

THE NEOPULSE® FRAMEWORK

概要



AI Dynamics Inc.
formerly DimensionalMechanics Inc.

2018 156th Ave NE, Building F, Suite 119
Bellevue, WA 98007



日本国内販売代理店
neopulse@nippon-rad.co.jp

イントロダクション

NeoPulse® Frameworkは、あらゆるビジネスにおいてAIを利用可能なものとし、機械学習を活用する際に直面する課題の大部分を解決する強力な技術の集積体です。具体的には、バージョン管理、セキュリティ、スケーリング、ロギング、課題の発見難度といった問題に配慮し、AI展開の制御・管理を可能にする機能をフルに備えたAIマネジメントシステム(AIMS)です。

また、NeoPulse® FrameworkはAWS・Azureのクラウドサービスからオンプレミス、さらにはARM64をベースとしたエッジデバイスに至るまで、幅広いエンドポイントへ柔軟に展開できます。

中でもIoTとエッジコンピューティングは、NeoPulse® Frameworkが対象デバイスへ推論モデルを容易に展開でき、低パワー・低メモリ環境でも推論モデルを実行できるようモデルに制約をつけられるため、特に注力している分野です。

一つのプラットフォーム上でデータベースを管理するのと変わらないほど簡単に、AIソリューションの生成・展開・管理が可能となります。

目標

NeoPulse® Frameworkには3つの主な目標があります。

AIの作成を容易にすることにより、あらゆるビジネスがAIを活用できるようにする

AIを、どこでも、どのようなデバイスにも展開できるようにする

企業がAIを完全にコントロールできるようにする

NeoPulse® Frameworkはユニークなプロデューサ-コンシューマアーキテクチャの採用により、推論モデルの生成プロセスと展開・管理プロセスを分離し、これらの目標を実現します。

NeoPulse® Frameworkの構成要素

NeoPulse® Frameworkは、主に4つの構成要素から成り立っています。



NeoPulse® AI Studio: AIを構築するAI

洗練されたAIモデルの生成プロセスを自動化する、強力なAI「the oracle」を有するサーバアプリケーションです。



ポータブル推論モデル (Portable Inference Models: PIM)

AI Studioにより作成されたニューラルネットワークモデルを格納した小さなファイルで、NeoPulse® Query Runtimeと合わせて使用することでAIの活用が可能となります。PIMは一つのAIモデルとみなすことができます。



NeoPulse® Query Runtime

PIMを利用するためのランタイムライブラリで、様々なデバイスやプラットフォームに対応しており、ビジネスにおけるあらゆるアプリケーションへの組み込みによってRESTful APIやコマンドラインインターフェースを経由するAIモデルへのアクセスを可能にします。



NeoPulse® モデリング言語 (NeoPulse® Modeling Language: NML)

Dimensional Mechanics Inc.が開発した直感的に利用できる専用言語です。NeoPulse® AI StudioがAIモデル生成を自動化する際の様々な設定を記述します。

NeoPulse® AI Studio (NAIS) はサーバアプリケーションであり、「Oracle」と呼ばれる強力なAIが搭載されています。「Oracle」は700,000–900,000超のアルゴリズムを有する巨大なデータベースと、さまざまなエンドポイント用のカスタム推論モデルを生成、最適化およびパッケージ化可能な学習・最適化エンジンに支えられています。NeoPulse® AI Studioは対象エンドポイント上で最適化実行されるためのオブジェクトであるポータブル推論モデル (Portable Inference Model “PIM”) を出力し、これにはメタデータ、実行アルゴリズム、推論モデル等の関連データが含まれます。PIMは対象デバイスまたは対象環境でNeoPulse® Query Runtime (NPQR) さえ動作していれば利用できます。NPQRは、プログラムがPIMに対してクエリ実行を行えるようにRESTインターフェースを自動生成します。

NeoPulse®モデリング言語 (NeoPulse® Modeling Language “NML”) はチューリング完全なドメイン特化言語 (DSL) で、Deep Learning用に設計されており、モデル作成時の多くの反復的な時間のかかるタスクを省略します。たった14行のNMLコードで、AI Studioはほとんどのデータタイプをカバーした強力な推論モデルを生成することが可能です。

