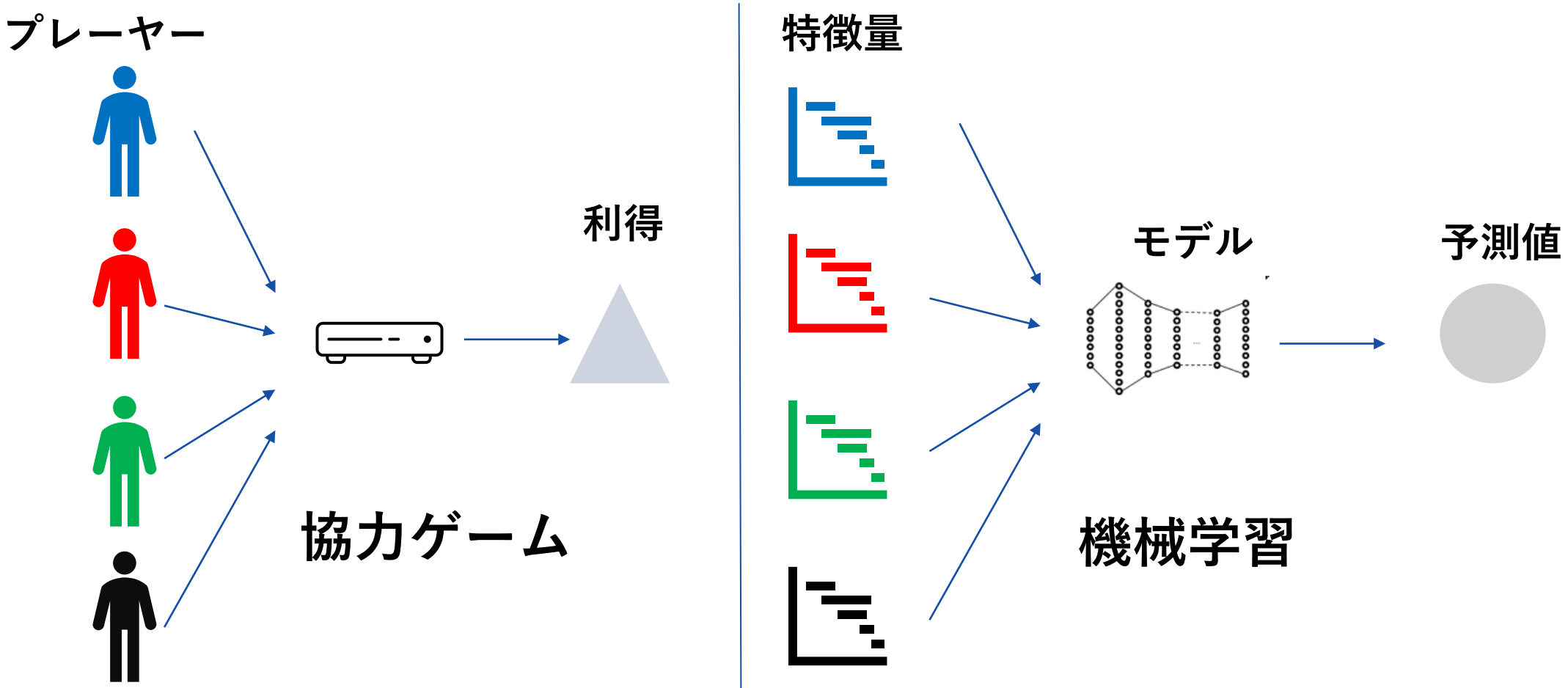
赤：　プラスに寄与

青：　マイナスに寄与

Scott M Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In Advances in Neural Information Processing Systems 30, pages 4768–4. 2017.

SHAPは, ゲーム理論のShapley 値を求める手法を使って
各特徴量の寄与度を計算する手法

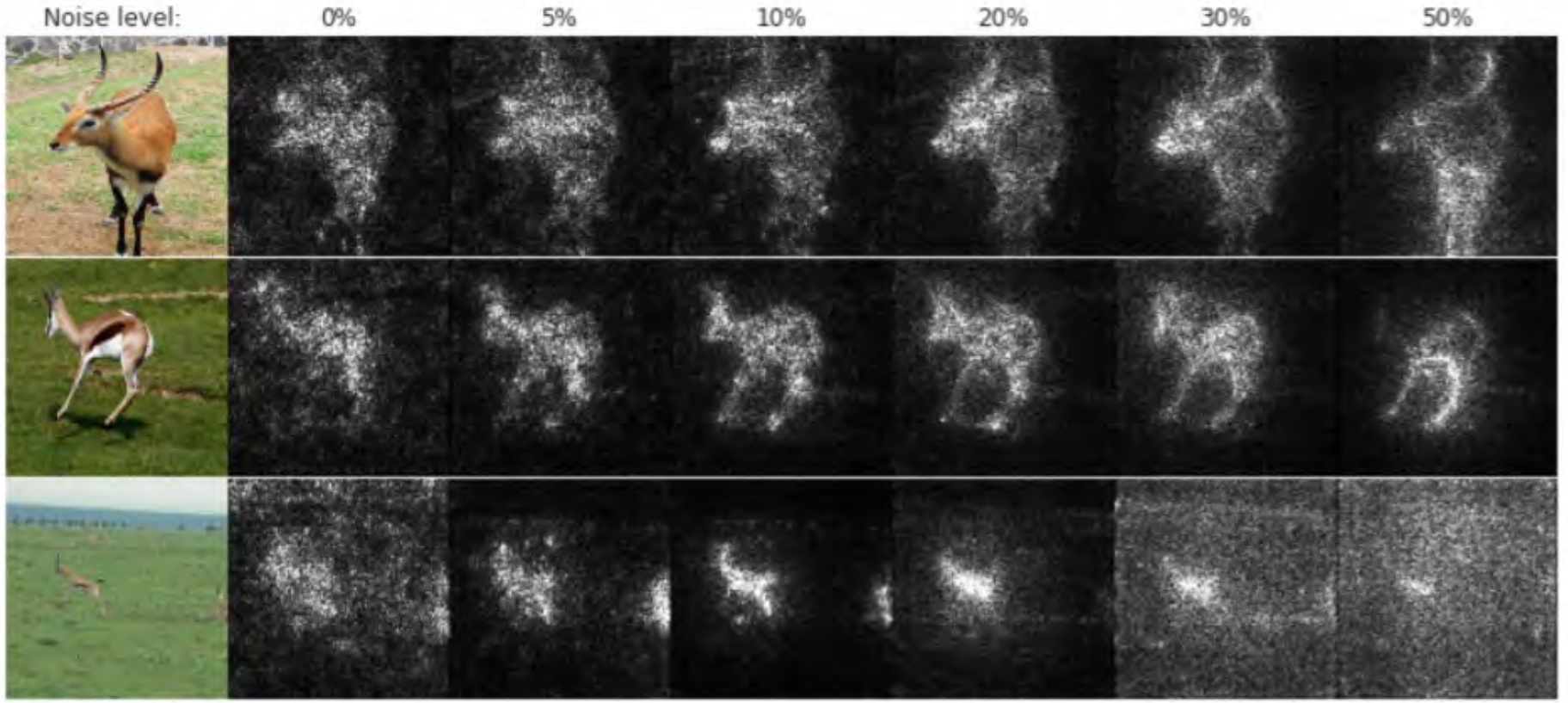
協力ゲームと機械学習



協力ゲーム理論のShapley値は, プレーヤー間での 報酬を公平に分配する方法.

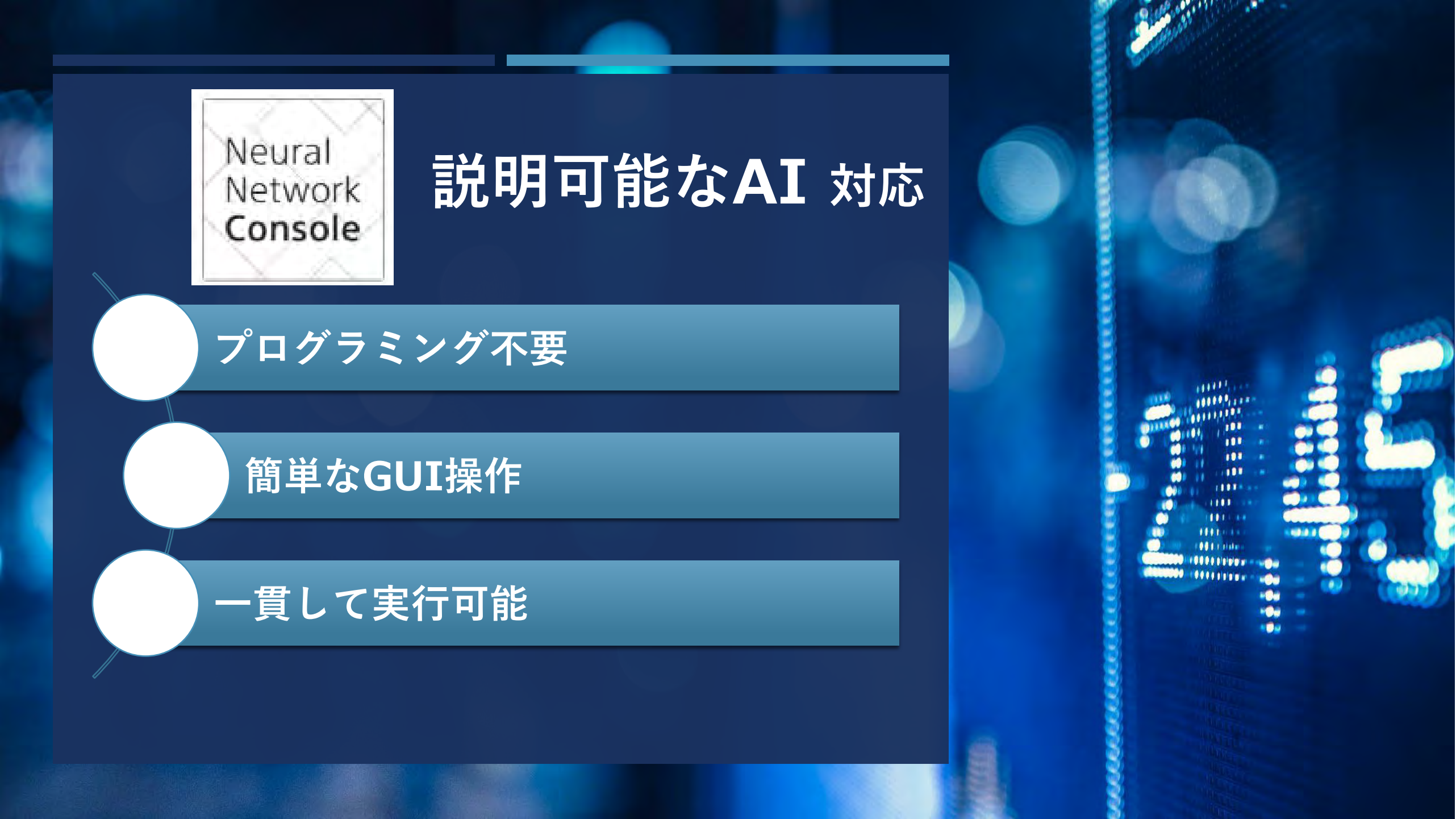
この考え方を機械学習に適用し, 各特徴量の寄与度を算出する.

