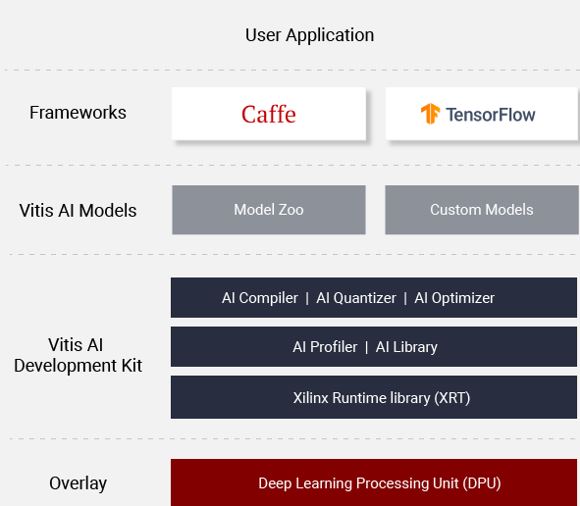
## 

## **Guide Book**

## Deep Neural Network Inference on FPGA using TensorFlow



**מבצעים**: אלון נמירובסקי ועמית שטאובר  
**מנחים**: אינה ריבקין, עוז שמואלי

אביב תשפ"א / אפריל 2021

תוכן עניינים

[מבוא ...........................................................................................................................3](#_Toc69473980)

[הקמת עמדה 4](#_Toc69473981)

[צריבת SD-Card 8](#_Toc69473982)

[Setup של כרטיס FPGA 9](#_Toc69473983)

[הקמת רשת, קומפילציה והרצת הרשת Partial Flow For Resent-50 : 13](#_Toc69473984)

[ שלב הקוונטיזציה 13](#_Toc69473985)

[ שלב הקומפילציה 15](#_Toc69473986)

[הקמת רשת, קומפילציה, אופטימיזציה וסיווג: Full Flow For Mnist 18](#_Toc69473987)

[ שלב הקוונטיזציה 19](#_Toc69473988)

[ שלב הקומפילציה 19](#_Toc69473989)

[ שלב הסיווג/הרצה על הboard 22](#_Toc69473990)

[ אופטימיזציה וביצועים 23](#_Toc69473991)

[טיפים, בעיות ופתרון בעיות 26](#_Toc69473992)

[תובנות כלליות 27](#_Toc69473993)

מבוא   
  
**רקע כללי**

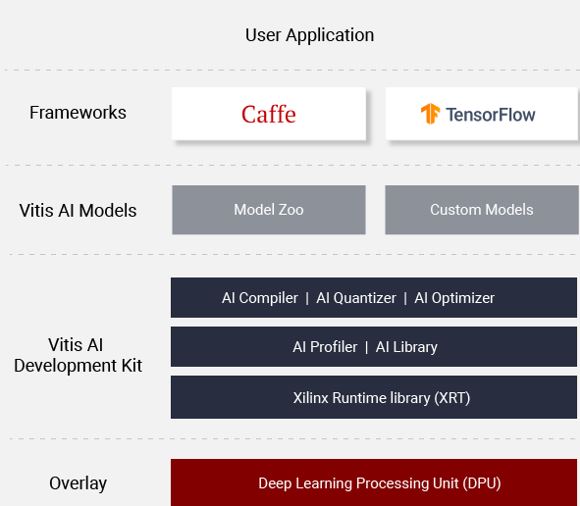
* תחום המערכות הלומדות המבוסס על רשתות נוירונים צובר תאוצה בשנים האחרונות, בעקבות ההצלחה של מערכות אלו בפתרון בעיות מורכבות.
* יישום של רשתות מתבצע ב-2 שלבים – שלב האימון ושלב הסיווג.
* שלב האימון הוא שלב מקדים המתבצע פעם אחת לכל רשת נתונה ולרוב מבוצע על מערכת בעלת עוצמת חישוב גבוהה כגון GPU. כיוון ששלב זה מבוצע פעם אחת לכל מערכת אין חשיבות עקרונית לזמן וכוח חישוב ששלב זה דורש.
* לעומת זאת בהרבה אפליקציות כגון רובוטים, מל"טים, רשתות חיישנים, טלפונים סלולאריים, רכבים אוטונומיים וכו' שלב הסיווג הוא שלב המתבצע בשטח ע"י רכיבים המוגבלים בכוח החישוב, בהספק, ולעתים גם יש דרישה לעבודה בזמן אמת ולמחיר זול.
* כיוון שאחד המאפיינים הבולטים של שלב הסיווג ברשתות נוירונים הוא מקביליות, מימוש שלב זה דווקא על חומרה כגון FPGA מאפשר להגיע לביצועים גבוהים מאוד תוך שמירה על צריכת הספק נמוכה ומחיר זול. כמו-כן, מימוש החומרה על FPGA נותן גמישות ויכולת התאמה של הרשתות עבור בעיות שונות.
* כדי לקדם מימוש רשתות נוירונים בחומרה על רכיבי FPGA יצרני רכיבים אלו פתחו סט כלים חדש שמאפשר מימוש שלב הסיווג ואופטימיזציה של רשת החל משלב ההגדרה עד שלב המימוש בפועל, תוך שימוש בכלים סטנדרטיים.