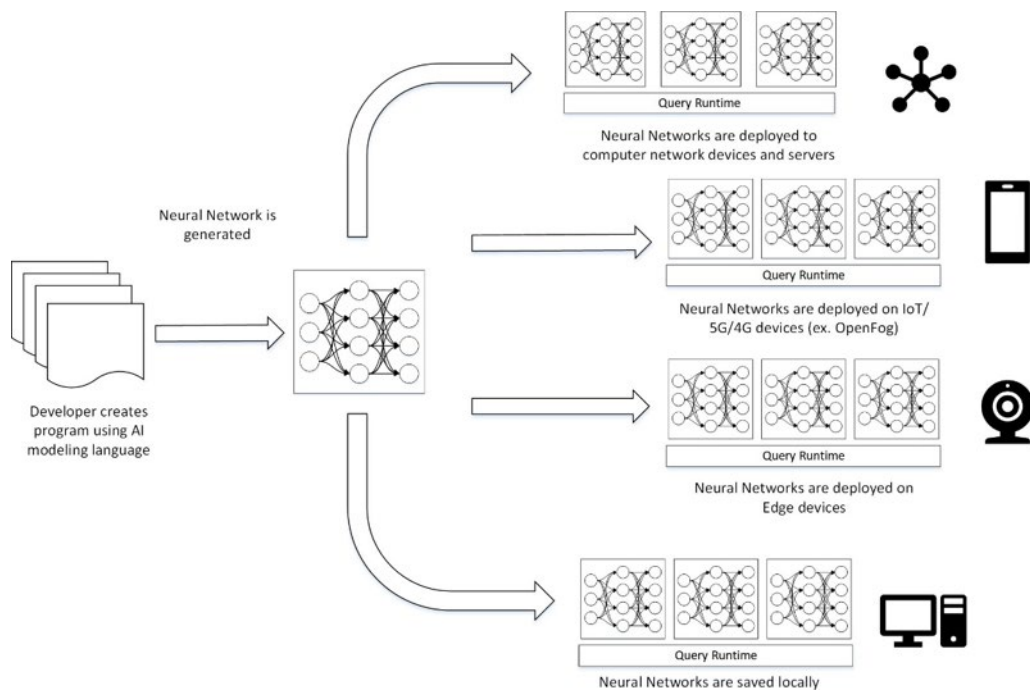


Figure 1 University deployment across a range of devices and environments

NeoPulse® Frameworkは考え得るデータタイプに対してさまざまな推論モデルを生成することが可能です。今後も新たなデータタイプを追加していきます (Table 1 参照)。

Table 1 NeoPulse® supported data types

データタイプ	説明	サポート状況
IMAGES	画像(大半のコーデックに対応)	完全サポート
VIDEOS	映像(大半のコーデックに対応) 1フレーム毎の推論モデル、マルチフレーム推論モデルの双方をサポート	完全サポート
AUDIO	音声(大半のコーデックに対応)	完全サポート
TEXT	Unicode表記されたすべての言語におけるすべてのテキスト分析及びテキストデータのクラス分類	完全サポート(日本語対応予定)
VECTOR	数値データ及びその他ベクトルデータ形式	完全サポート
DICOM	医療用画像データ	特定パートナー向けのPoC(概念実証)のみ対応
MOLECULES	バイオテクノロジー及び化学産業で用いられる可能性のある化学構造の広範囲にわたる推論モデルの生成をサポート	特定パートナー向けのPoC(概念実証)のみ対応
PROTEINS	バイオテクノロジー産業で用いられる可能性のあるタンパク質構造の広範囲にわたる推論モデルの生成をサポート	近日利用可
GENES	バイオテクノロジー産業で用いられる可能性のある遺伝子構造の広範囲にわたる推論モデルの生成をサポート	近日利用可

ワシントン大学の2人の学生が行った研究により、「Oracle」はVGGNetと同等の画像推論モデルを、格段に少ない学習エポック数で生成できることが実証されています (Figure 2 参照)。

Case Study: VGG Net

What is VGG Net?

2014 ImageNet Large Scale Visual Recognition Challenge(ILSVRC) Winner, Computer Vision Milestone ILSVRC: Image Classification Task

- Develop an algorithm that classifies photographs into different categories (e.g. Dog, Cat, Car, Person).
- Data set given by competition has ~150,000 photographs and 1000 image categories.
- Due to time and budget restrictions, we randomly chose 20 categories from the data set to train our algorithms.

Technical Approach

Recreate VGG Net Architecture and use Neopulse Studio's full auto mode to solve the ILSVRC classification task, compare the process with other popular machine learning(ML) platforms.

VGG Net architecture defined in Neopulse and custom functions to formulate architecture:

Define Architecture of VGG in Neopulse NML script and formulate parts of the architecture into functions for modular use.

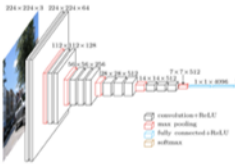


Figure 9. Macro architecture of VGGNet [3].

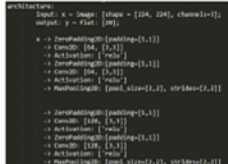
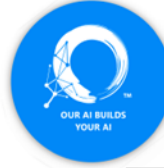


Figure 10. Architecture of VGG in Neopulse script.



Figure 11. Creating custom architecture similar to VGG Net utilizing Neopulse function construct.



Comparison With Other Popular Platforms

Detailed Codes are available at <https://github.com/GriffinTheOne/2018CapstoneFair.git>

- Platforms used: Neopulse Studio, Caffe, TensorFlow

Platform Installation

Neopulse: Pre-installed on AWS instance, no local installation required, built-in multi-GPU support.
Caffe: Directly built from source. Has many dependencies. Overall difficult.
TensorFlow: Hard to install with early Ubuntu environments. Need to install dependencies for GPU support.

Training Scripts

Neopulse: Uses call_auto mode and build architecture with Keras library API.
Caffe: Architecture available since it was originally developed on Caffe. Intuitive CLI commands.
TensorFlow: Requires python and TensorFlow basic understanding before training.

Model Query (Application)

Neopulse: Uses Portable Inference Model (PIM) which is portable with easiness.
Caffe & TensorFlow: Required to reload all the weights into a potentially huge architecture before use.

