

1
oneAPI
<HACK>ATHON

BUILD SOLUTIONS TO
UNLOCK THE POTENTIAL OF
HETEROGENEOUS COMPUTING

LEAP powered by Intel® oneAPI AI Analytics Toolkit

Problem Statement : **Open Innovation in Education**

Team Name : C5ailabs

Team Members : Rohit Sroch, Sujith R Kumar, Mohan K Rachumallu, Shubham Jain



MOOCs

(Massive Open Online Courses)

200K

Users in 2012

380M

Users in 2020

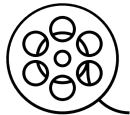
34.26%

CAGR 2022-27*

5% -10%

Completion rate

Key Challenges



Lengthy
videos



Instructor
Availability



Slow response
from forums



No real time
Q&A/Mentor

Approach

LEAP

(Learning Enhancement and Assistance Platform)



AI based
platform



Powered by
Intel OneAPI



Quality
Education



All time
Availability

Key Features of LEAP



Ask Question/Doubt



Conversational AI Examiner

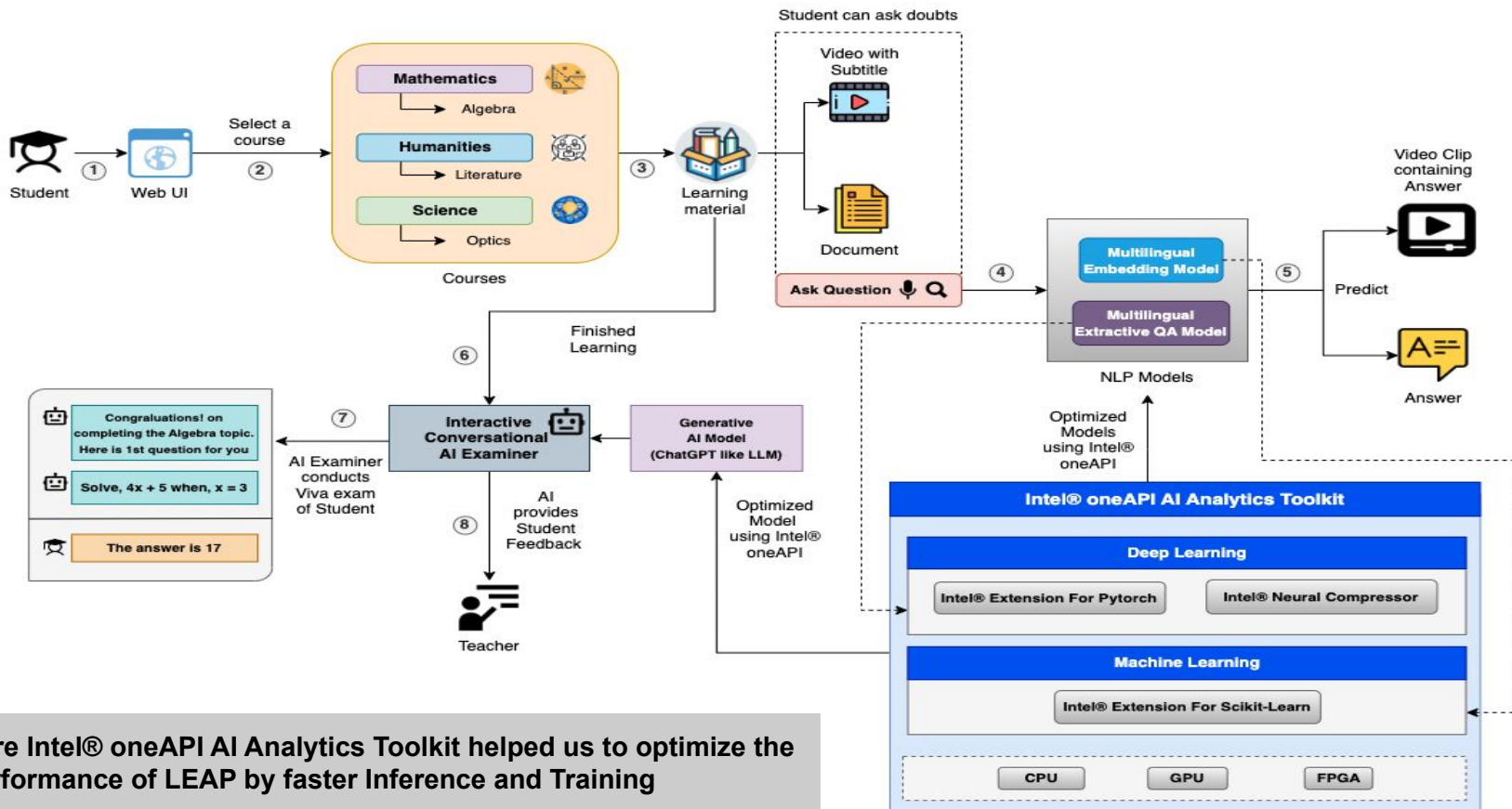


Feedback from AI Examiner

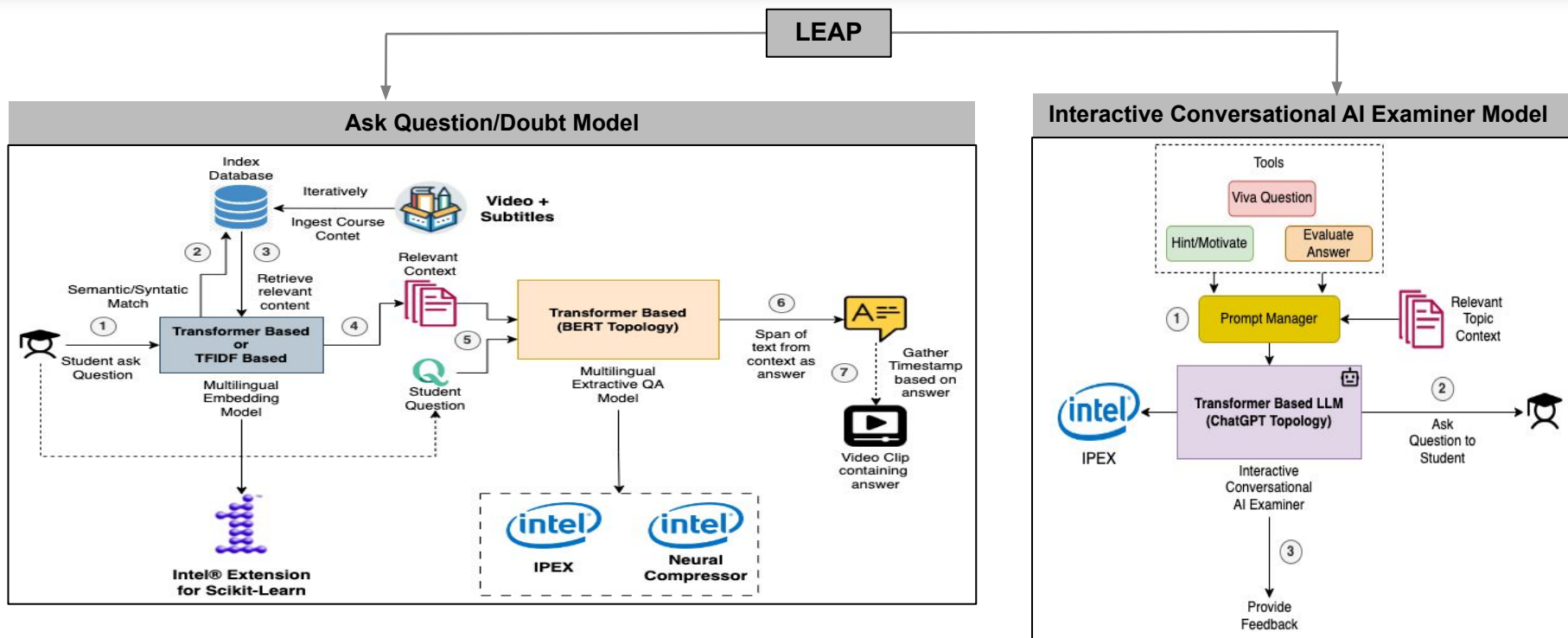


Multilingual Support

Reference: [PRNewswire](#) , [Edtechreview](#); [holonig](#)



Here Intel® oneAPI AI Analytics Toolkit helped us to optimize the performance of LEAP by faster Inference and Training