Smooth Grad



Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viegas, and Martin Wattenberg. SmoothGrad: removing noise by adding noise. arXiv:1706.03825, 2017.

SmoothGradは，入力画像にガウシアンノイズを載せ，複数回の勾配計算をした後に平均を取ることによって，判断根拠となる箇所の可視化画像を 生成する手法．

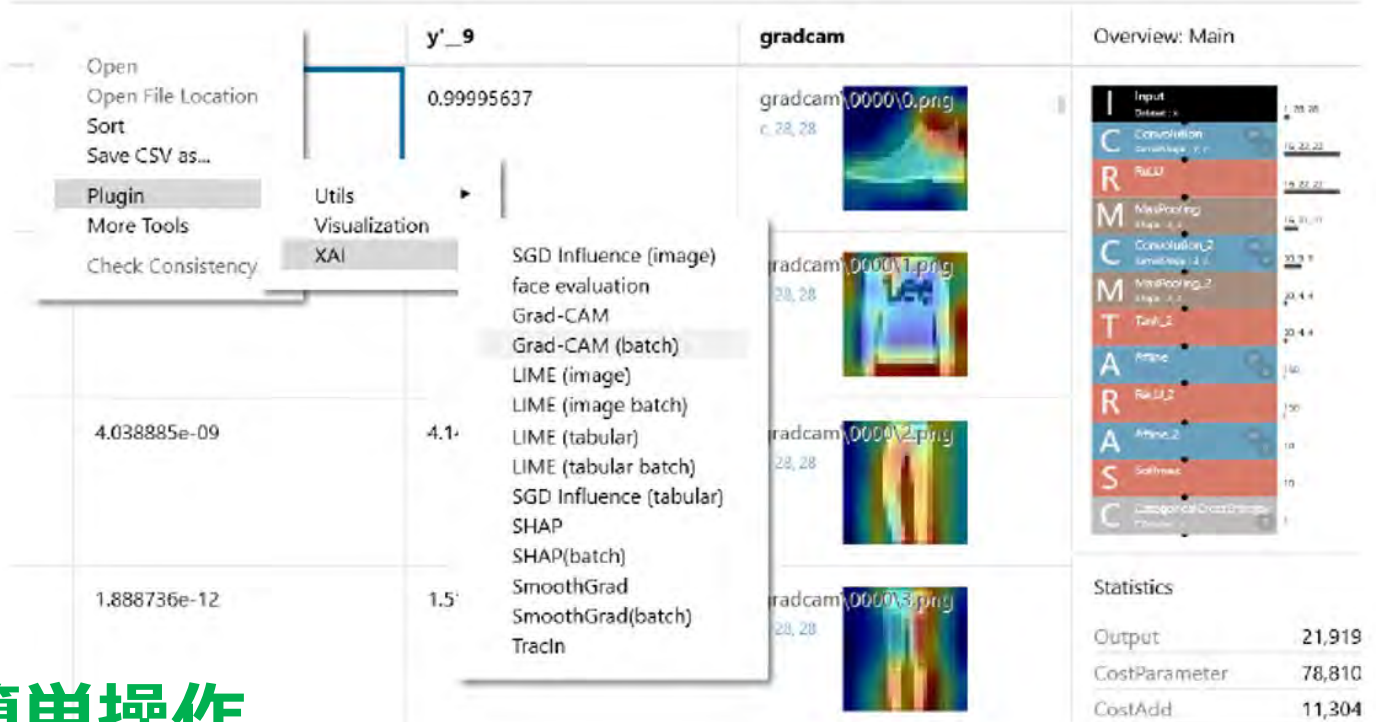


説明可能なAI 対応

- プログラミング不要
- 簡単なGUI操作
- 一貫して実行可能

ディープラーニング・説明可能なAIの統合開発フレームワーク

Neural Network Console



GUIによる簡単操作

Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. Han Xiao, Kashif Rasul, Roland Vollgraf. arXiv:1708.07747

https://dl.sony.com/ja/

モデルの構築、学習、説明可能なAIまで,
Neural Network Consoleにて一貫して開発できる.

画像分類における判断の根拠の可視化

(a) Original　(b) Grad-CAM　(c) LIME　(d) SHAP　(e) SmoothGrad

靴の上部にて判断していることが分かる

(a) Ankle boot

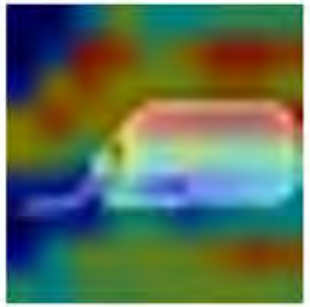
違いは、靴の上部

(b) Sandal

データセット内での似たような画像

それぞれの手法による判断根拠の可視化は，必ずしも一致しない。
このように説明手法が異なれば，得られる説明も異なる.

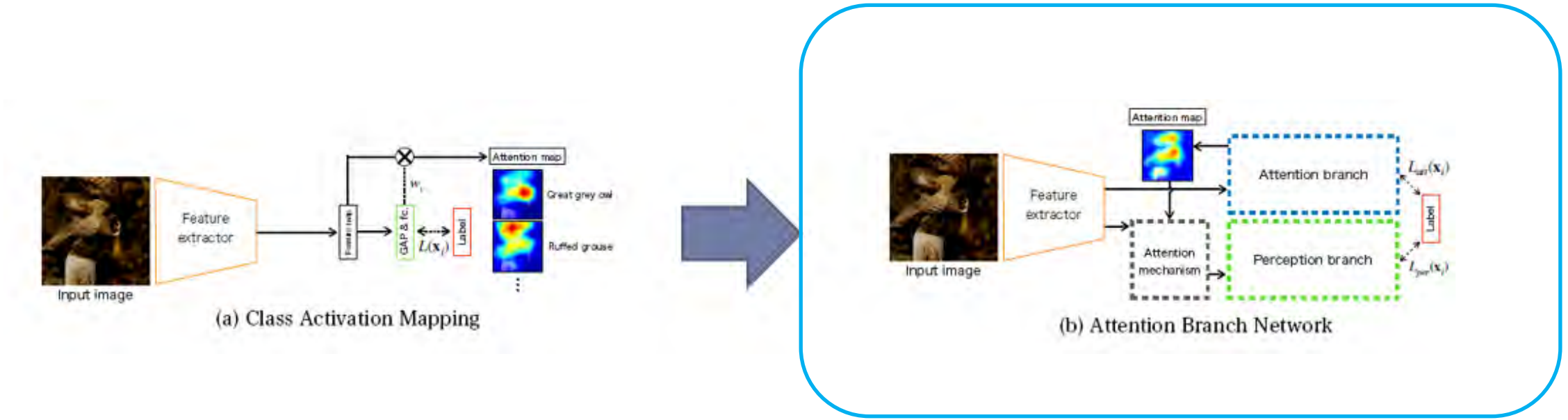
画像分類における誤判断の根拠の可視化

真のラベル	判断	(a) Original	(b)Grad-CAM	(c) LIME	(d) SHAP	(e)Smooth grad
Ankle boot 	Pullover 					
Bag 	Ankle boot 					

AIの判断根拠を活用する技術

判断根拠の可視化だけでなく、精度の向上

Attention Branch Network



H. Fukui, T. Hirakawa, T. Yamashita, and H. Fujiyoshi. Attention branch network: Learning of attention mechanism for visual explanation. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2019.

- 判断根拠を用いてCNNの性能を向上させる初めての試み
- より注目すべき場所を示すことで効率よく学習しモデルの精度を向上
- ベースのモデルにAttention Branchを導入するため様々なモデルに適用可能

機械学習モデルへのデータの影響度

- データクレンジング
- ミスラベル検知

判断根拠

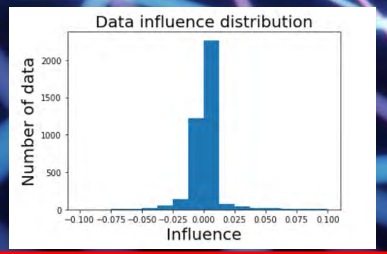
Basis for judge

- Grad-CAM
- LIME
- SHAP
- Smooth Grad
- Attention Branch Network

データの影響度

Data influence

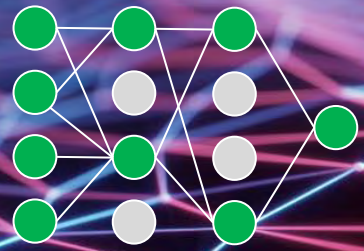
- SGD influence
- Influence functions
- TracIn



データの不確実性

Data uncertainty

- MC dropout



データクレンジング

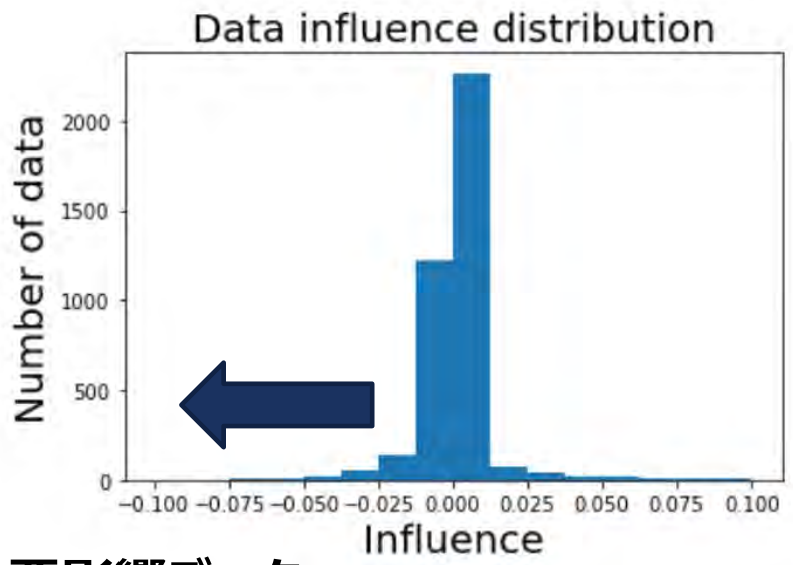
Satoshi Hara, Atsushi Nitanda, and Takanori Maehara. Data cleansing for models trained with SGD. In *Advances in Neural Information Processing Systems*, pages 4215-4224, 2019

DNNへ悪影響を及ぼすデータを削除

データを追加することなく精度向上

データの影響度 Data influence　SGD influence

データ影響度分布



悪影響データ
Negative influence data

悪影響データの除去
Remove bad influence data

DNNの精度向上
Accuracy up of DNN

SGD influenceは，悪影響データを抽出することができる.