Data Preparation

Neopulse: Describe input shape/type and validation split ratio. Can specify pre-process parameters for auto processing. ~ 10 lines of code.
Caffe & TensorFlow: Manual Selection of training/validation data, file conversion and pre-processing of images required. ~70-185 lines of code.

Hyperparameter Fine-tuning

Neopulse: Fully automated and optimized process with no extra requirement from user. Utilizes spectral optimization combined with reinforcement learning.
Caffe: Full manual adjustment required. Time consuming and requires deeper understanding of ML before use.
TensorFlow: Has some default parameters to start the process.

Training Results

Results of training in Neopulse, Caffe and TensorFlow are shown in the graphs below.

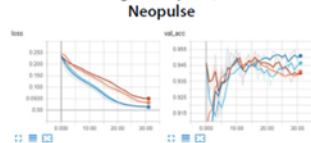


Figure 12. Training result of the Neopulse platform. ~95% Accuracy after 30 epochs.

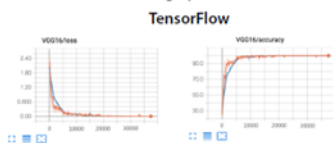


Figure 13. Training result of the Neopulse platform. ~95% Accuracy after 20,000 iterations.

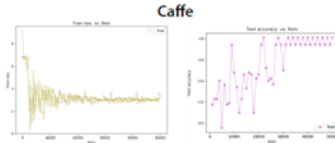


Figure 14. Training result of the Neopulse platform. ~85% Accuracy after 30,000 iterations.

Figure 2 Comparison between VGGNet and the NeoPulse® Framework

性能

NeoPulse® Frameworkは、Table 1に挙げられる様々なデータタイプに基づいた機械学習モデルを生成することが可能です。教師あり学習(クラス分類、回帰分析など)と、いくつかの教師なし学習(GAN、オートエンコーダなど)の両方をサポートしています。さらに、NeoPulse® Frameworkは企業向けの機能も幅広くサポートしています。

NeoPulse® Frameworkは、その企業向けの特色により、Figure 3に示したようにコストと時間を劇的に削減することが可能です。

企業向けの特色

- AIエンジニアリングの自動化
- マルチプラットフォーム (クラウド、PC、ARM64)
- RESTful インターフェース
- 拡張ロギング
- ビジネスワークフローとの統合
- 事業者向けスケールング
- Nvidia CUDA GPU コンピューティング
- モデル完全性、データ来歴などのセキュリティ特性

機械学習用に設計されたNMLによって、KerasとPythonを使用した場合400行必要とされたコード量が、たった14行のNMLコードとなり、コード量が50分の1に削減されます。AIの専門家でもなくともAIソリューションの生成を可能にするとともに、「Oracle」が理想的なAIモデル設計を自動選定することで必要な時間とエンジニアリング作業が劇的に少なくなります。さらに、生成されたモデルをシンプルにコピーするだけで容易に各種デバイスに展開可能となり、自動生成されるRESTインターフェースにより各種システムとの統合も容易です。

このようにしてNeoPulse® Frameworkは劇的なコスト削減をもたらします。

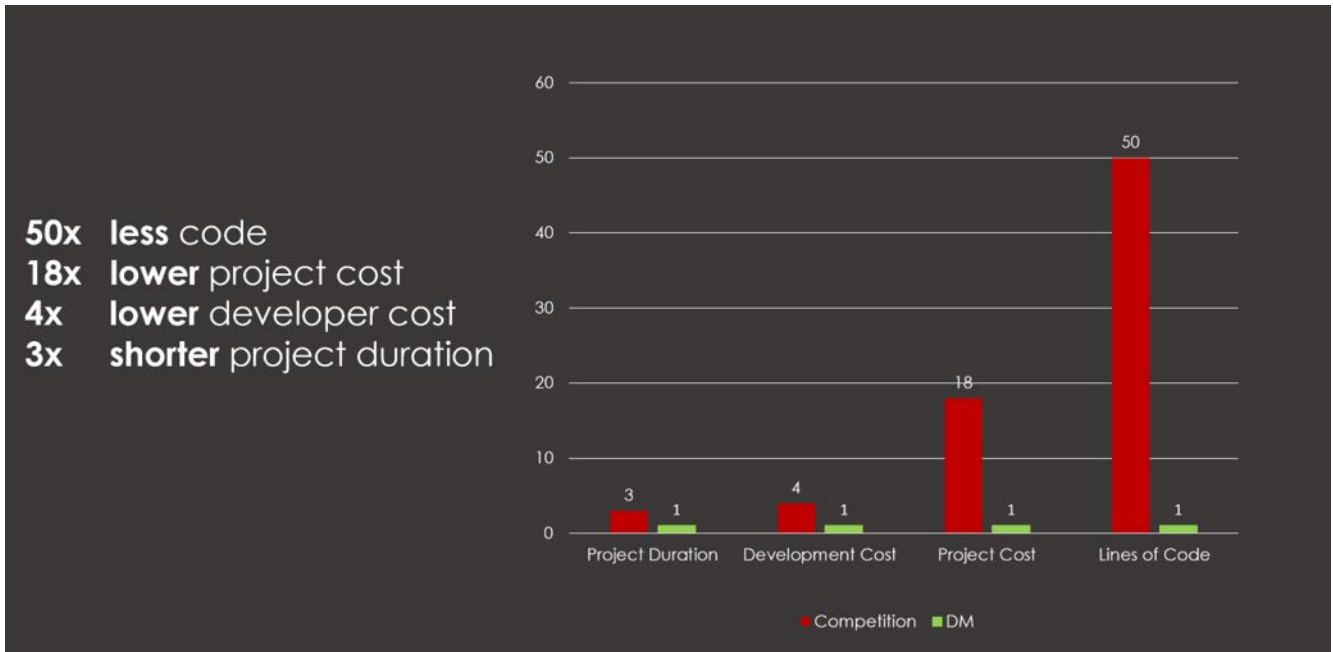


Figure 3 Cost and time savings associated with using the NeoPulse® Framework in the enterprise AI process

