Published in axinc

Kazuki Kyakuno Follow

Mar 10 · 6 min read

ailia SDK 1.2.14をリリース

クロスプラットフォームで利用できるGPU対応の高速AI推論フレームワークであるailia SDKのバージョン1.2.14のご紹介です。ailia SDKについてはこちらをご覧ください

ailia SDK 1.2.14の新機能は下記となります。

新しく対応するレイヤー

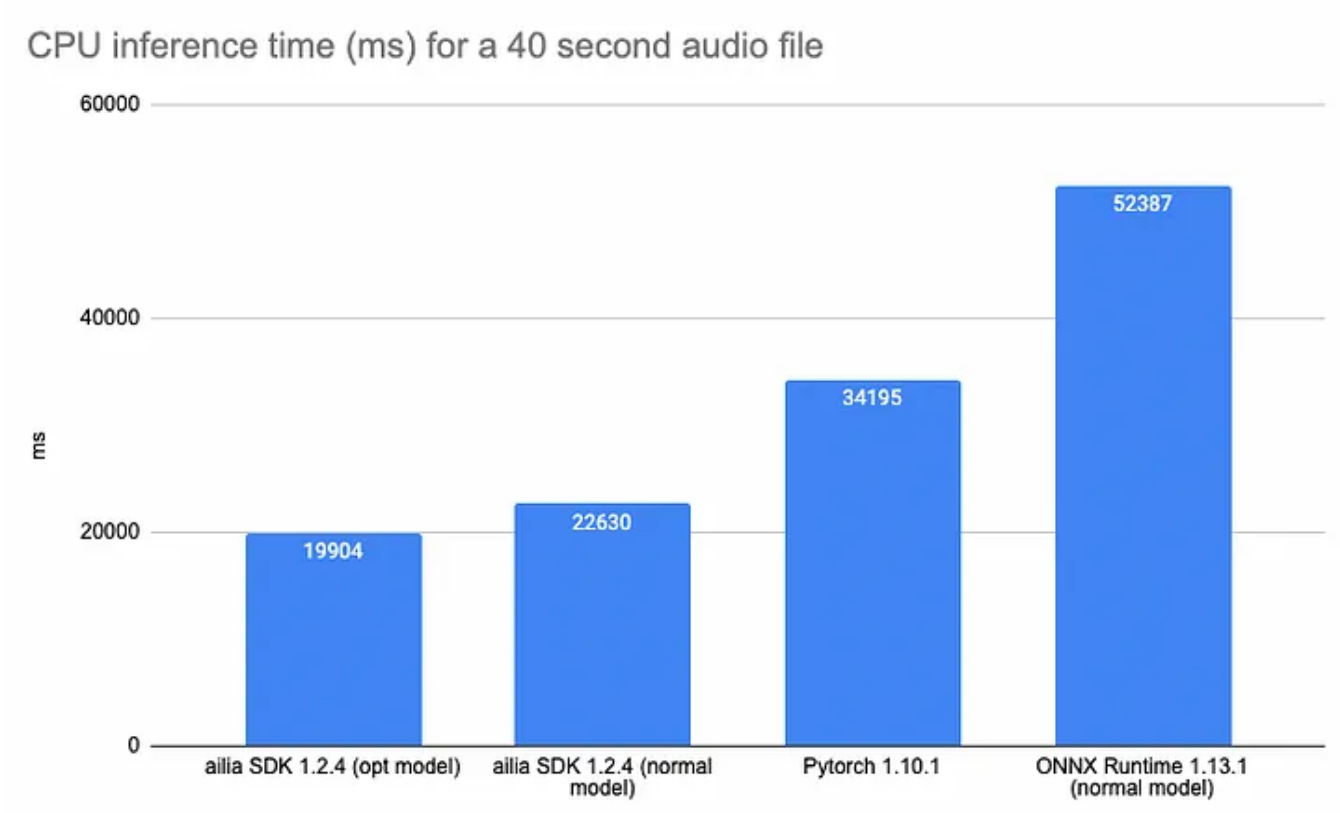
DFT、Mean、LpPool、Upsampleレイヤーに対応します。DFTはopset=17で追加されたレイヤーで、離散フーリエ変換を行うレイヤーとなります。将来的に、音声の前処理などへの活用が期待されています。

AVX512対応

GemmとConvolutionのAVX512対応を行いました。サイズの大きいGemmなどで最大で1.61倍程度、高速化されます。

Whisper向けの最適化

CPU推論においてWhisper向けの最適化を行い、Whisperの推論速度が大きく向上しました。M1 Max + macOSでの評価において、ONNX Runtimeと比較して、2.3倍程度、高速に推論可能です。また、Whisper公式のPytorchと比較して、1.7枚程度、高速に推論可能です。



Whisperのベンチマーク（40秒の日本語音声の変換時間、Whisper Small、Beam Size 1での評価）

グラフは縦軸が推論に要した時間であり、短いほど高速に動作していることを示しています。normal modelはPytorchから変換したONNXモデル、opt modelはONNXモデルにailia Optimizerを通したONNXモデルです。

Whisperはgithubのailia MODELSからお試しいただけます。

ailia-models/audio_processing/whisper at master · axinc-ai/ailia-models

Audio file Recognized speech text He hoped there would be stew for dinner, turnips and carrots and bruised potatoes and...

github.com

省メモリモードのメモリ使用量の削減

CPU推論でメモリ再利用モード（AILIA_MEMORY_REDUCE_INTERSTAGE）を使用した場合のメモリ使用量を削減しました。特に、DeticやWhisperなどの複雑なモデルで、従来よりもメモリ使用量が削減されます。例えば、WhisperのDecoder Smallにおいては、ailia SDK 1.2.13で1886MB使用していたのが、ailia SDK 1.2.14で1252MB程度まで削減されます。