データの質をどのように向上させるか

Deep learning requires a large amount of data and labeling

Current issues

- **Require a large amount of data**
- **Uniformly correct of high-quality data**
- **Need to label data**
- **Require domain knowledge for labeling**

Goal

Automatically clean data of DNN

Method

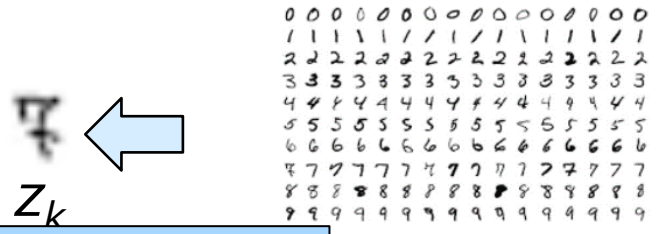
$$\left\langle u, \theta_{-k}^{[T]} - \theta^{[T]} \right\rangle$$

Results

Use data cleansing on Neural Network Console

DNNへのデータ影響度の計算

Which data should be removed to improve accuracy ?
How can we know the influence of a piece of data on DNN ?



z_k

Identify negatively influential data

Remove negatively influential data

Retrain

Accuracy up

Removed Data	Loss function
Positively influential	increase ⇧
Negatively influential	decrease ⇩

θ :model parameter
D, z : data
l : loss function

Small loss ≈ Linear influence

$$\text{influence } \tilde{r}_k = \left\langle u, \theta_{-k}^{[T]} - \theta^{[T]} \right\rangle$$

$$\hat{\theta}_{-k} = \arg\min_{\theta} \sum_{z \in D \setminus \{z_k\}} l(z;\theta) \quad \hat{\theta} = \arg\min_{\theta} \sum_{z \in D} l(z;\theta)$$

Issue

It takes an enormous amount of time to remove the data one by one and retrain.

Approach

Estimate $\hat{\theta}_{-k} - \hat{\theta}$ without retraining

データ影響度計算のアルゴリズム

Algorithm 1 Training phase

Initialize parameter θ [1]

Initialize sequence as null : A $\leftarrow 0$

for $t = 1, 2, ..., T-1$ **do**

$\theta^{[t+1]} \leftarrow \theta^{[t]} - \frac{\eta_t}{|S_t|} \sum_{i \in S_t} g(z_i; \theta^{[t]})$

end for

A $\leftarrow \theta$ // store parameters

Algorithm 2 Inference phase

Require: $u^{[T-1]} \in \mathbb{R}^p$

Initialize influence: $\hat{L}^{[T]}_{-j}(u) \leftarrow 0, \forall j$

$\theta \leftarrow$ A // load parameters

for $t = T-1, T-2, ..., 1$ **do**

// update linear influence of z_j

$\hat{L}^{[T]}_{-j}(u^{[t]}) += \langle u^{[t]}, \frac{\eta_t}{|S_t|} g(z_j; \theta) \rangle, \forall j \in S_t$

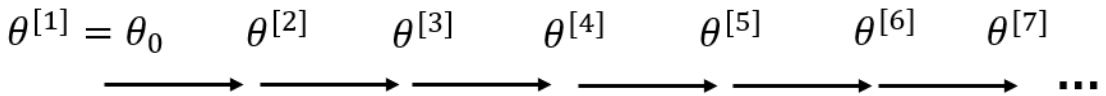
$u^{[t-1]} = u^{[t]} - \eta_t H^{[t]} u^{[t]}$ // update u

end for

ストレージに対して効果的なデータ影響度計算手法

従来手法　[Hara+NeurIPS2019]

全てのイテレーションのパラメータを利用



最終エポック

我々の提案手法

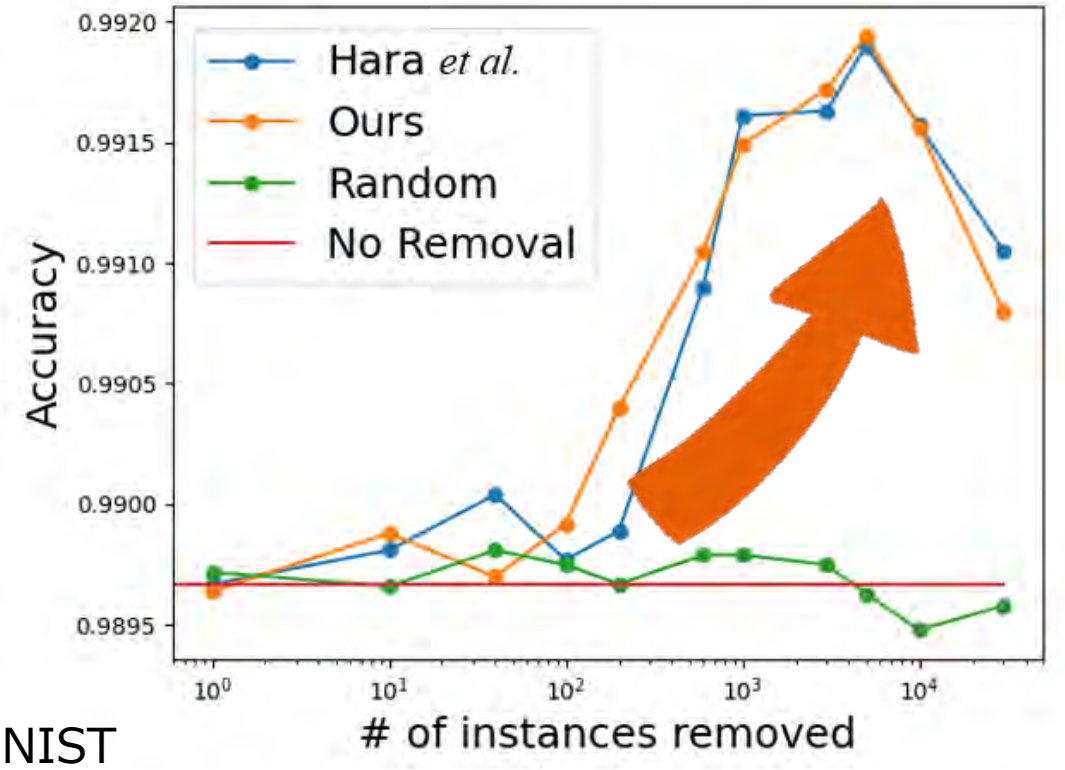
最後のパラメータのみを利用

θ

キャッシュサイズを1/1,563へ削減

Table 1 : Cache size of the parameters in training

Methods	Cache	Cache size (GB)	
		MNIST	CIFAR10
All parameters	$\theta \times T \times k$	38.64	30.90
Hara *et al.*	$\theta \times T$	1.932	1.545
Ours	θ	0.001236	0.001236

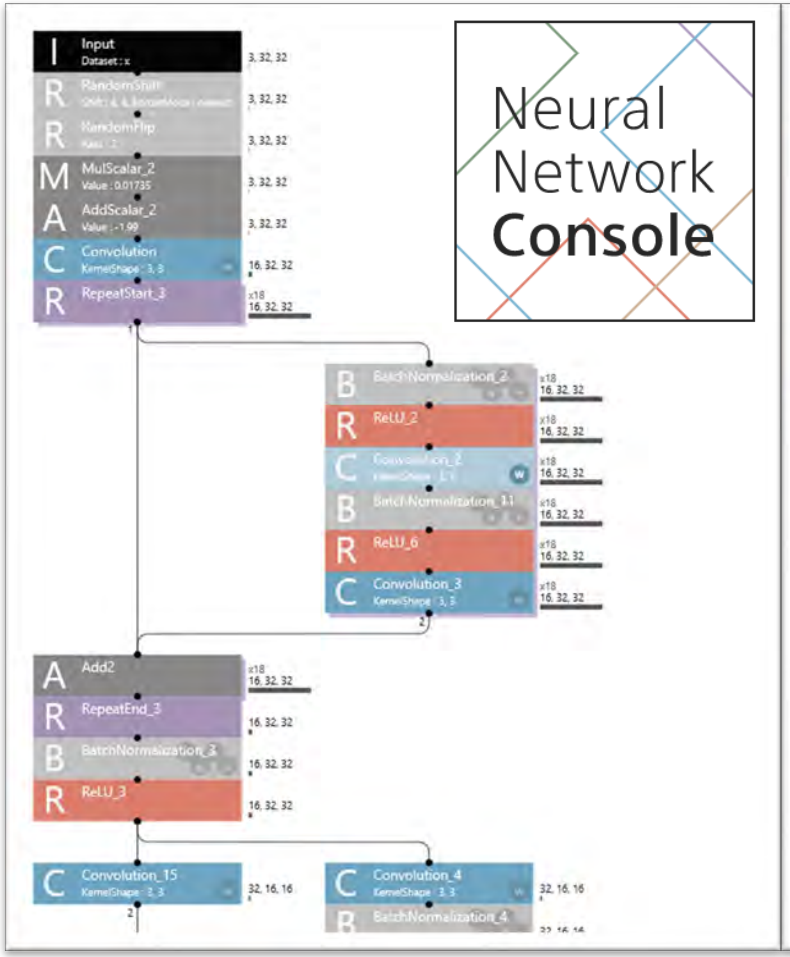


データセット　MNIST

従来手法と同等のデータクレンジング性能

Neural Network Consoleへ実装

https://dl.sony.com/ja/



Neural Network **Console**

x:image	y:label	influence
... training\8\39330.png c, 32, 32	8	-0.012459796509163769
... training\0\32878.png c, 32, 32	0	-0.006328233351282367
... CIFAR10\training\0\6370.png c, 32, 32	0	-0.0063282314659545745

データクレンジングについての論文

Computer Science > Machine Learning

[Submitted on 22 Mar 2021 (v1), last revised 1 Jun 2021 (this version, v2)]

Data Cleansing for Deep Neural Networks with Storage-efficient Approximation of Influence Functions

Kenji Suzuki, Yoshiyuki Kobayashi, Takuya Narihira

Identifying the influence of training data for data cleansing can improve the accuracy of deep learning. An approach with stochastic gradient descent (SGD) called SGD-influence to calculate the influence scores was proposed, but, the calculation costs are expensive. It is necessary to temporally store the parameters of the model during training phase for inference phase to calculate influence sores. In close connection with the previous method, we propose a method to reduce cache files to store the parameters in training phase for calculating inference score. We only adopt the final parameters in last epoch for influence functions calculation. In our experiments on classification, the cache size of training using MNIST dataset with our approach is 1.236 MB. On the other hand, the previous method used cache size of 1.932 GB in last epoch. It means that cache size has been reduced to 1/1,563. We also observed the accuracy improvement by data cleansing with removal of negatively influential data using our approach as well as the previous method. Moreover, our simple and general proposed method to calculate influence scores is available on our auto ML tool without programing, Neural Network Console. The source code is also available.