ワークフロー

NeoPulse® Frameworkによりモデルを簡単に作成できます。初めのステップは、CSVファイルで学習データを準備することです。

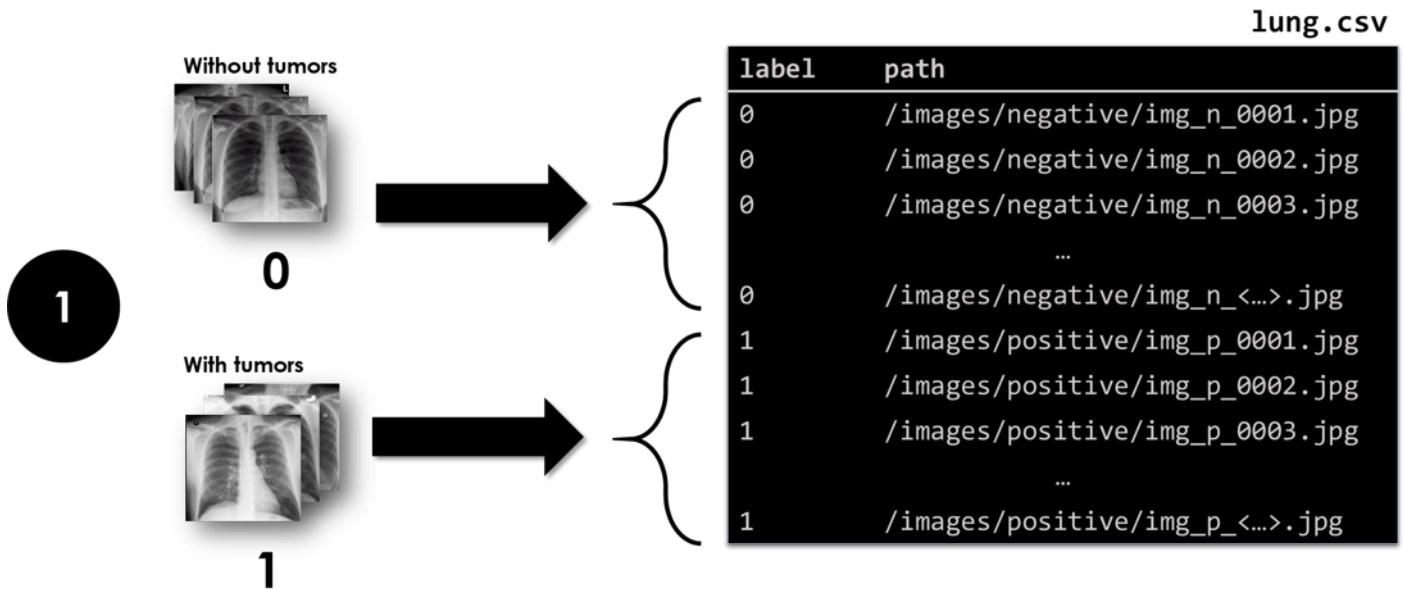


Figure 4 Step one: curating the training data

第2ステップではNMLファイルを作成します。NMLでは、学習データの抽出や前処理のような複雑なプロセスが1行で実行されます。また、NMLにキーワード「auto」を用いれば、コンパイラは「oracle」を使用して自動でアーキテクチャを選定します。

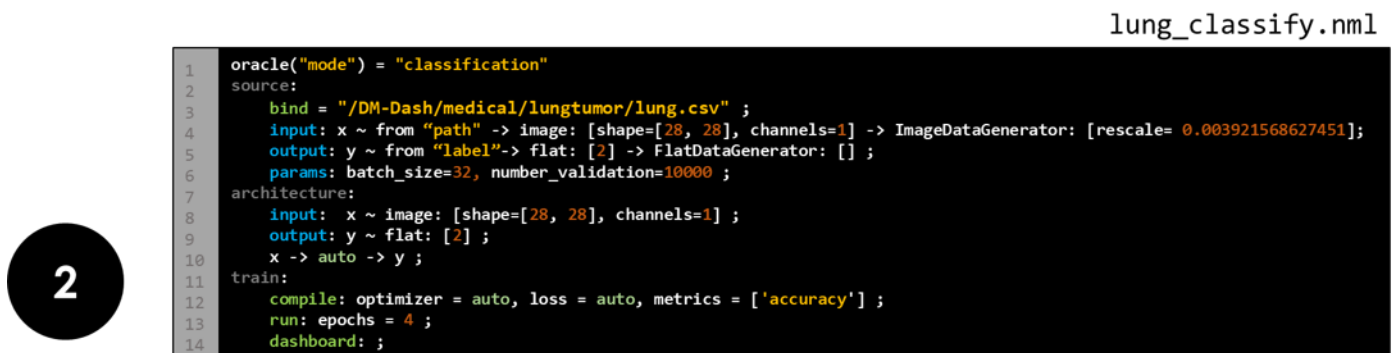


Figure 5 Step two: creating the NML file.