Android NDKのバージョン間の互換性の修正

soにlibc++_staticのシンボルが公開されている問題を修正し、ailia SDKとは異なるバージョンのAndroid NDKを使用した場合の互換性を修正しました。

Vulkanのバージョンが混在する場合の問題の修正

Vulkan1.0のデバイスとVulkan1.1のデバイスが混在する場合に、デバイスの列挙に失敗する問題を修正しました。

プロファイルの改善

プロファイルモードでレイヤー別の消費時間を出力する機能を追加しました。これにより、ボトルネックの解析が容易になります。出力例は下記となります。

```
====Profile(Grouped by LayerType)====
LayerType       TotalPredictTime(Average)[us]   TimeRatio[%]
Convolution     1225009 43.46
Convolution/ReLU[Fused] 1003087 35.59
Eltwise 204641  7.26
ReLU    151380  5.37
Transpose       95091   3.37
Resize  51491   1.83
Concat  39713   1.41
BatchNorm       31717   1.13
MatMul  15263   0.54
Softmax 849     0.03
Convolution/Sigmoid[Fused]      394     0.01
```

```
Reshape 137     0.00
ConvertValue    47      0.00
Unsqueeze       8       0.00
Ailia_ConvBN_Convert    0       0.00
Shape   0       0.00
Slice   0       0.00
```

Unity Plugin

AiliaModelクラスでIDisposableを継承するように変更しました。これにより、Editorの開発中にメモリ不足になる可能性を低減します。

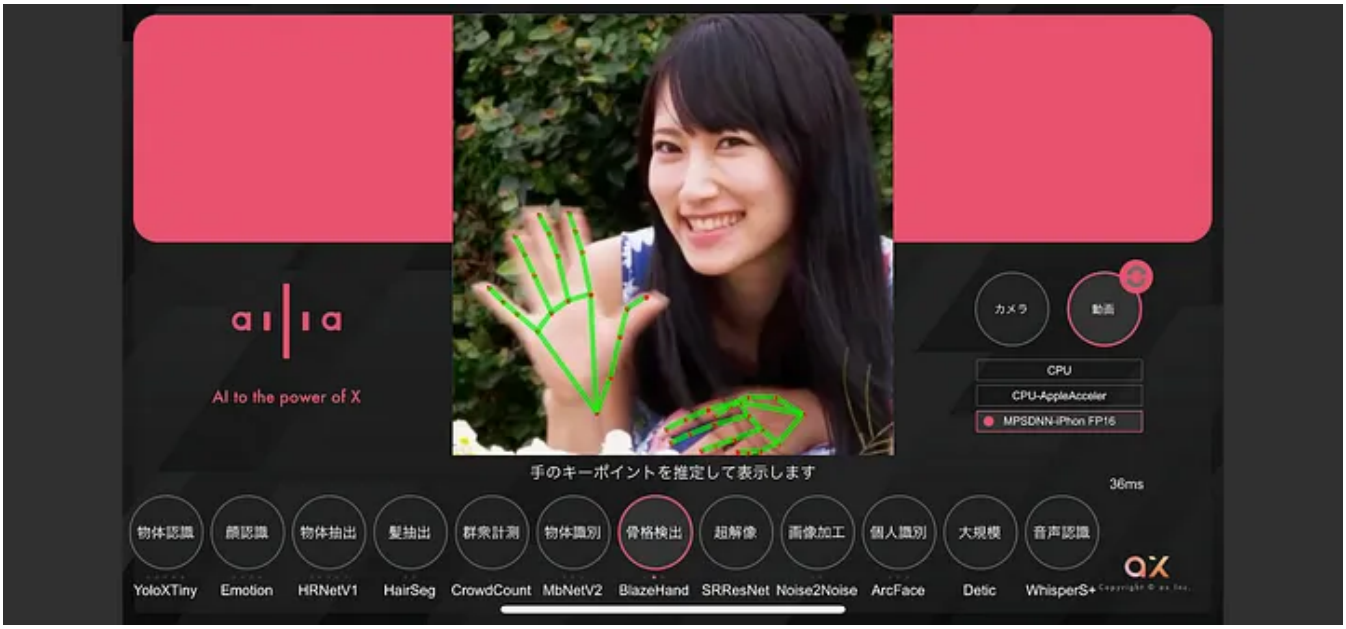
また、AppleSiliconではdylibをbundleにリネームしても読み込めない問題を修正するため、ailia.audioにbundleビルドを追加しました。

ailia AI showcaseのアップデート

ailia AI showcaseをailia SDK 1.2.14相当にアップデートし、Detic、Whisper、BlazeHand、FaceMesh、RoadSegmentationAdasが使用可能になりました。また、Android環境において、一部のモデルのailia TFLite Runtimeを使用したNPU推論に対応しました。iOS版はAppStoreから、Android版はGooglePlayからダウンロード可能です。

Deticによる物体検出

Whisperによる音声認識



BlazeHandによる手のキーポイント検出

RoadSegmentationAdasによる路面検知

ax株式会社はAIを実用化する会社として、クロスプラットフォームでGPUを使用した高速な推論を行うことができるailia SDKを開発しています。ax株式会社ではコンサルティングからモデル作成、SDKの提供、AIを利用したアプリ・システム開発、サポートまで、 AIに関するトータルソリューションを提供していますのでお気軽にお問い合わせください。

Ailia Sdk