Subjects: **Machine Learning (cs.LG)**; Artificial Intelligence (cs.AI)
Cite as: arXiv:2103.11807 **[cs.LG]**
(or arXiv:2103.11807v2 **[cs.LG]** for this version)

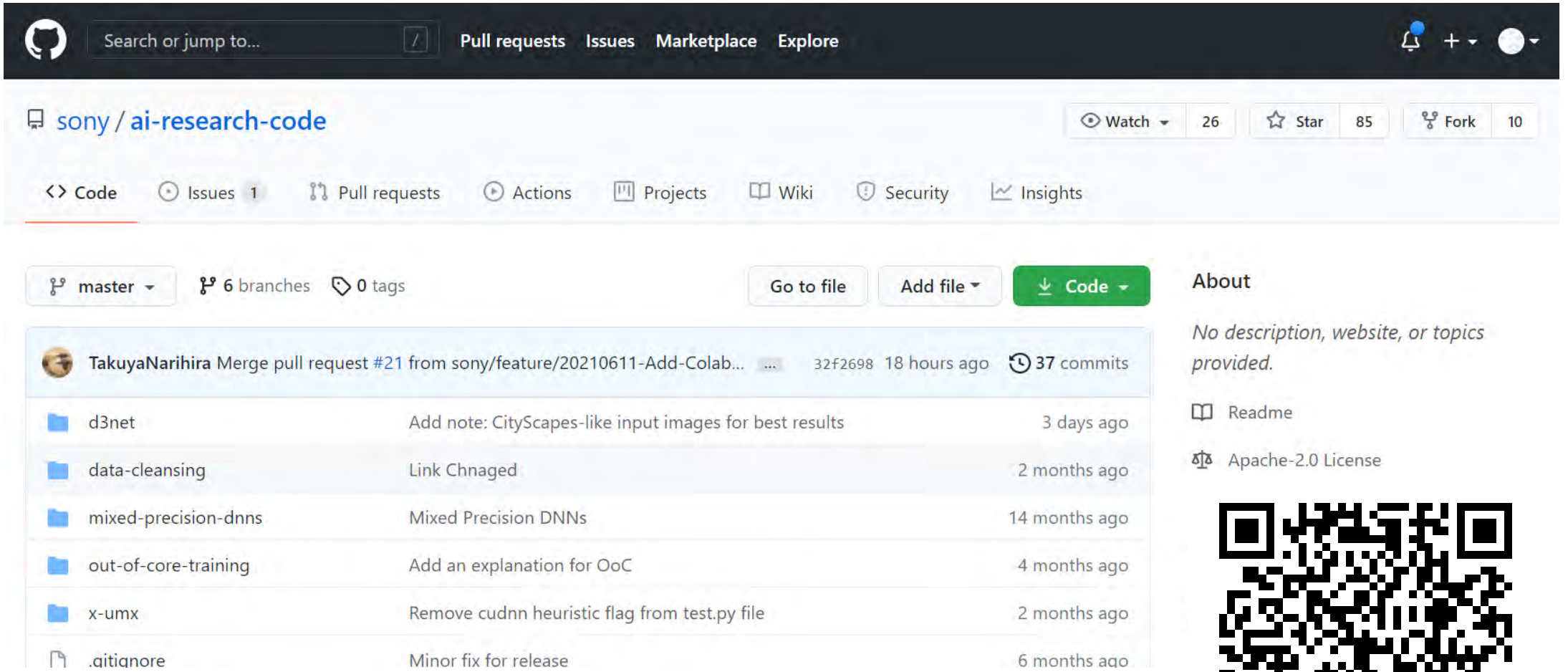
Submission history

From: Kenji Suzuki [view email]
[v1] Mon, 22 Mar 2021 13:08:46 UTC (921 KB)
[v2] Tue, 1 Jun 2021 08:23:25 UTC (960 KB)

https://arxiv.org/abs/2103.11807v2

詳しくは, 論文をご覧ください.

GitHubへコード公開



Neural Network **Libraries**

データクレンジング　まとめ

✓提案
- ストレージに対して効果的なデータ影響度計算手法を提案

✓結果
- キャッシュサイズを1/1,563へ削減(MNISTでの実験)
- 従来手法と同等のデータクレンジング効果あり

✓実装
- Neural Network Consoleへ実装
- GitHubへコードを公開

ミスラベル検知

課題

- **実社会のデータの教師あり学習では, ミスラベルが起きやすい.**

解決策

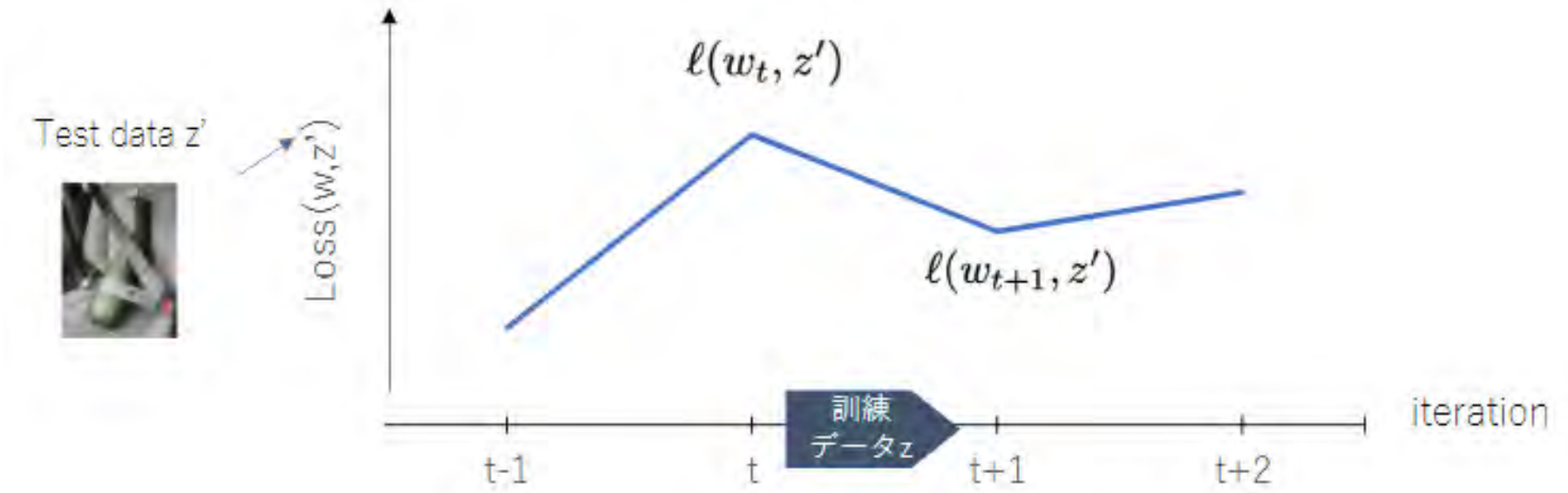
- **訓練データのミスラベルを検知する.**

手法

- **パラメータ更新による損失の変化を追跡し, 訓練データの影響を定式化.**

Garima Pruthi, Frederick Liu, Mukund Sundararajan, Satyen Kale
Estimating Training Data Influence by Tracing Gradient Descent, NeurIPS2020.

訓練データの影響の定式化(TracIn)

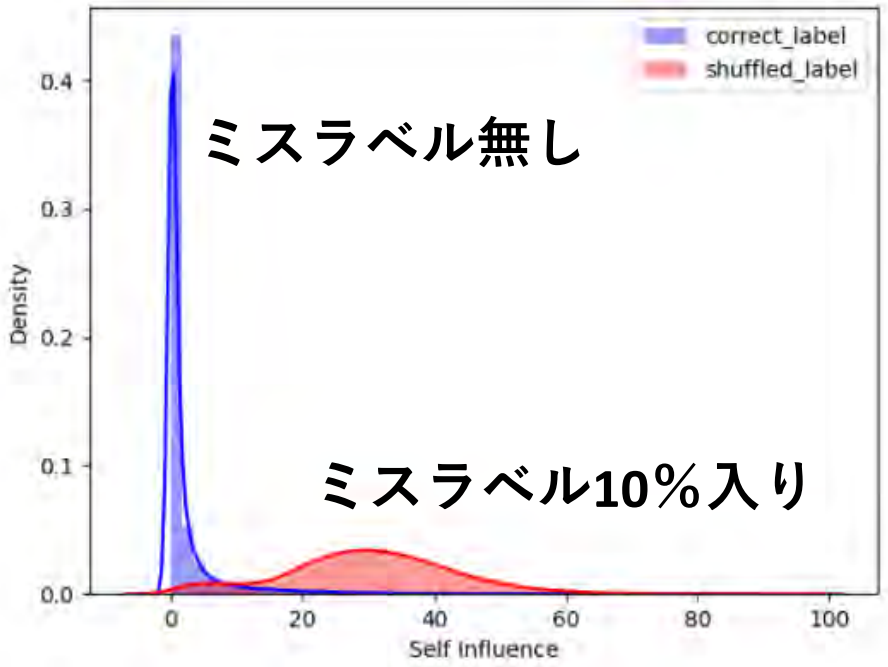


訓練過程でのTest dataのLossの変化をみる

Lossが減少 → 訓練に寄与したデータは良い影響を持っている

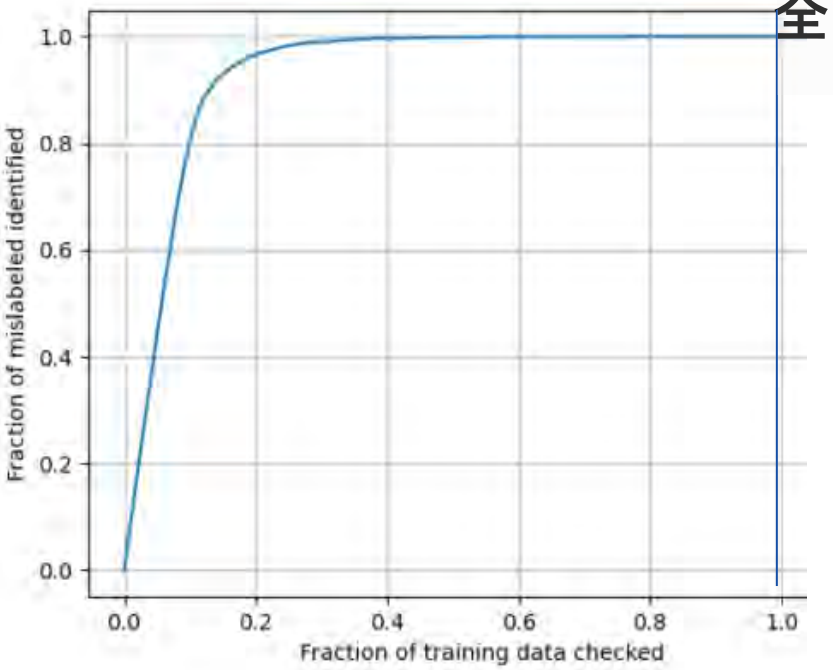
ミスラベル検出の結果

影響度の分布



ミスラベルの検出

全データチェック



https://github.com/sony/nnabla-examples/tree/master/responsible_ai/tracin

データの不確実性

判断根拠

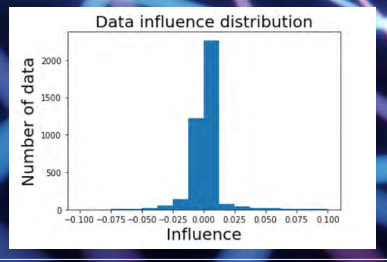
Basis for judge

- Grad-CAM
- LIME
- SHAP
- Smooth Grad
- Attention Branch Network

データの影響度

Data influence

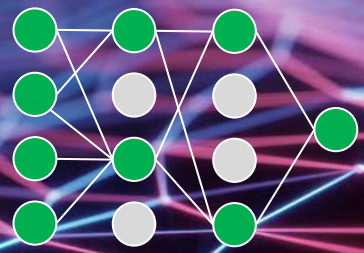
- SGD influence
- Influence functions
- TracIn