NMLファイルとCSVデータがコンパイルされ、モデルの学習が開始します。NeoPulse®はキュー方式を採用しているため、NeoPulse® AI Studioがビジュー状態の場合でも学習ジョブをサブミットすることが可能です。

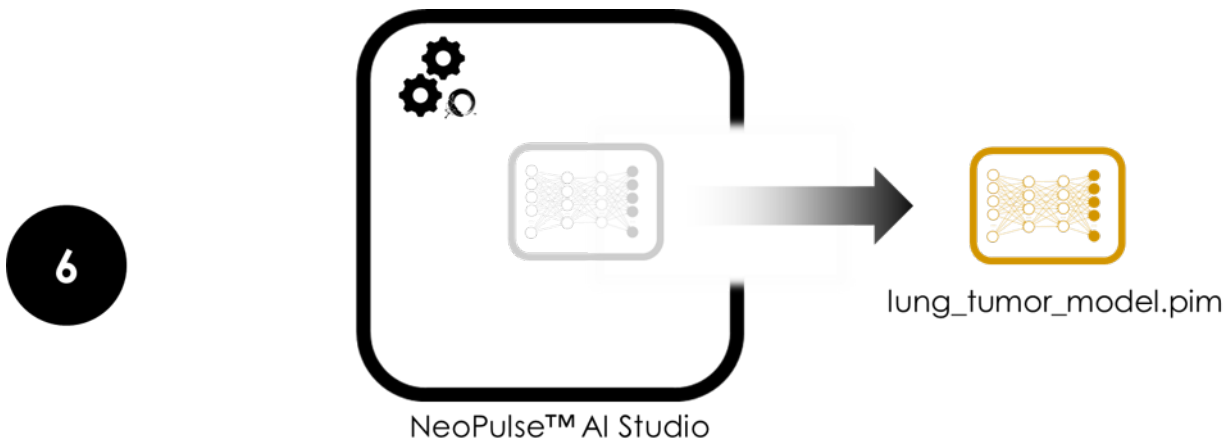


Figure 9 Exporting a model

モデルが一度エクスポートされれば、その結果生成されるPIMファイルはNPQRをインストールしたあらゆる環境に移植することが可能です。PIMファイルはNPQRにインポート可能で、NPQRは自動でRESTインターフェースを生成し、クエリ実行が可能となります。

