ONNX

ONNX Model Zoo/Tutorials Sig Updates

Presenter:
Jacky Chen (Microsoft US)

Outlines

- **Introduction**
- **ONNX Model Zoo**
  - Latest ONNX Model Zoo models
  - Test coverage
  - Improvements
  - ONNX Model Zoo x Hugging Face
  - Roadmap
- **ONNX Tutorials**
  - Improvements
  - Roadmap
- **Welcome to contribute!**

Introduction

- ONNX Model Zoo
  - A collection of pre-trained, state-of-the-art models in the ONNX format
  - 40 kinds of ONNX models and 168 models (with different ONNX version) in total
    - 35 vision-based models including classification, object detection, super resolution
    - 5 models about machine comprehension
- ONNX Tutorials
  - Tutorials demonstrating how to use ONNX in practice for varied scenarios across frameworks, platforms, and device types

ONNX

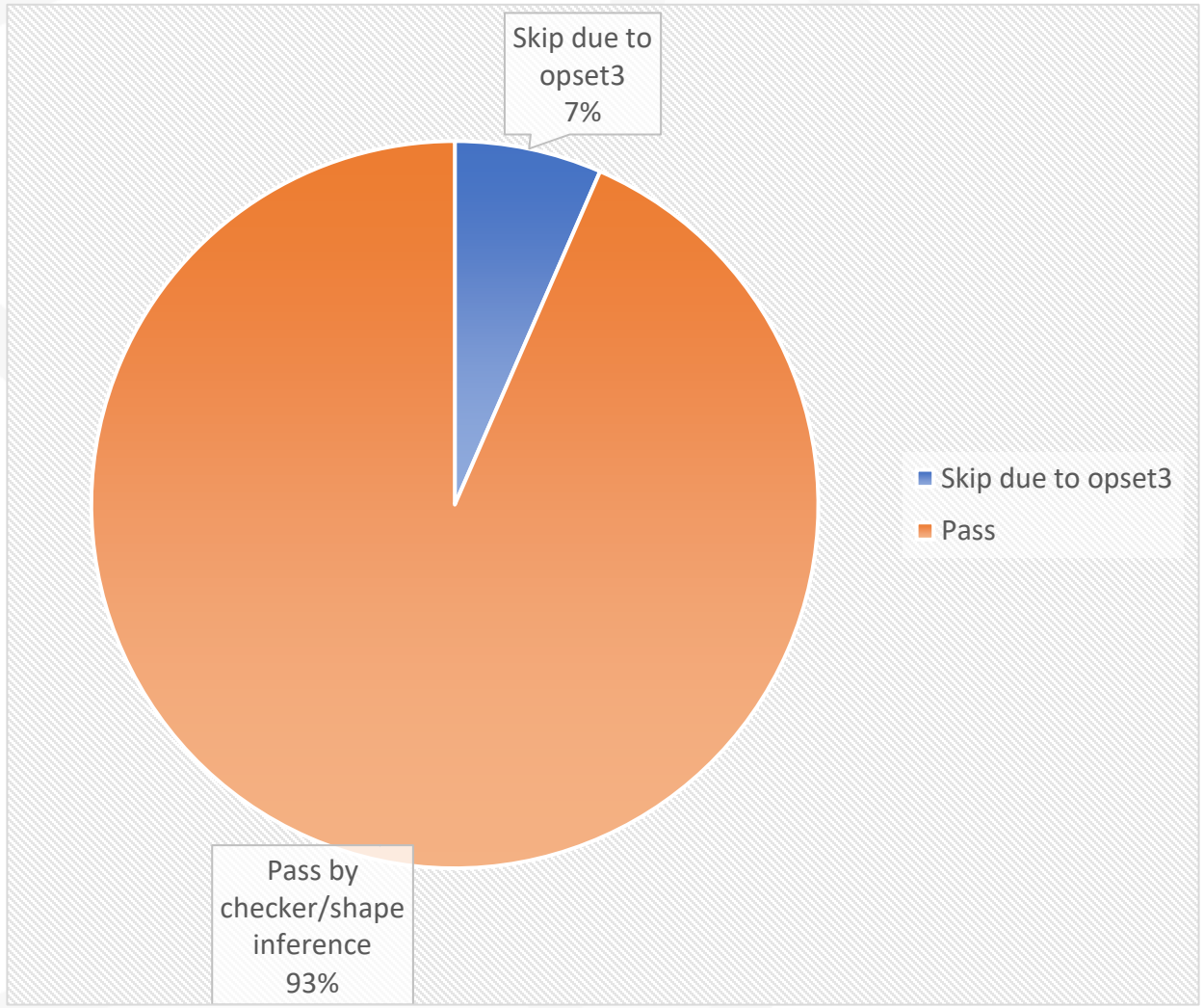
ONNX Model Zoo

a collection of pre-trained, state-of-the-art models