データの不確実性

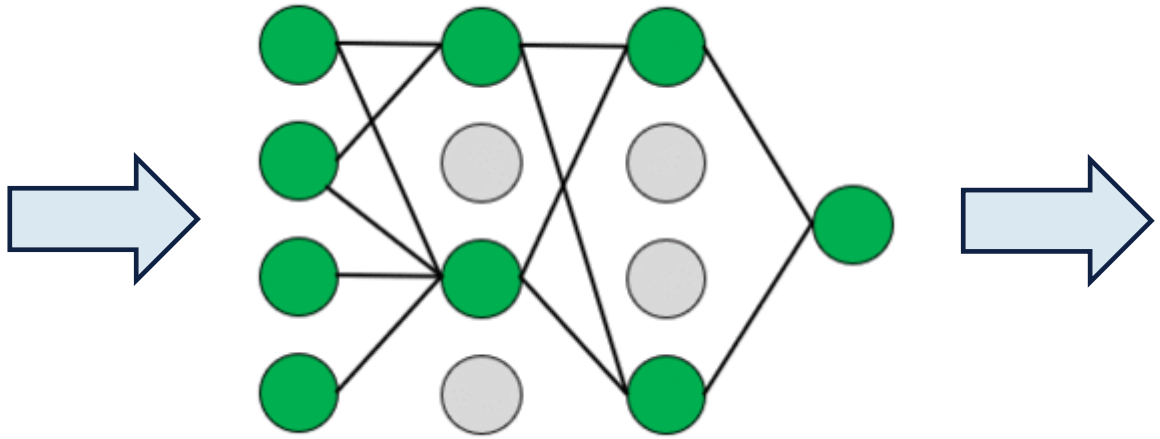
Data uncertainty

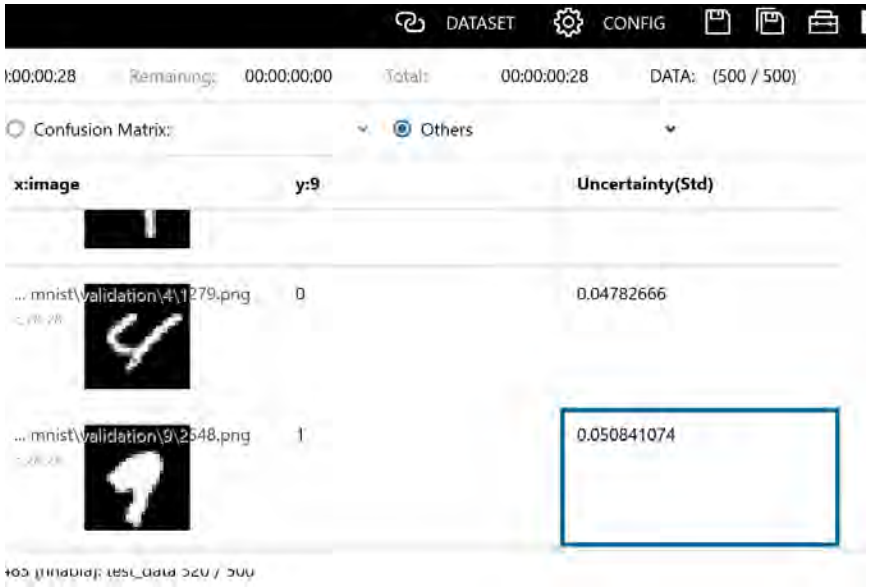
- MC dropout



データの不確実性 Data uncertainty

MC dropout





近似ベイズ推論 Approximate Bayesian inference

MC dropoutは、データの不確実性を示すことができる.

データの不確実性

近似ベイズ推論による不確実性を伴った予測

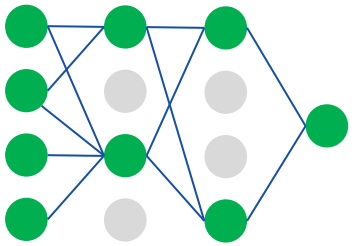
Deep learning

$$y = f(w, x)$$

パラメーターwは定数. 入力xによりyが決まる.

Bayesian Deep Learning

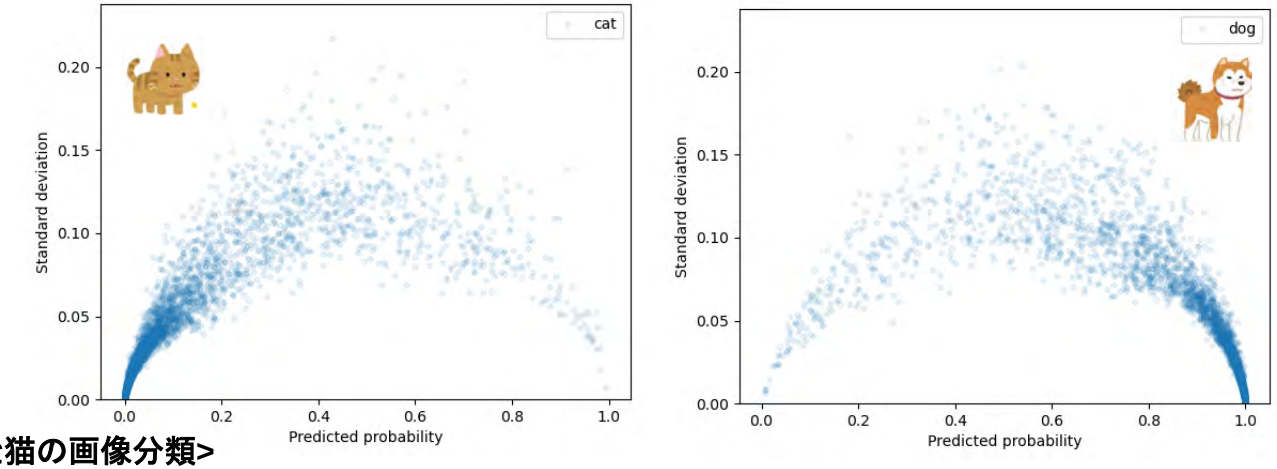
$$q(\mathbf{y}^*|\mathbf{x}^*) = \int p(\mathbf{y}^*|\mathbf{x}^*, \omega) q(\omega) \mathrm{d}\omega$$



確定な点推定パラメーターではなく,全ての推定確率分布を利用

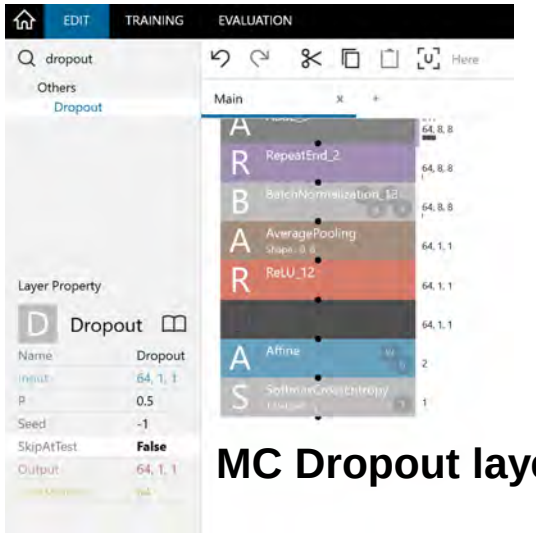
推定値と不確実性(標準偏差)の関係

Relationship between estimates and uncertainties (standard deviation)