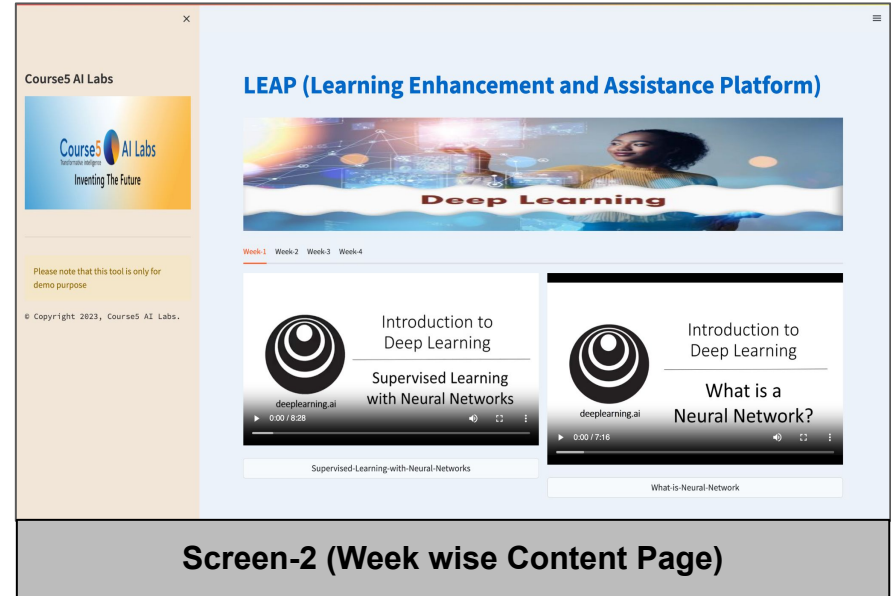
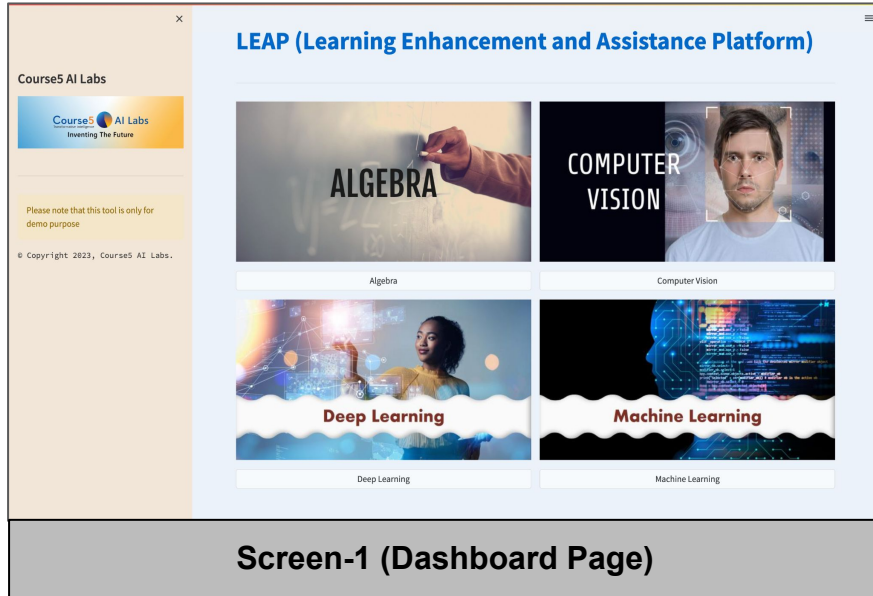


| Ask Question/Doubt Model | | | | |
|-------------------------------------|-----------------------|-----------------------|-----------------|-------------------------|
| Extractive Question Answering Model | | | | |
| | Pytorch (Base) - FP32 | Pytorch (IPEX) - FP32 | Static-QAT-INT8 | Static-Smooth-QA T-INT8 |
| Latency (milli sec) | 64.513 | 39.329 | 14.514 | 15.24 |
| Throughput (samples/sec) | 15.501 | 25.427 | 68.9 | 65.616 |
| F1 Score (SQuAD-v1) | 76.11 | 76.11 | 75.72 | 75.72 |

| Interactive Conversational AI Examiner Model | | |
|--|---------------------|----------------------------------|
| TFIDF Embedding Model | | |
| | Scikit-Learn (Base) | Intel Extension For Scikit-Learn |
| Latency (milli sec) | 0.761 | 0.752 |
| Throughput (samples/sec) | 1313.63 | 1330.49 |