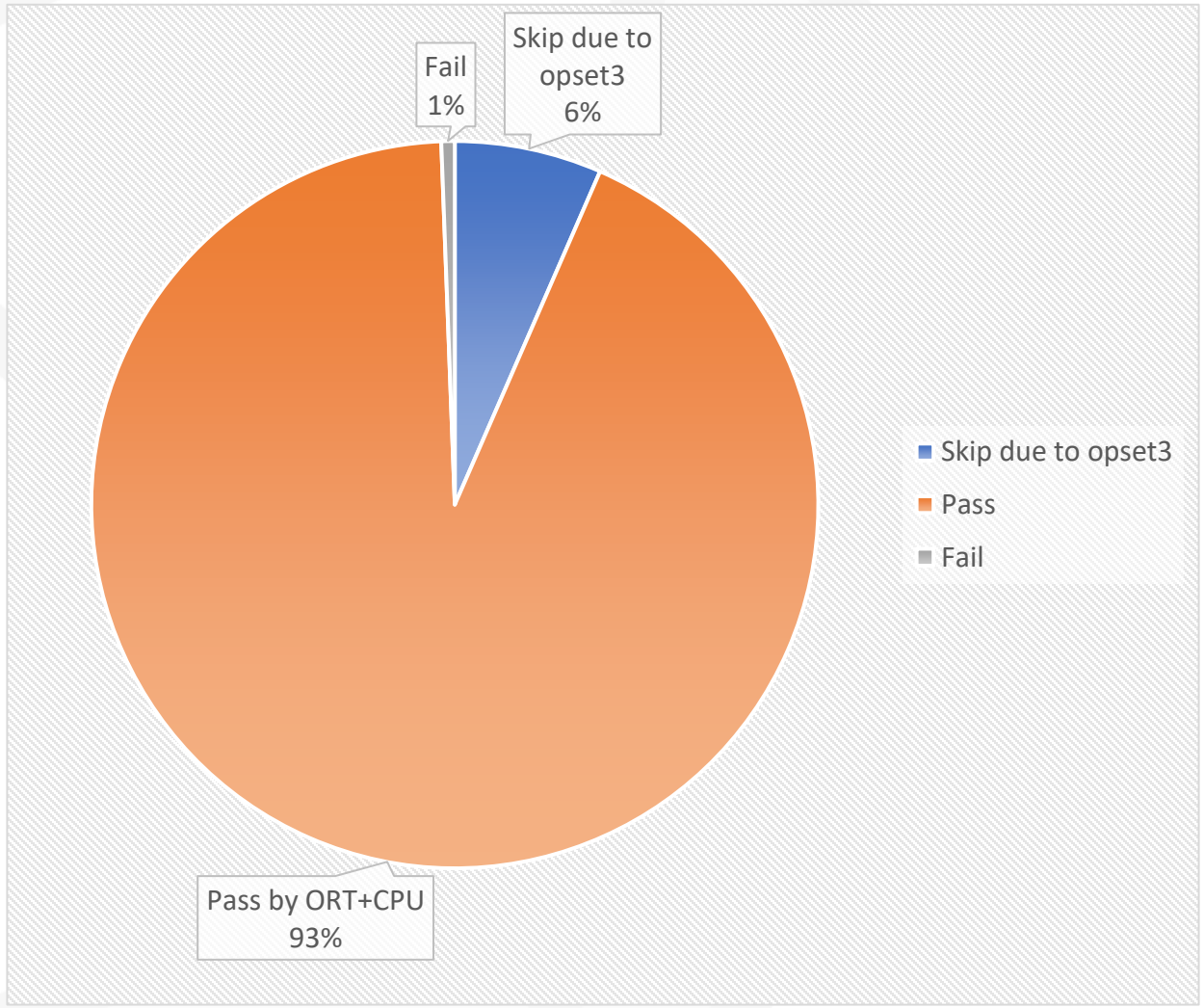
Latest ONNX Model Zoo models

- More quantized models (int8)
  - Vision: AlexNet, CaffeNet, GoogleNet, SqueezeNet, ZfNet, EfficientNet, Inception, SSD, FCN, MobileNet, Faster-RCNN, Yolo, Mask_RCNN, DenseNet
  - Text: Bert-squad
- Thank Mengni from Intel for her contribution!

Test coverage by ONNX 1.12



Test coverage by ORT 1.11



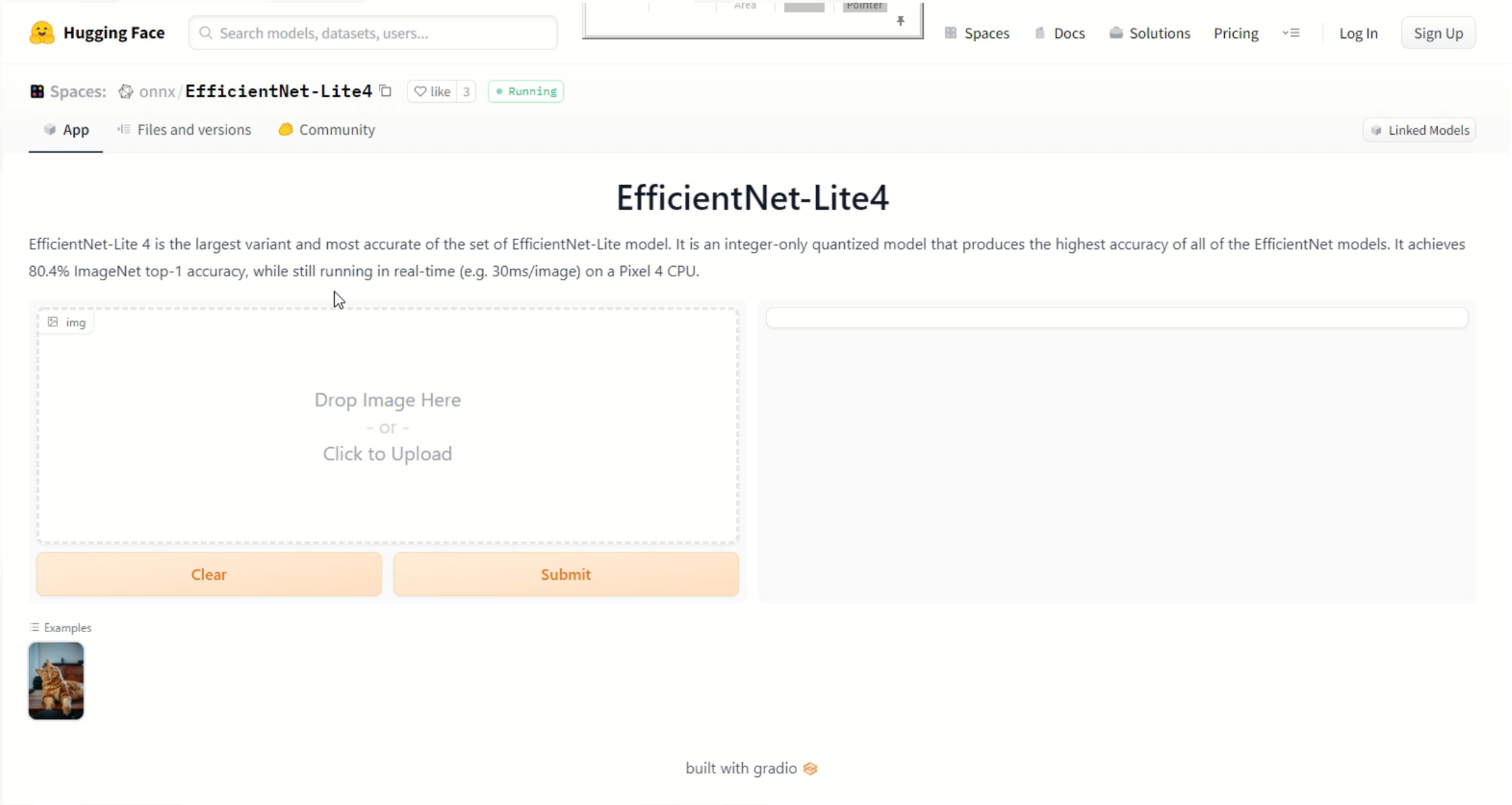
Improvements (ONNX Model Zoo)