<犬と猫の画像分類>
<Dog and cat image classification>

不確実性は、推定値が0や1では無く、0.5に近いほど大きい

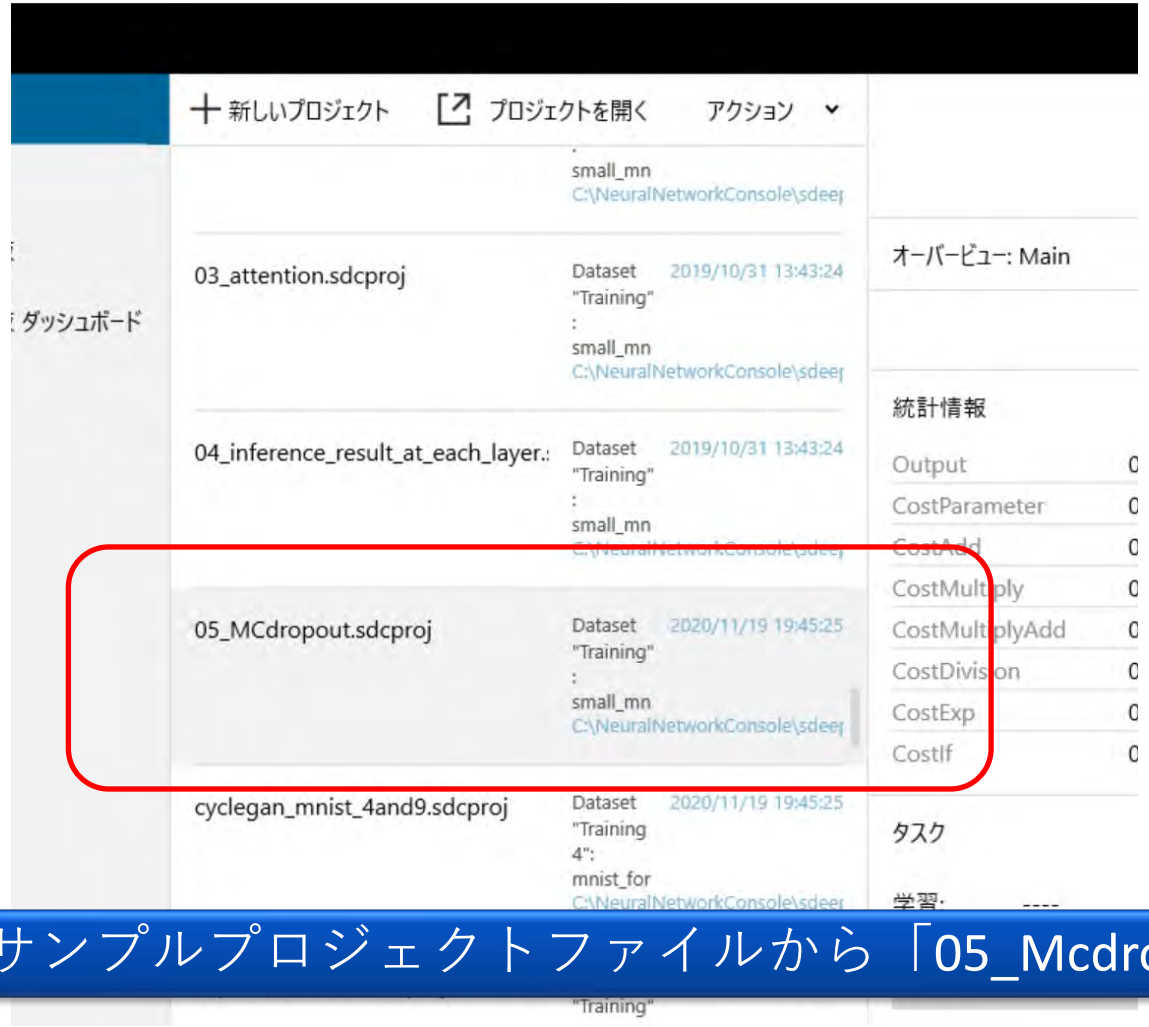


MC Dropout layer

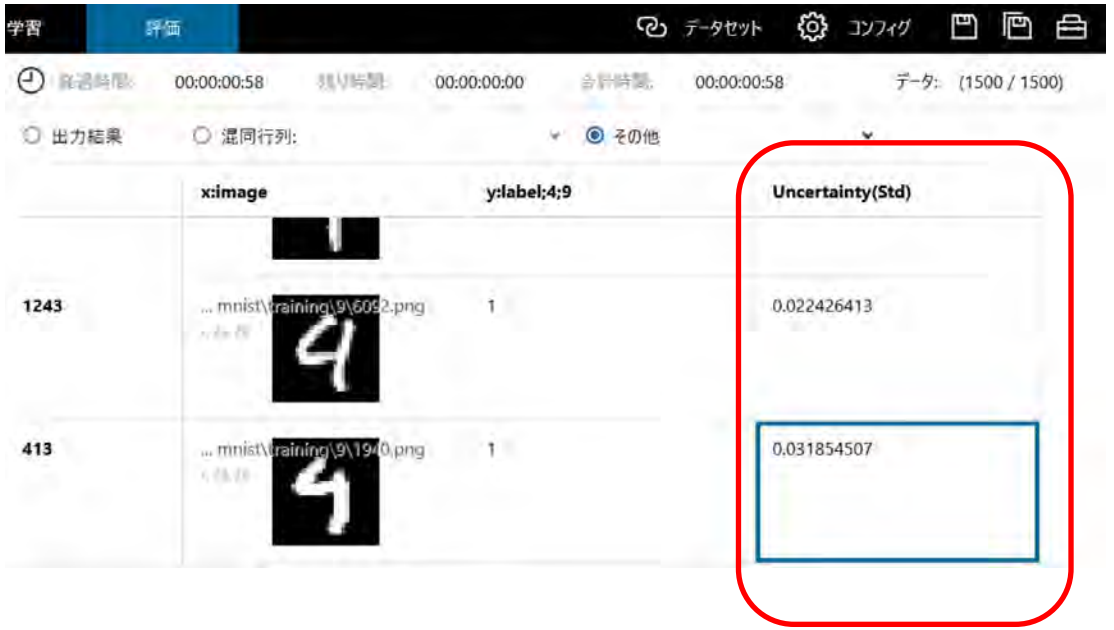
Neural Network Console

Neural Network Consoleへの実装

Neural Network Consoleでのデータの不確実性の使い方



- 学習と評価を実行



各データの不確実性(Uncertainty)が表示される

サンプルプロジェクトファイルから「05_Mcdropout.sdcproj」を選択

近年の動向や課題

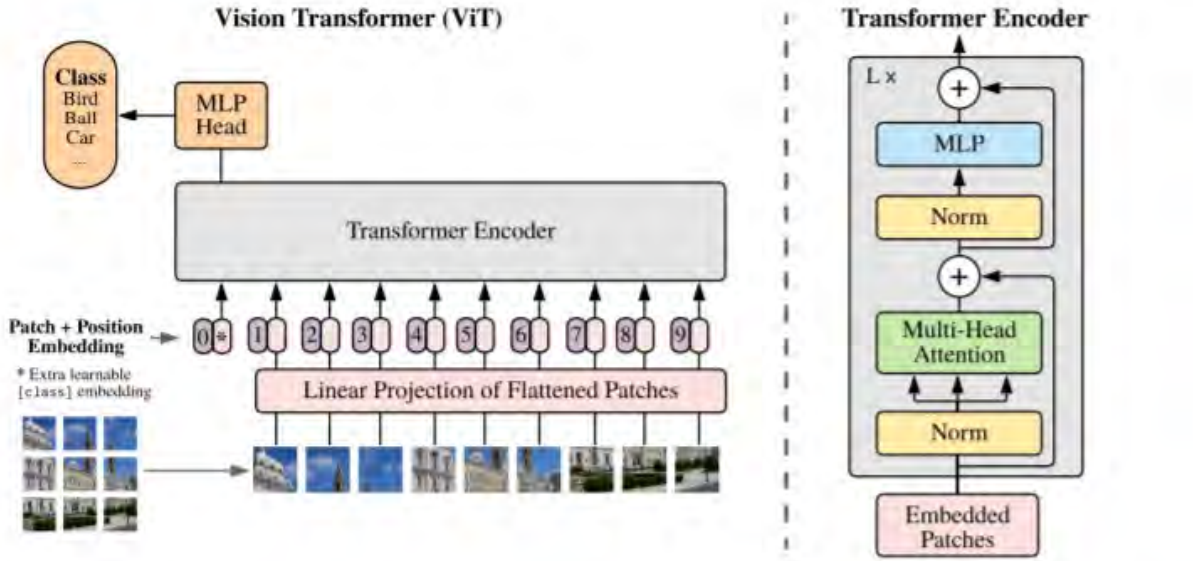
Visual Transformer (ViT)の出現

Visual Transformerは，ImageNetの画像認識精度を90%台へ

出典　https://paperswithcode.com/sota/image-classification-on-imagenet

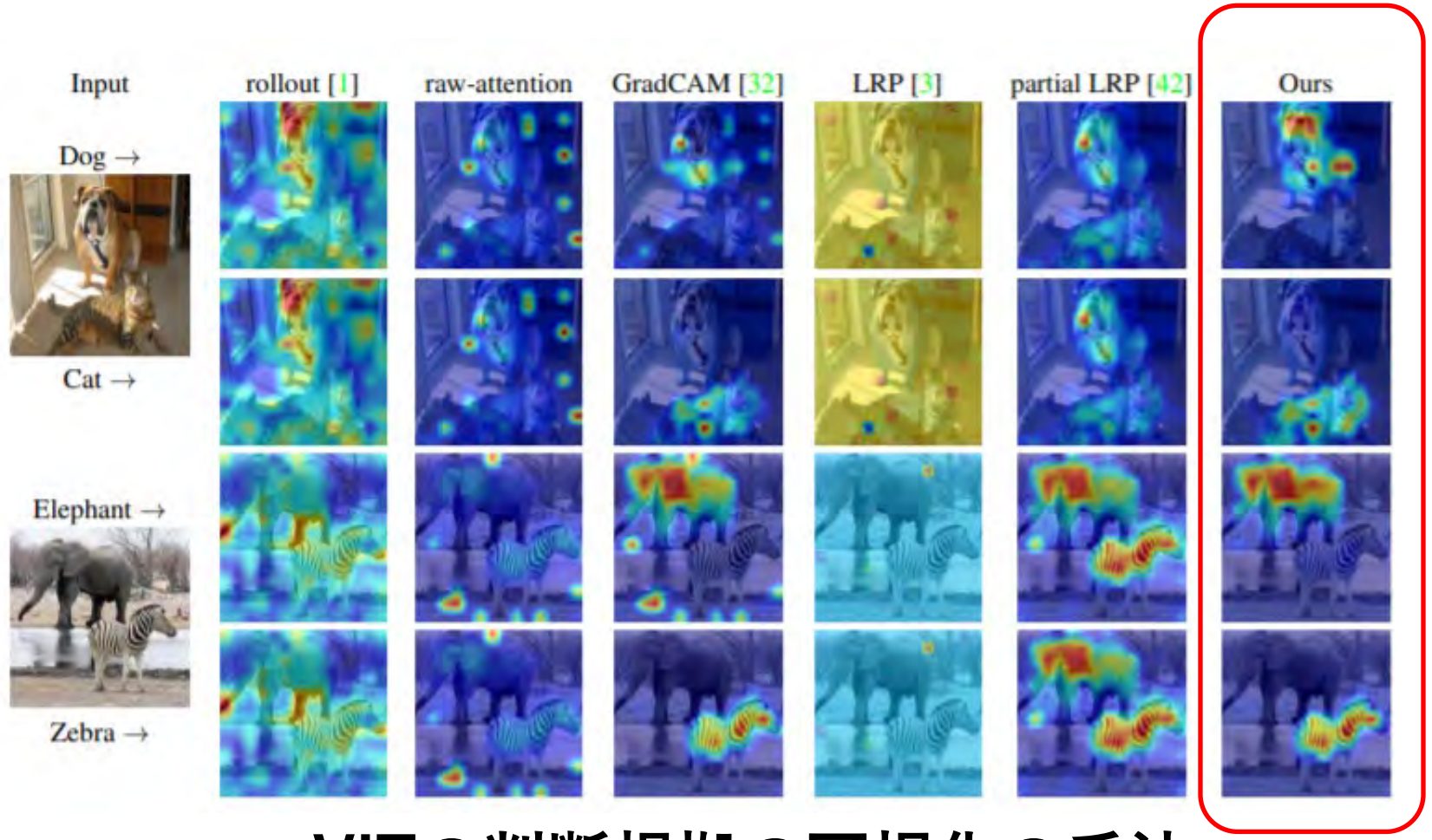
Vision Transformer(ViT)

- Transformerは，大域的な空間情報を捉えることができる.
- 画像パッチを単語のように取り扱う.
- 巨大なデータセット(300M)で事前学習をする.



Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby
An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, ICLR 2021

Vision Transformer (ViT) Explainability



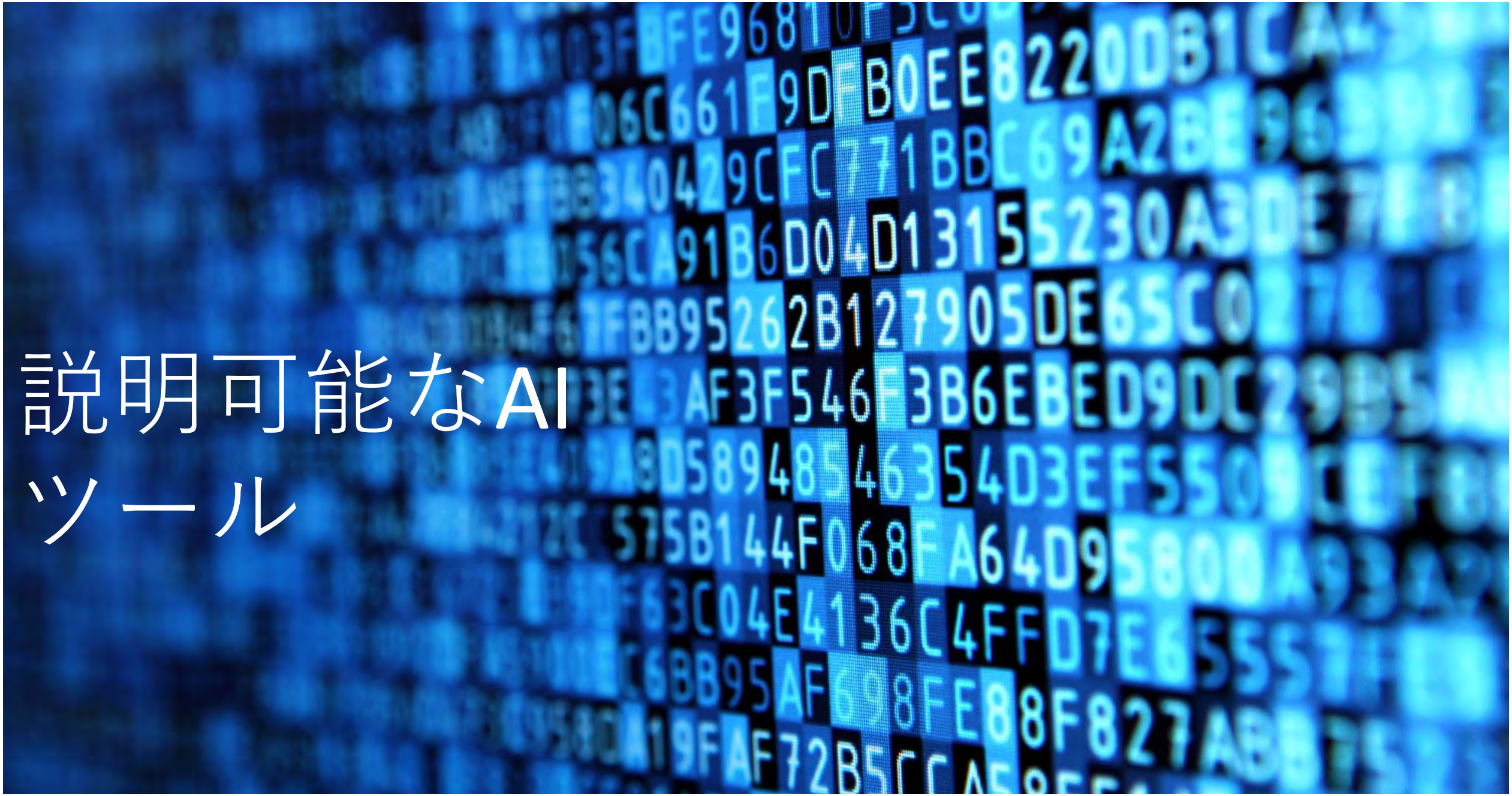
ViTの判断根拠の可視化の手法

Hila Chefer, Shir Gur, Lior Wolf, Transformer Interpretability Beyond Attention Visualization, CVPR 2021