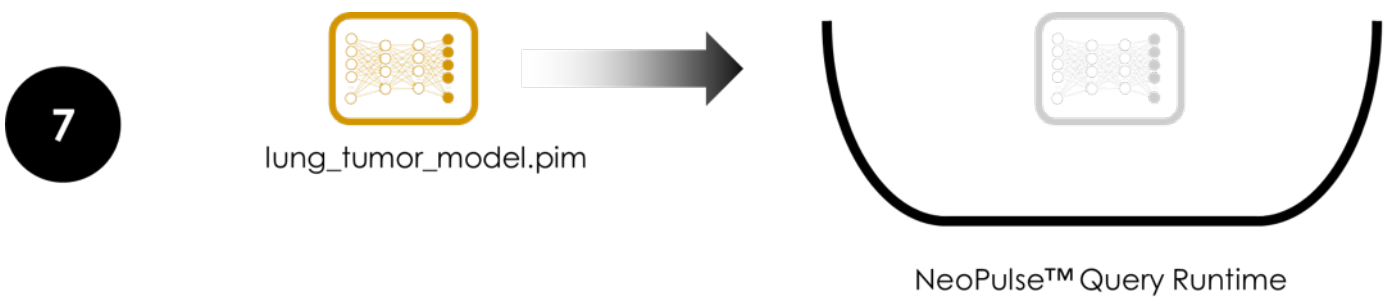


Figure 10 Importing the PIM file

最終的に、PIMファイルをホスティングしているNPQRに対して呼び出しを行うアプリケーションを作成できます。

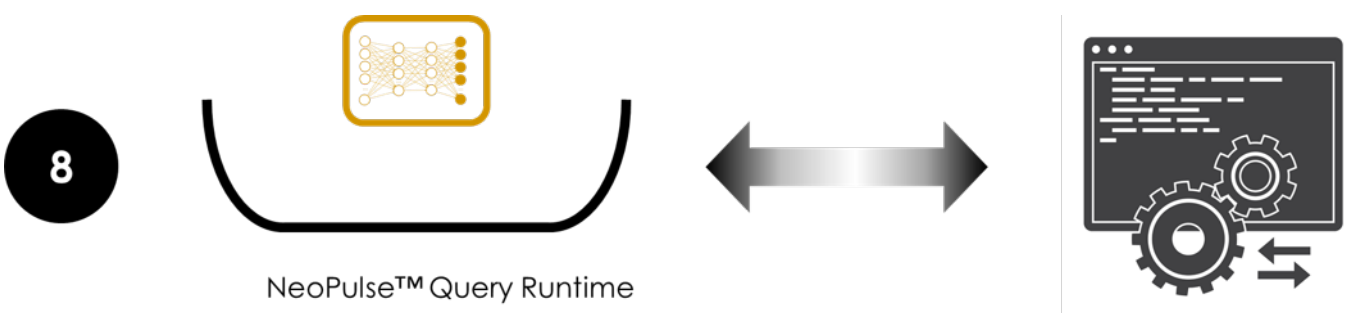


Figure 11 Integrating the Inference Model via NPQR

ワークフローの全容をFigure 12に示しています。

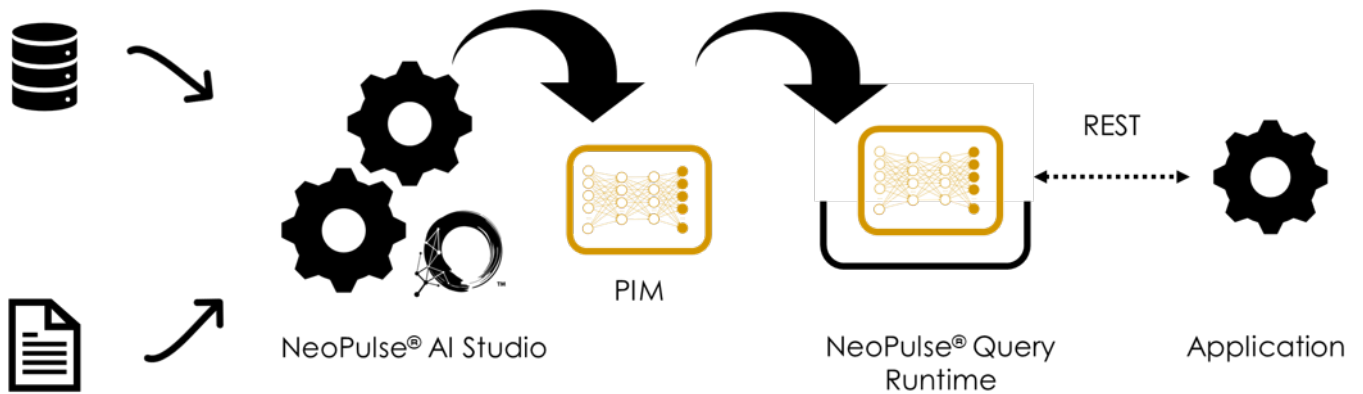


Figure 12 The NeoPulse® workflow

統合

NeoPulse® Frameworkは、NAIS及びNPQRによって提供されるAPIセットとPIMに保持されたメタデータを通して、ビジネスアプリケーションに統合されるように設計されています。アプリケーションからは、豊富なAPIセットによってNAIS及びNPQRに直接的に接続し、PIMを生成・管理・クエリ実行することが可能です。PIMにはアプリケーションが推論モデルの内容確認とバリデーションを可能とするようなメタデータが含まれています。