Table: Latency/Throughput/Speed-Up Benchmark result for **our Extractive Question Answering ALBERT Model (Multilingual) and TFIDF Embedding Model** on Intel® Dev Cloud machine (**Intel® Xeon® Platinum 8480+ (4th Gen: Sapphire Rapids) - 224v CPUs 503GB RAM**) with optimization using IPEX-FP32, Static-QAT-INT8 using Intel® Neural Compressor and TFIDFVectorizer using Intel® Extension for Scikit-Learn.

Link: <https://www.youtube.com/watch?v=M51BFcoJa3k>



Course5 AI Labs


Course5 AI Labs

Inventing The Future

Please note that this tool is only for demo purpose

© Copyright 2023, Course5 AI Labs.

LEAP (Learning Enhancement and Assistance Platform)




Introduction to Deep Learning

Supervised Learning with Neural Networks

Video Transcript:

00:00:33.320 — 00:00:34.890
There's been a lot of hype about neural networks. And perhaps some of that hype is justified, given how well they're working. But it turns out that so far, almost all the economic value created by neural networks has been through one type of machine learning, called supervised learning. Let's see what that means, and let's go over some examples. In supervised learning, you have some input x , and you want to learn a function mapping to some output y . So for example, just now we saw the housing price prediction application where

Ask Doubt:



Screen-3 (Ask Question/Doubt Page)

Course5 AI Labs


Course5 AI Labs

Inventing The Future

Please note that this tool is only for demo purpose

© Copyright 2023, Course5 AI Labs.

what is ReLU




ReLU function which stands for rectified linear units.

Get More Info

Supervised Learning

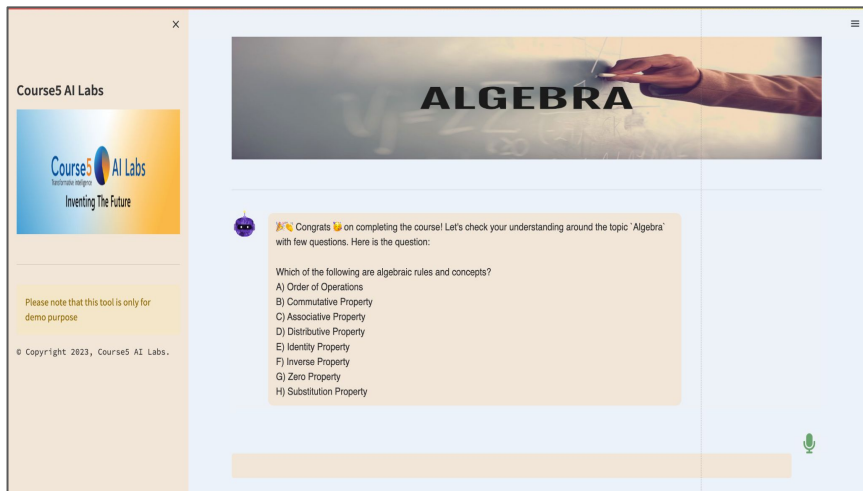
| Input(x) | Output (y) | Application |
|---------------|---------------------|--------------------|
| Home features | Price | Real Estate |
| Ad, user info | Click on ad? (0/1) | Online Advertising |
| Image | Object (1.....1000) | Photo tagging |

1:50 / 8:28



takes a max of zero, and then outputs the estimated price. And by the way in the neural network literature, you see this function a lot. This function which goes to zero sometimes and then it takes of as a straight line. This function is called a ReLU function which stands for rectified linear units. So ReLU. And rectify just means taking a max of 0 which is why you get a function shape like this. You don't need to worry about ReLU units for now but it's just something you see again later in this course.