- CI improvements
  - Tested uploaded .onnx and .tar.gz (including test_data_set)
  - Detected avx512 support in CI machines for quantized models
- Fixed all of broken test data
- Migrated to main branch
- Introduced Hugging Face Spaces for ONNX Model Zoo

ONNX Model Zoo x Hugging Face

- Collaborated by Gradio, HF Spaces, ONNX Runtime, ONNX Model Zoo
- Gradio: fast Python Web App to demo ML model
- Simplified the complex development process and demonstrate accurate inference results with a friendly Web UI
- Added a lot of ONNX Model Zoo models in HF spaces
- Thank Ahsen Khaliq from Hugging Face for his contribution!
- Reference
  - Gradio app tutorial: Gradio And ONNX On Hugging Face
  - Microsoft cloud blog: Live demos of machine learning models with ONNX and HF Spaces

Demo


Hugging Face
Search models, datasets, users...
Spaces
Docs
Solutions
Pricing
Log In
Sign Up
Spaces: onnx / EfficientNet-Lite4
like 3
Running
App
Files and versions
Community
Linked Models
EfficientNet-Lite4
EfficientNet-Lite 4 is the largest variant and most accurate of the set of EfficientNet-Lite model. It is an integer-only quantized model that produces the highest accuracy of all of the EfficientNet models. It achieves 80.4% ImageNet top-1 accuracy, while still running in real-time (e.g. 30ms/image) on a Pixel 4 CPU.
img
Drop Image Here
- or -
Click to Upload
Clear
Submit
Examples
built with gradio

Roadmap (ONNX Model Zoo)

- Deal with legacy operators and models
- Improve CI to verify models, sample codes and test data
- More quantized and mixed precision models
- Provide training example models

ONNX

ONNX Tutorials

Tutorials demonstrating how to use ONNX in practice for varied scenarios across frameworks, platforms, and device types

Improvements (ONNX Tutorials)

- Introduced CI pipelines
  - Add CIs to validate URLs for new PRs
  - Add CI to validate all URLs weekly
  - Run flake8 check in CI to ensure Python code quality in existing notebooks
- Removed old invalid URLs
- Fixed existing flake8 failures

Roadmap (ONNX Tutorials)

- Polish old/deprecated tutorials
- Prefer URL redirection to other tutorials

Welcome to contribute!

