Architecture Challenges for Transformer Models

Cheng Wang
CTO & SVP, Software & Architecture
Flex Logix
Blind person and the elephant – Receptive field is a big limitation
CNNs are limited by receptive field

Smaller receptive field
Smaller features & objects

Larger receptive field
Larger features & objects
Transformers use context from whole image

CNN: Elephant?

Transformers: Elephant!

Receptive Field

MAYBE

Receptive Field

VERY HARD

Receptive Field
DETR 2020 – The de-facto vision transformer model

- Uses CNN backbone for feature extraction & transformer “head”
- Transformer Encoder extracts features from all patches for context
- Decoder makes predictions based on all extracted features
- Transformer Encoder/Decoder operations are very different from CNN
InferX provides flexibility essential for transformers:

- Each TPU can stream data with: TPU, L1 weight mem, L2 Data mem & DDR
- TPU natively supports mixed precision
- Flexible activations in EFLX eFPGA
- More data bandwidth vs Network-on-Chip based AI
  - Also more flexible data manipulation
Diving into vision transformers

• First stage is positional encode:
  • PE values are stored eFPGA ROMs
  • EFLX lookup PE “on the fly” to add to the K/Q matrix into the attention head

\[
PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right) \\
PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right)
\]
• Second stage multiplies input with 3 matrices for each head (Q/K/V):
  • Each matrix maps to TPU weights
Main part of multi-head attention layer is a challenge on traditional edge accelerators:
- The \((Q, K, V)\) for each matrix is activation data
- \(Q \times K^T\) multiplies 2 activation data:
  - InferX can load activation into weight memory
Diving into vision transformers (4)

- Softmax and normalization operators are also challenging on int8 datapaths
  - InferX mixed-precision includes BFloat16 format
  - Enables softmax & normalization computation without going to a separate floating-point unit

\[
\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}
\]

\[
\text{softmax}(\frac{QK^T}{\sqrt{d_k}}) = Z
\]

© 2023 Flex Logix Technologies, Inc.
Normalization is executed in BF16 due to its large dynamic range.

\[
\|W\|_1 = \sum_{i} |\omega_i|
\]

\[
\|W\|_2^2 = \sum_{i} \omega_i^2
\]

Add and Feed-forward Network (FFN) operators are similar those in CNNs.
InferX IP is linearly scalable (5nm, batch=1)

<table>
<thead>
<tr>
<th>Model</th>
<th>1280×1280</th>
<th>1024×1024</th>
<th>640×640</th>
<th>512×512</th>
<th>256×256</th>
<th>128×128</th>
<th>64×64</th>
<th>32×32</th>
</tr>
</thead>
<tbody>
<tr>
<td>DETR 2020 (1024×1024)</td>
<td>12 IPS</td>
<td>23 IPS</td>
<td>70 IPS</td>
<td>127 IPS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YOLOv5s (640×640)</td>
<td>75 IPS</td>
<td>175 IPS</td>
<td>396 IPS</td>
<td>850 IPS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YOLOv5L6 (1280×1280)</td>
<td>5 IPS</td>
<td>10 IPS</td>
<td>24 IPS</td>
<td>48 IPS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ResNet50 (1024×1024)</td>
<td>18 IPS</td>
<td>38 IPS</td>
<td>102 IPS</td>
<td>158 IPS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
How to deploy InferX models in your system?

Standard training and transfer learning workflow

- Application data
- Transfer learning
- FP32 model

> (Optional)
> - Refine model for int8

Flex Logix model developer workflow with InferX MDK

- Model conversion
- Edge model

InferX model developer kit

- Optimization
- Quantization
- Compilation

Your Inference Application

Your SoC with InferX accelerator
InferX compiler is available for evaluation

Model Formatting & Quantization – Ready & Under Test

Inference Runtime Engine – Ready & Under Test

GRAPH COMPILER – Ready & Under Test

OPERATOR COMPILER – Ready & Under Test

EFLX Compiler – Ready & In Production

© 2023 Flex Logix Technologies, Inc.
Come see us for more information!

- InferX IP provides flexibility to future-proof your AI solution, including state-of-the art CNN & transformers workloads

- Please come visit our booth for demos and brochures

- Please visit www.flex-logix.com for more information