Screen-4 (Ask Question/Doubt Page)

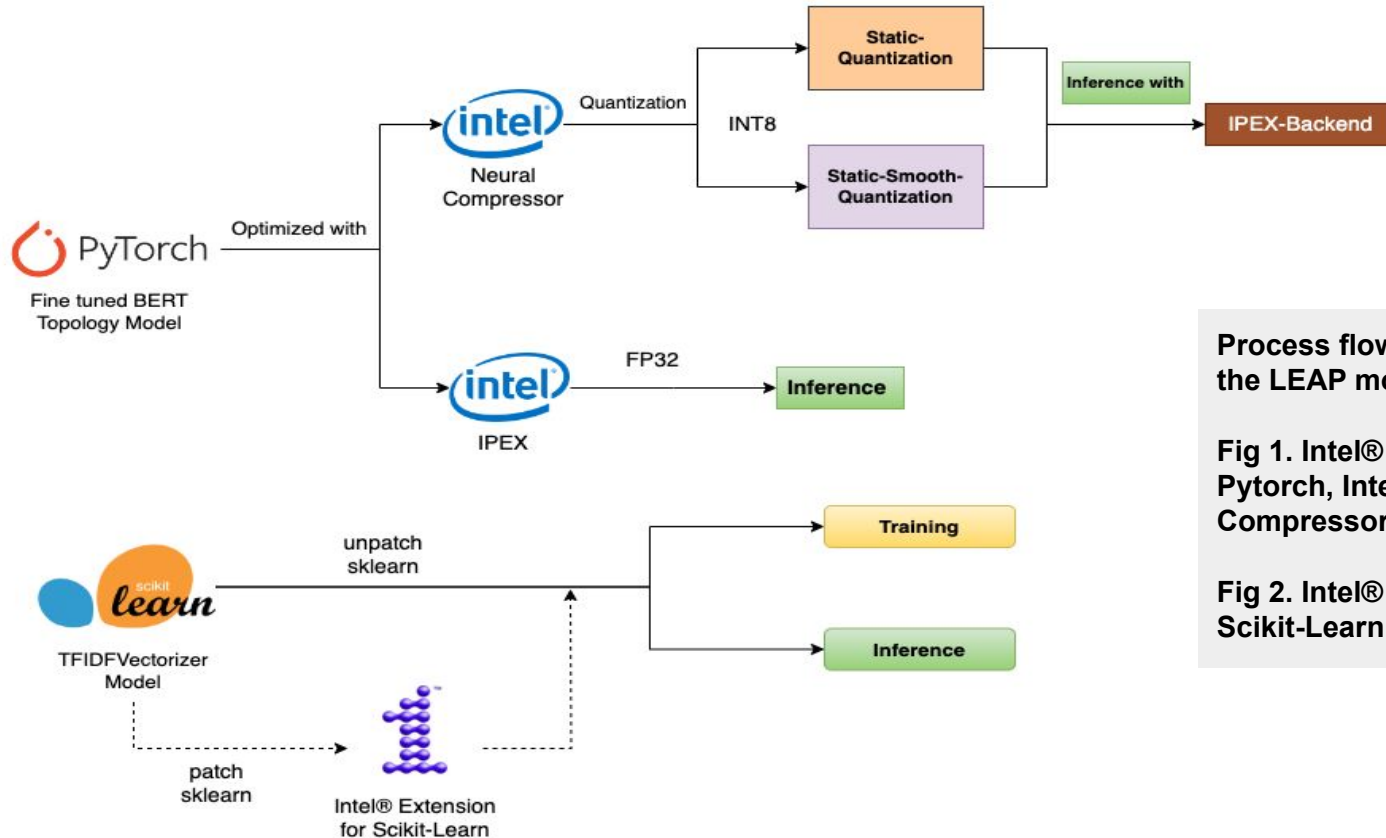


Screen-5 (Interactive Conversational AI Examiner Asks Question to student)



Screen-6 (Interactive Conversational AI Examiner provides hints and motivates a student in case of a wrong answer)