- Upload new ONNX models

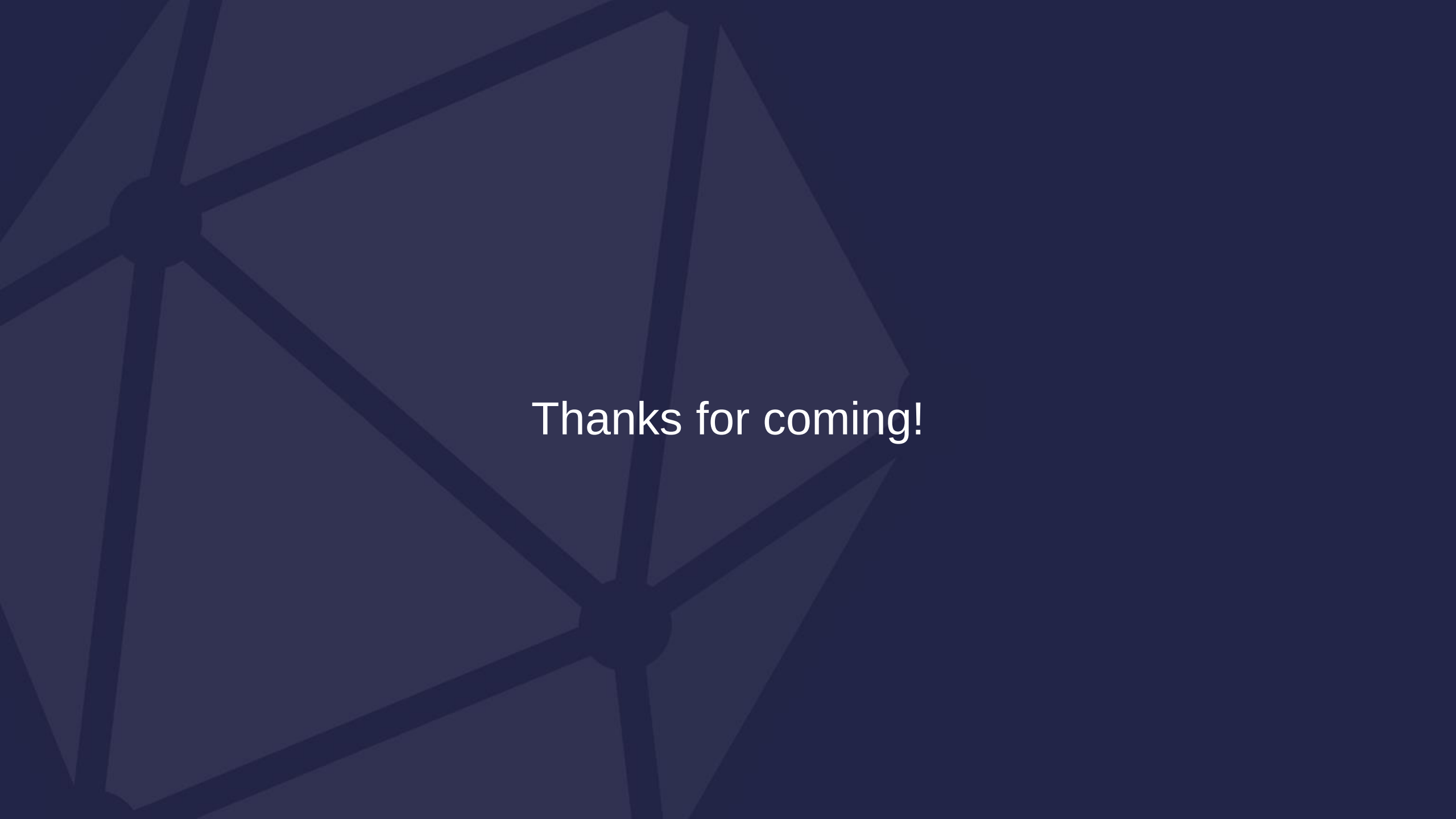
Files needed for PR

- ONNX Model file
- Test input/output data
- Readme.md
- (Optional) Inference example/tutorial
- (Optional) Hugging Face Space link

Model verification

- Check by onnx.checker/shape_inference
- ORT inference test on test data with CPU EP

- Discussion: join us on Slack in #onnx-modelzoo channel
- Help to review pull requests (a few open are waiting)
- Look for volunteered approvers for SIG-modelstutorials


Thanks for coming!