説明可能なAIの課題

- ✓ 人間が理解できても, 納得がいく説明になっているか.
- ✓ 様々な説明手法により結果が異なり，人間が理解できない.
- ✓ 様々なタスクに対して，どの説明可能なAIの手法を使うのが妥当か.
- ✓ 乱数パラメーターを利用する手法での再現性に注意.
- ✓ 計算コストが大きいケースもある.

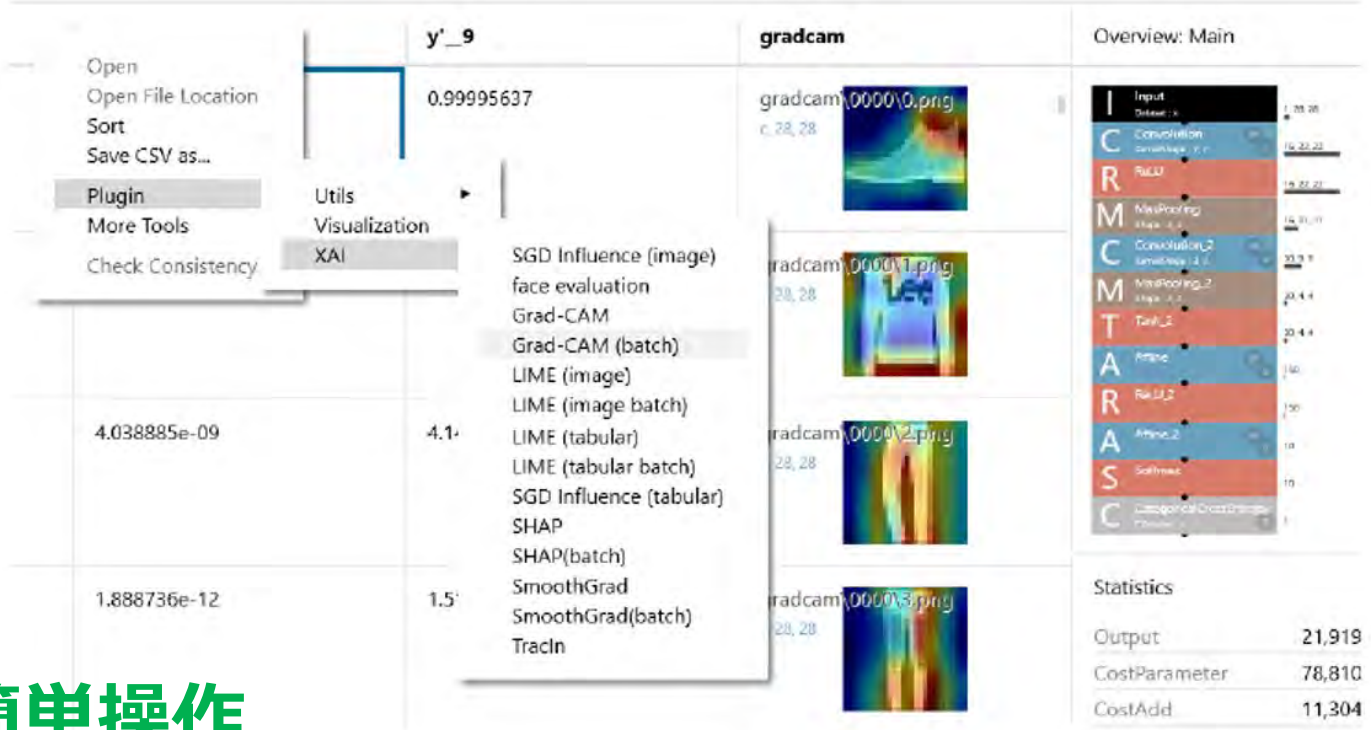
説明可能なAIへの過度な期待は禁物



説明可能なAIツール

Neural Network Console による説明可能なAI

Neural Network Console



GUIによる簡単操作

Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. Han Xiao, Kashif Rasul, Roland Vollgraf. arXiv:1708.07747

https://dl.sony.com/ja/

モデルの構築、学習、説明可能なAIまで,
Neural Network Consoleにて一貫して開発できる.