<https://github.com/rohitc5/intel-oneAPI>

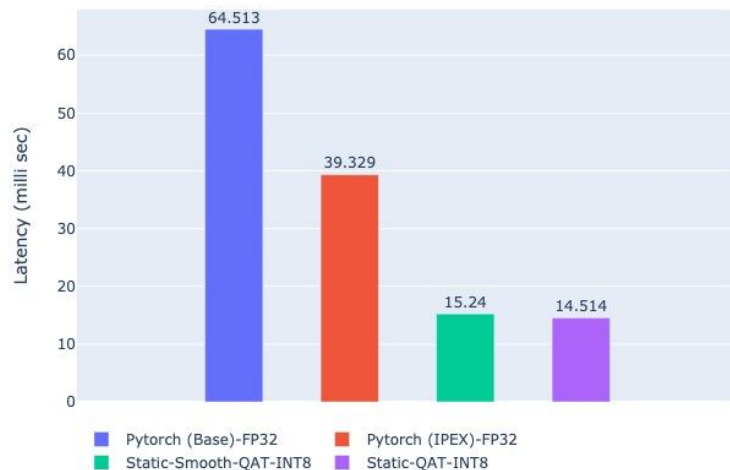


Process flow to optimize the LEAP models by using

Fig 1. Intel® Extension for Pytorch, Intel® Neural Compressor and

Fig 2. Intel® Extension for Scikit-Learn

Extractive QA Model Latency Comparison



Extractive QA Model Speed Up Comparison

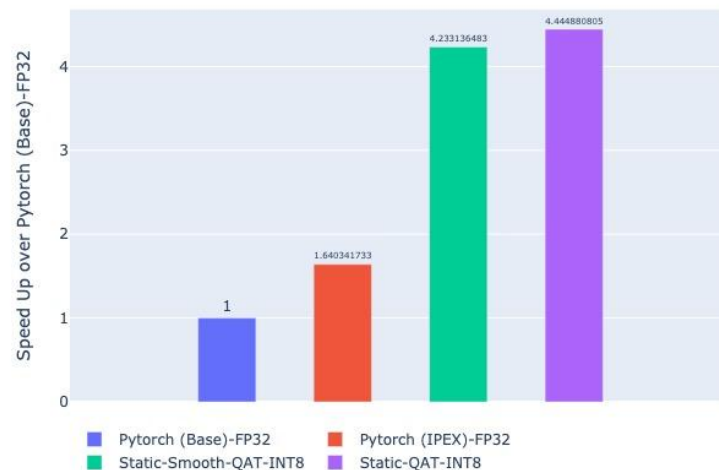
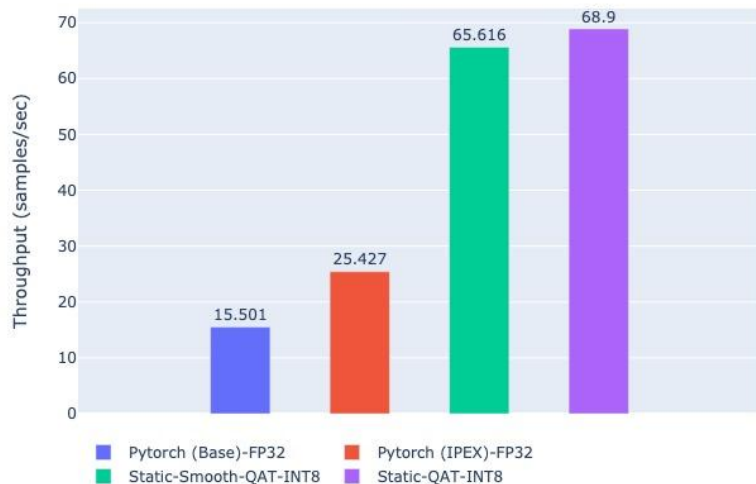


Fig: Latency/Speed-Up Benchmark result for our **Extractive Question Answering ALBERT Model (Multilingual)** on Intel[®] Dev Cloud machine (**Intel[®] Xeon[®] Platinum 8480+ (4th Gen: Sapphire Rapids) - 224v CPUs 503GB RAM**) with optimization using IPEX-FP32 and Static INT8-Quantization using Intel[®] Neural Compressor.