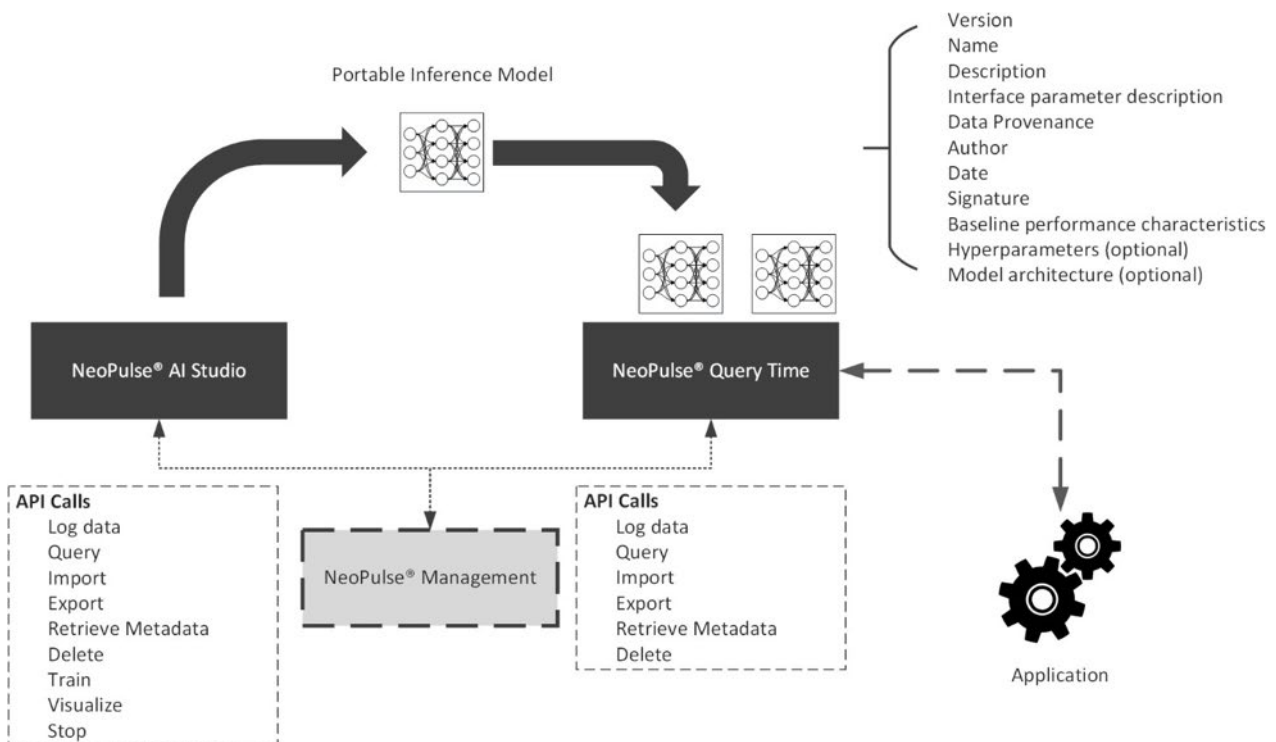


Figure 13 Integration with the NeoPulse® Framework

Figure 13は典型的なワークフローの例です。

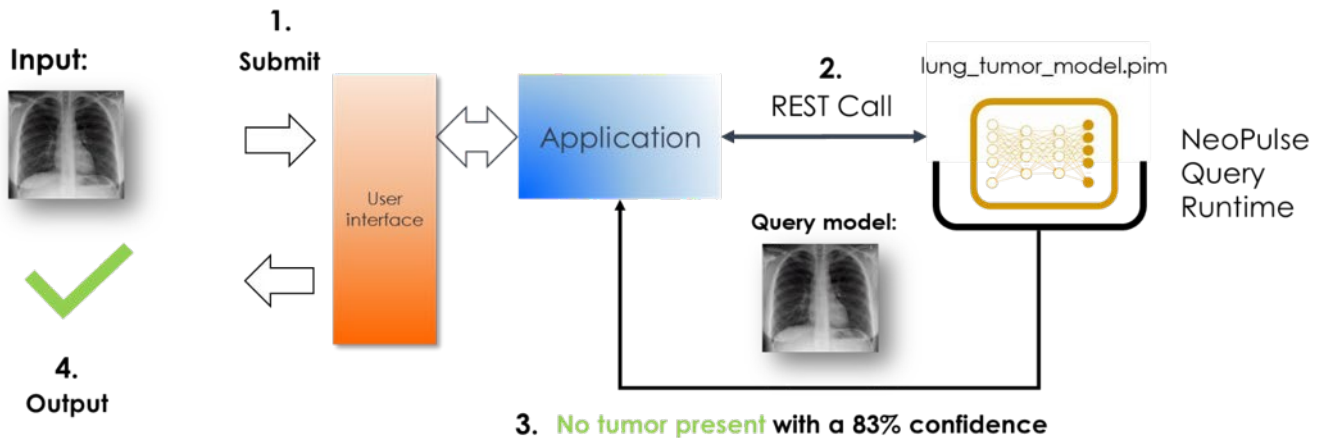
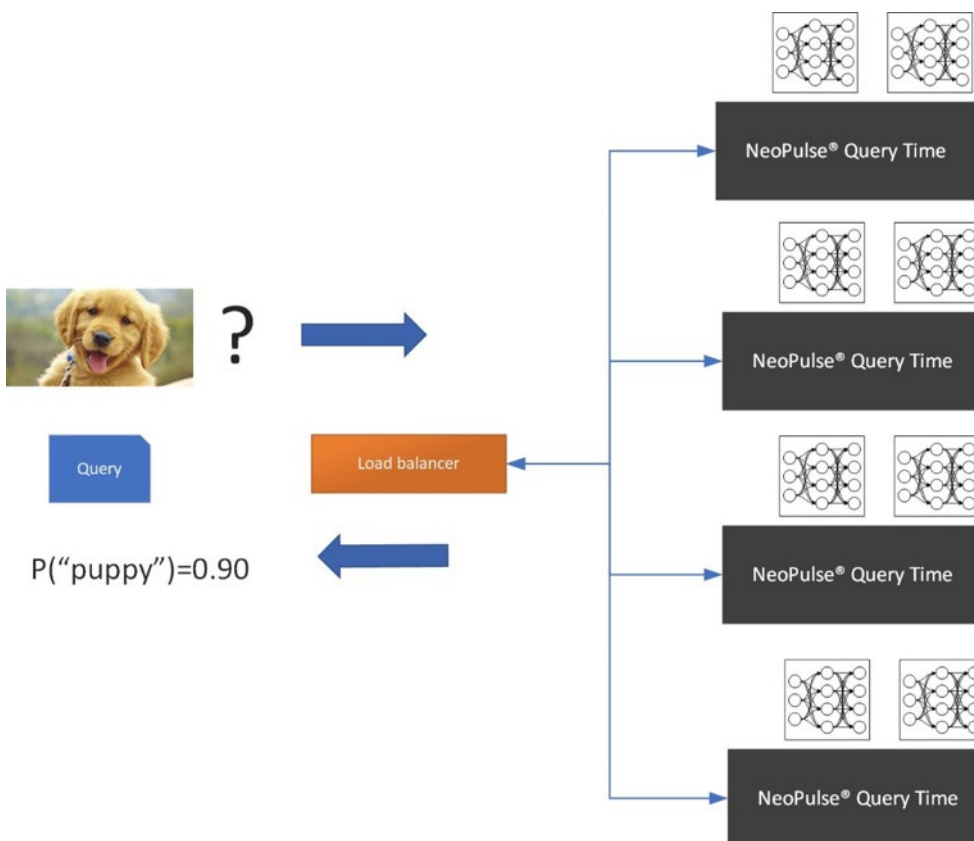