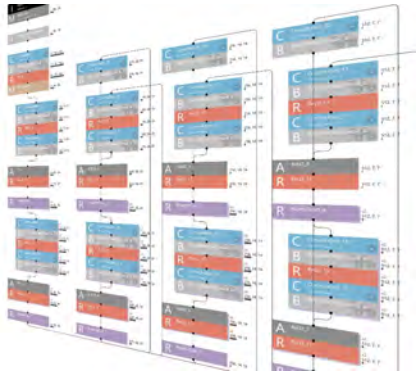
説明可能なAIのリリース形態

説明可能なAI

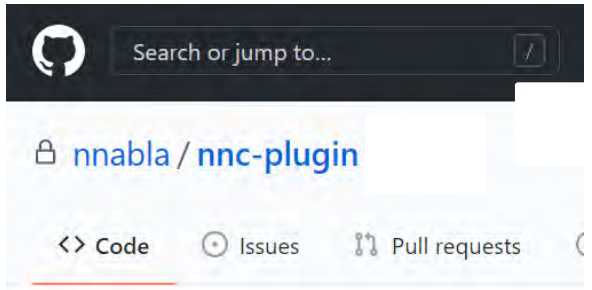
Neural Network Libraries

GUI

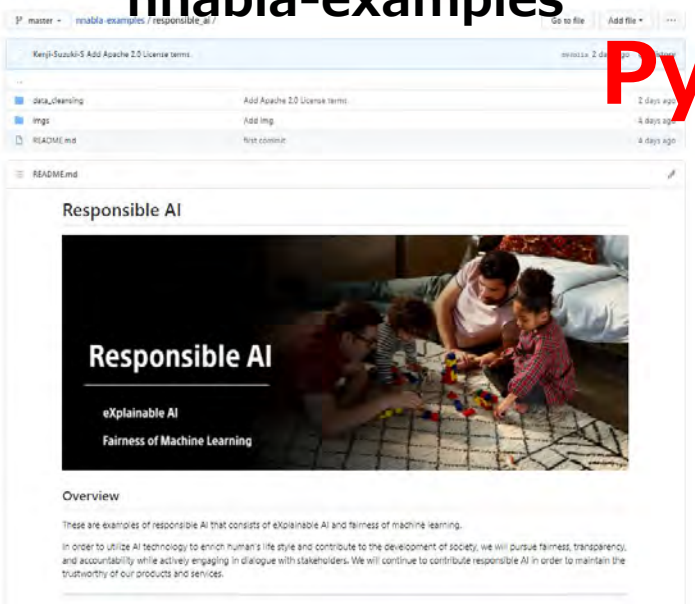
Neural Network Console



オープンソースソフトウェア



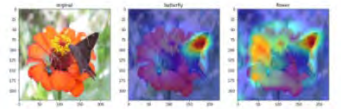
nnabla-examples

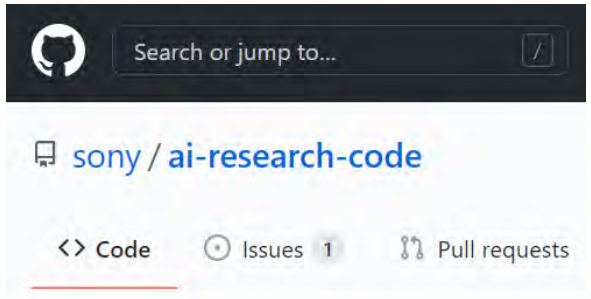


Python

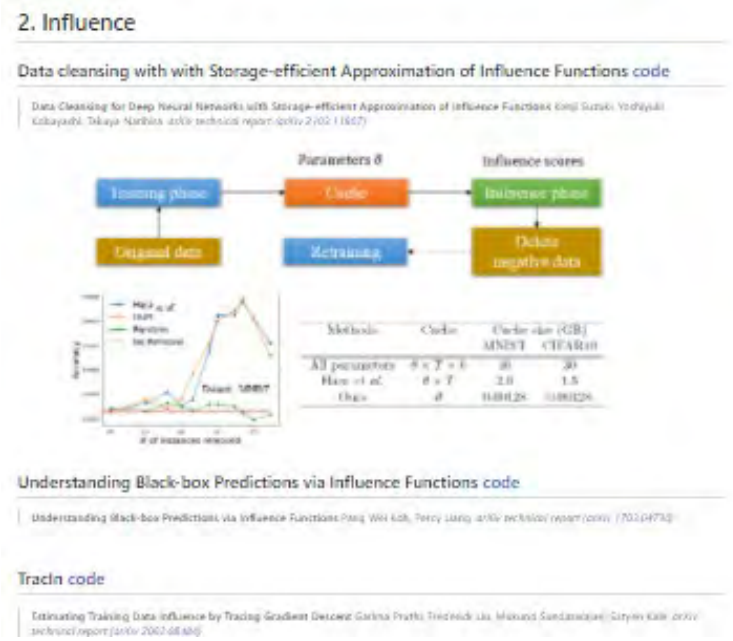
Colab

eXplainable AI

Name	Notebook	Task	Example
Grad-CAM	Open in Colab	Grad-CAM	



Responsible AI ライブラリ

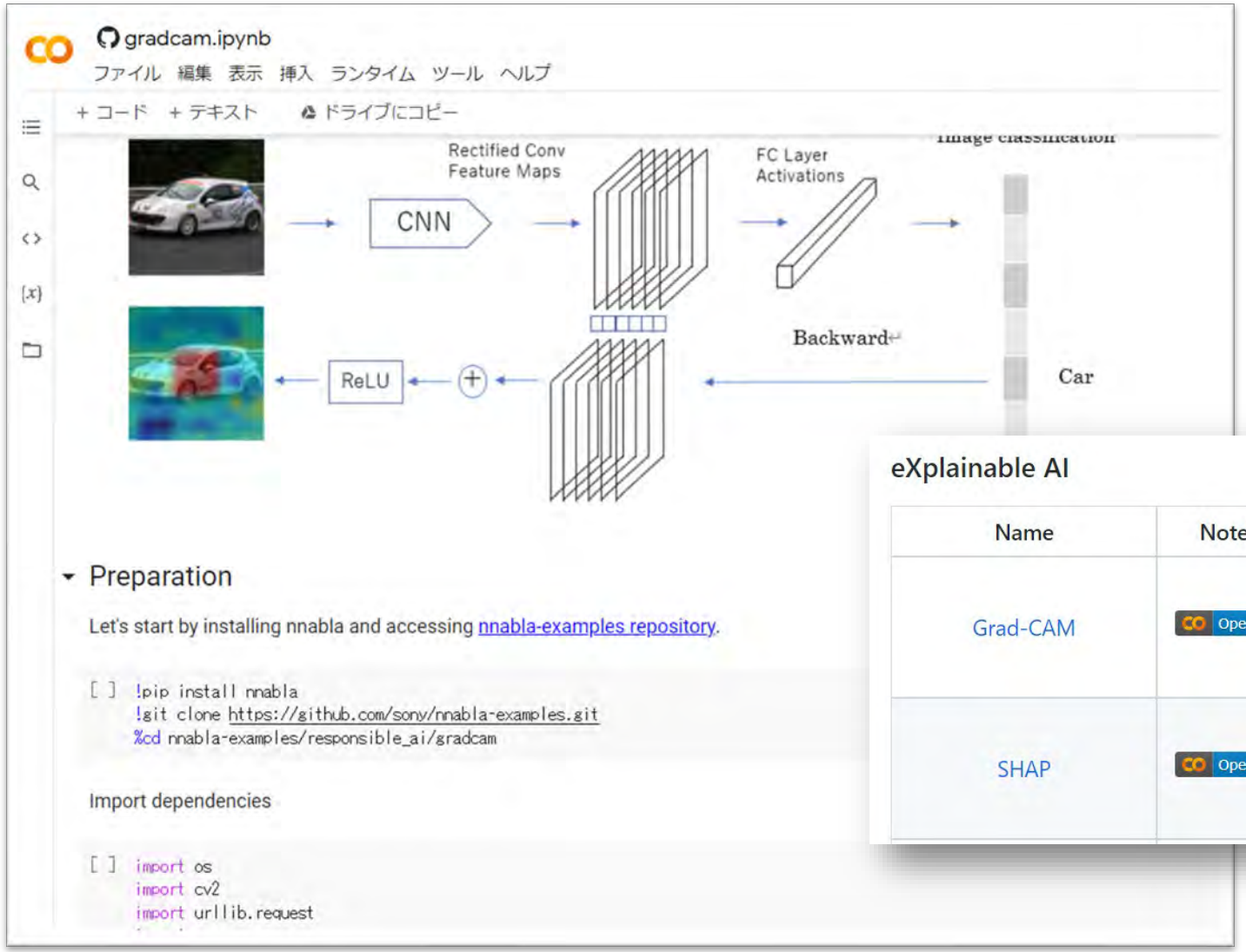


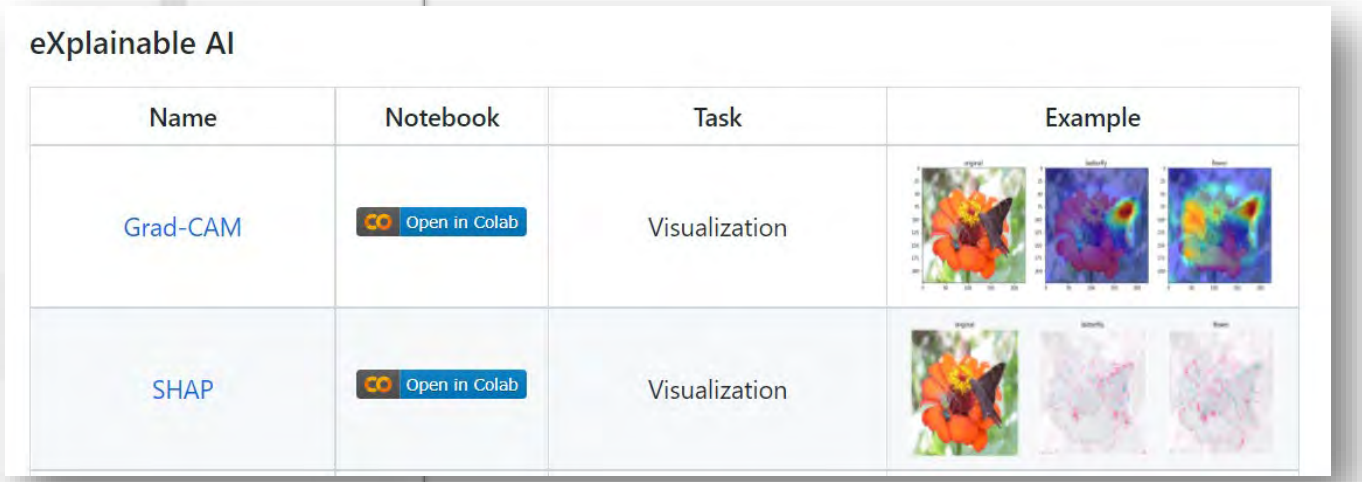
Neural Network Libraries

https://github.com/sony/nnabla-examples/tree/master/responsible_ai

可視化、データ影響度、Colab形式のサンプルが充実

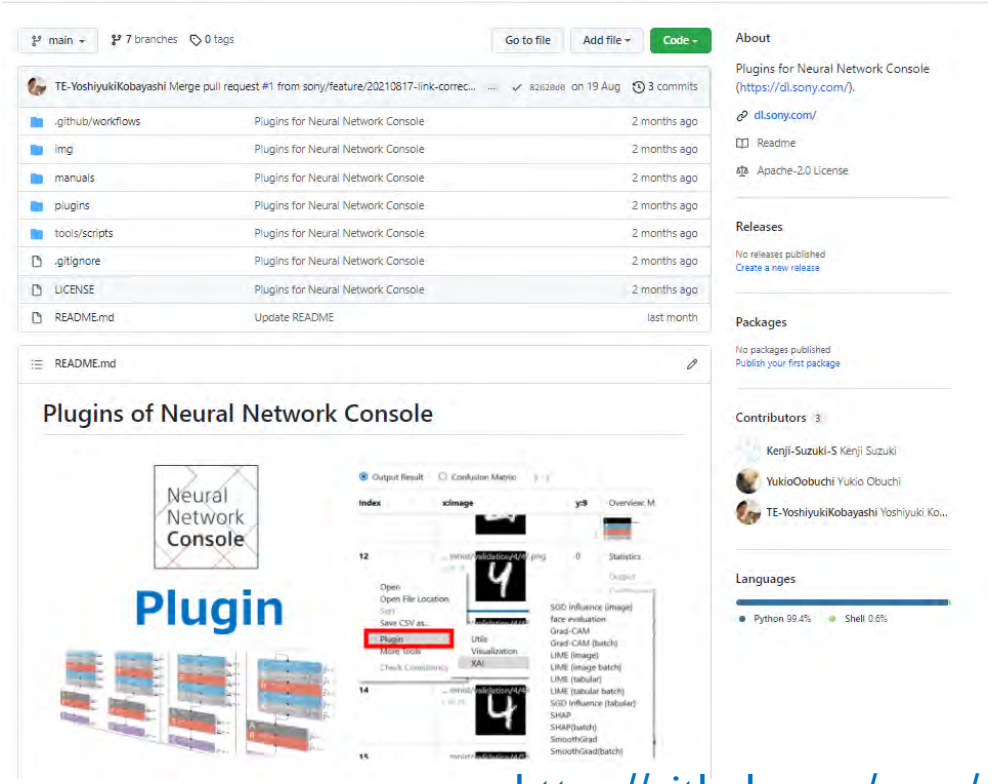
分かりやすいノートブック形式　Colabチュートリアル

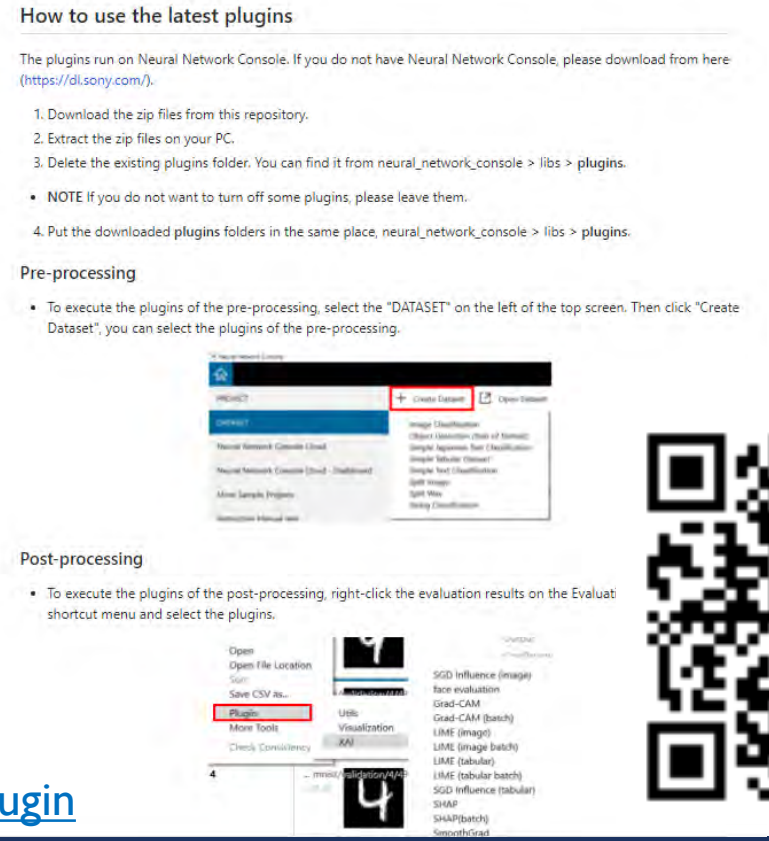




https://github.com/sony/nnabla-examples

Neural Network Console 最新Pluginコード OSS公開





https://github.com/sony/nnc-plugin

最新版のプラグインをOSS公開サイトからダウンロードして
Neural Network Consoleで使える.

YouTube 動画公開

New!!

- Neural Network Consoleのチュートリアル
- ディープラーニング技術の入門的解説
- 約25,000人のチャンネル登録者

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

- Neural Network Librariesのチュートリアル
- ディープラーニング技術の先端論文・技術解説
- 2021年3月に開設後、約2,500人のチャンネル登録

https://www.youtube.com/channel/UCOELxR-yS2EbjBxQ0hx4yBw

まとめ　「説明可能なAI」

説明可能なAIとは, 人間がAIの判断理由を理解することができるようにする技術である.

社会的な要請により, 説明可能なAIは, AI利活用に必要な技術である.

AI倫理のためだけではなく, AIの判断理由を明らかにすることにより, 潜在的なAIの能力を引き出すことができる.

Neural Network Consoleにより, 専門的な知識を必要とせず, 説明可能なAIを手軽に利用することができる.

SONY