For Ask Question/Doubt Extractive QA Model

Extractive QA Model Throughput Comparison



Extractive QA Model F1 Score (SQuAD-v1) Comparison

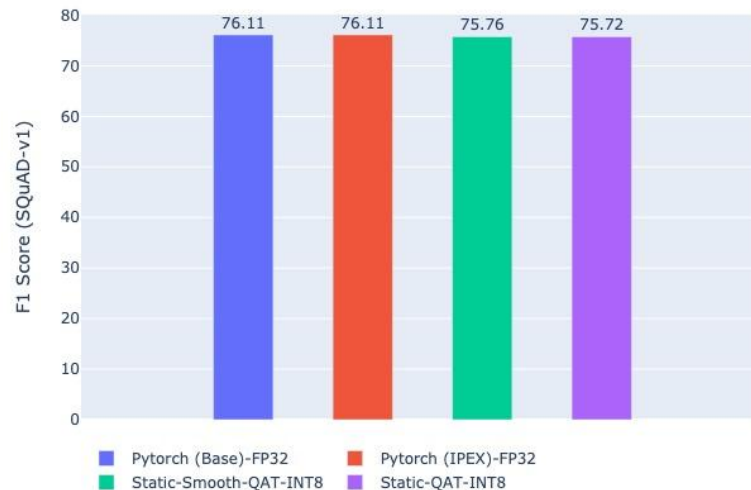
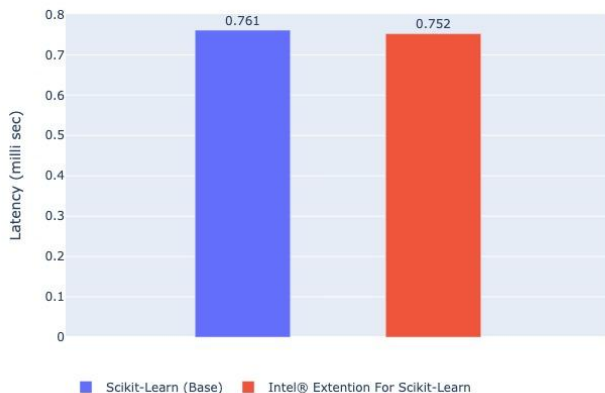


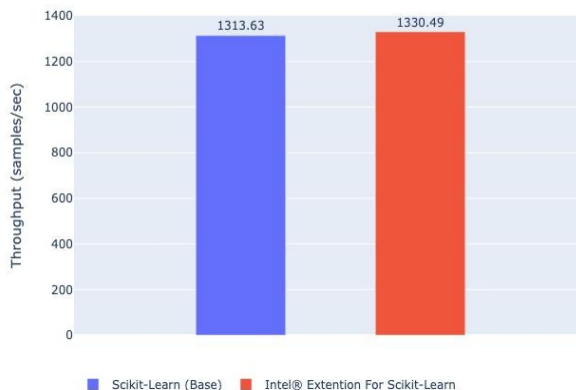
Fig: Throughput/F1 Score Benchmark result for **our Extractive Question Answering ALBERT Model (Multilingual)** on Intel[®] Dev Cloud machine (**Intel[®] Xeon[®] Platinum 8480+ (4th Gen: Sapphire Rapids) - 224v CPUs 503GB RAM**) with optimization using IPEX-FP32 and Static INT8-Quantization using Intel[®] Neural Compressor. Also, the model (<https://huggingface.co/ai4bharat/indic-bert>) was fine-tuned on SQuAD-v1 dataset.

For Ask Question/Doubt Extractive QA Model

TFIDF Embedding Latency Comparison



TFIDF Embedding Throughput Comparison



TFIDF Embedding Training Time Comparison

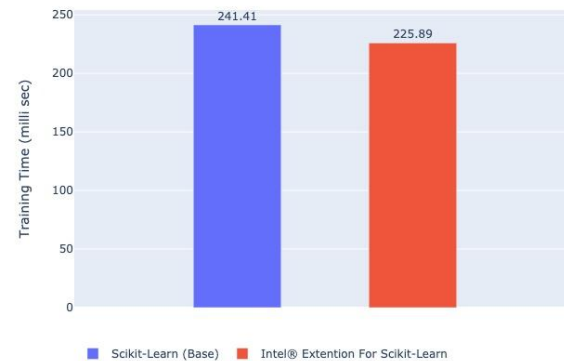


Fig: Benchmark results for **TFIDFVectorizer** Embedding model during training and inference on Intel® Dev Cloud machine (Intel® Xeon® Platinum 8480+ (4th Gen: Sapphire Rapids) - 224v CPUs 503GB RAM). Please Note that we don't see much of a difference may be because we used a tiny dataset.

For Ask Question/Doubt Embedding Model

<https://huggingface.co/rohitsroch>

1
oneAPI
<HACK>ATHON

THANK YOU