Figure 14 Application integration with the NeoPulse® Framework

それぞれのランタイムはRESTインターフェースを通じたNPQRのファームの一部として同時のクエリ実行が可能のため、クエリ規模を簡単にスケールさせることができます。



PIMを生成する際に以下の項目を指定することで、特定の環境用にクエリ機能を明示的に最適化することが可能です。

- モデルの複雑度
- モデルサイズ
- 対象エンドポイントのパフォーマンス特性(メモリ、CPU/GPU速度)
- バッチサイズ

ガバナンス

AIシステムガバナンスは、企業が取り組むべき課題としてますます重要になってきています。AIモデルには他のほとんどのソフトウェア製品と同様にライフサイクルがあり、定期的なリフレッシュが必要となります。一方、従来のソフトウェアと比較していくつかの大きな違いもあります。

- AIモデルは比較的内容が不透明であり、学習時に使用されたデータに関する情報を補完することが重要となる
- 感度・特異度・F1値などのようなモデル特性は、データが古くならないように継続的にモニタリングしておく必要がある
- 学習モデルをだます手法が存在しており、それらがどのようなものか理解することが重要となる(例: Adversarial Attacks=データに含まれる小さなノイズにより誤認識を起こすAI特有の脆弱性のこと)

これらを考慮し、各PIMIはトラッキング、モニタリング、反証、更新、照合などに使用できるメタデータセットを含んでいます。

メタデータパラメータ	説明
NAME	PIMの名称
UNIQUE ID	PIMの一意なID
DESCRIPTION	PIMの説明
VERSION	PIMのバージョン
INTERFACE PARAMETER SIGNATURE	クエリパラメータの説明(例: 画像ファイルのパス, テキストの一部, ベクトルの長さやタイプ)
DATA PROVENANCE	データの起源の耐タンパ情報
AUTHOR	モデルの作成者
DATE	モデルの作成日
SIGNATURE	モデルの一意な署名
BASELINE PERF.	F1値、感度、特異度などのモデルパフォーマンス指標の基本コレクション
HYPERPARAMETERS	ハイパーパラメータのコレクション
MODEL ARCHITECTURE	モデル構造についての説明

PIMが一度ランタイムにインポートされれば、そのパラメータへのクエリ実行が可能になります。それに加えて、ランタイムがホスティングするモデルのコレクション、マシン特性、モデルのパフォーマンス特性やログ情報など、ランタイム自身に関する情報へのクエリ実行も可能になります。

これら要素の適正配置により、プロダクションに展開されたモデルは継続的なモニタリングが可能になります。さらに、NAISとNPQRの豊富なAPIセットのサポートにより、これらに含まれるモデルはクエリ実行、削除、エクスポート、可視化が可能となります。

サポート環境

NeoPulse® Frameworkは幅広い環境でサポートされています(Table 2 参照)。デバイス機器における新しいモデル作成には計算的に多大な負荷がかかるため、これらデバイス上のアプリケーションとしてはPIMIに基づく推論モデルへのクエリ実行を可能にするランタイムのみサポートされています。

AWS SageMaker上では、NeoPulse® Frameworkの購買によりバッチクエリとリアルタイムモードクエリの両方が可能になるため、ランタイムは明示的に必要とされることはありません。

Table 2 Availability of the NeoPulse® Framework

	Description	NeoPulse® AI Studio	NeoPulse® Query Runtime
Azure	Machine instance	Yes	Yes
AWS	Machine instance	Yes	Yes
	SageMaker	Yes	Not applicable
On premise	x86, x64, AMD64	Yes	Yes
Devices	ARM64	No	Yes
	FPGA	No	Soon

エッジサイドでPIMを柔軟に展開できるため、IoTやエッジデバイスのような低パワー機器上で実行する場合でもNeoPulse® Frameworkは非常に優れています。

参考

NeoPulse® Framework Documentation:

<http://docs.neopulse.ai>

Developer portal:

<https://dimensionalmechanics.com/ai-developer>

AWS AMI:

https://aws.amazon.com/marketplace/pp/B074NDG36S?ref=cns_srchrow

Azure AMI:

<https://azuremarketplace.microsoft.com/en-us/marketplace/apps/dimensionalmechanics-inc.neopulse-ai-studio>

SageMaker Mktplace:

https://aws.amazon.com/marketplace/pp/prodview-xy3ixkod6sxok?qid=1543606611522&sr=0-1&ref_=srh_res_product_title

アジア・ソリューション&テクノロジーパートナー 日本ラッド株式会社

日本ラッド株式会社は AI Dynamics Inc.のアジア・ソリューション&テクノロジーパートナー（日本国内における正規代理店）です。認定エンジニア(NPCE: NeoPulse Certified Engineer)がコンサルティングから導入、システムインテグレーション、運用保守までお客様をトータルでサポートいたします。

お問い合わせ： neopulse@nippon-rad.co.jp

日本ラッド株式会社（東証 JQ.4736） 〒105-0001 東京都港区虎ノ門2丁目2-5 共同通信会館ビル 7F

<https://www.nippon-rad.co.jp/neopulse/>