**מטרת הפרויקט**

לחקור ולהעריך סט כלים חדש **Vitis AI** עבור  DNN inference(סיווג ע"י רשת נוירונים) על רכיבי  FPGAשל חברת Xilinx.

בנית הרשת תעשה באמצעות TensorFlow

**דרישות הפרויקט:**

1. הקמת המערכת על מחשב ייעודי.
2. בחירה והרצה של מודל שאומן מראש תוך שימוש בכל כלי התוכנה של חברת Xilinx.
3. מימוש סביבת הבדיקה.
4. ביצוע שינויים והתאמות במודלים במטרה לבדוק את גמישותם ויכולתם של הכלים.

לדוגמא שינוי מבנה הרשת (model zoo vs Custom), טיפוסי המשתנים (FP vs Fixed point) וכו'.

1. השוואה בין המימושים השונים תוך בדיקה של דיוק, throughput, גודל ואתגרים במימוש ושימוש בכלים.
2. הסקת מסקנות ודיווח על מגבלות, יכולות ובעיות של הכלים השונים בהם נעשה שימוש.

# הקמת עמדה

חברת Xilinx מספקת לנו תשתית שלמה להרצת מערך הVitis-AI הדרוש להרצת מערכות בינה מלאכותית על הלוח ZCU104. נתאר כאן את תהליך הקמת התשתית על גבי המחשב האישי הכולל התקנת Virtual Machine ותוכנות שונות.

**סדר פעולות ראשוני:**

**הקמת VM על גבי המחשב האישי:**

הורדת קובץ ubuntu-18.04-desktop-amd64.iso ויצירת VM באמצעותVMware .

פירוט נוסף – חלק מהשלבים הראשונים בוצעו באמצעות VirtualBox ולא שונו לאחר שביצענו המרה לVirtualBox. **אין לנסות לעבוד עם VirtualBox כי אנחנו קיבלנו הודעת שגיאה שיש פקודות שלא נתמכות בVirtualBox.** התוכנה בה ביצענו שימוש: VMware player workstation.

הקצנו 4GB ראם לתוכנה ו-350GB להארד דיסק, 4 cpu.

**התקנת דוקר:**

לפי המדריך Program/References/PDFs\_Archive/1

תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי

הערה: ההרצה של הGPU העלתה שגיאה שכרגע נראית לא רלוונטית.

גרסת הדוקר:



פקודות להרצת הdocker:

./docker\_run.sh xilinx/vitis-ai-cpu:latest

conda activate vitis-ai-tensorflow

תקלה שנתקלנו בה (ופתרון כמפורט בתמונה):

תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי

תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי

**קישורים חשובים:**

Program/References/PDFs\_Archive/2

Program/References/PDFs\_Archive/3

Program/References/PDFs\_Archive/4

**התקנת VART:**

Program/References/PDFs\_Archive/5

הודעה לאחר סיום ההתקנה:

SDK has been successfully set up and is ready to be used.

Each time you wish to use the SDK in a new shell session, you need to source the environment setup script e.g.

$ . /home/projecta/petalinux\_sdk/environment-setup-aarch64-xilinx-linux

הpath:

~/petalinux\_sdk

הורדת qBitorrent:

sudo apt install qbittorrent

<https://itsfoss.com/best-torrent-ubuntu/>

**התקנת הPetalinux**

כחלק מהפרויקט בדקנו היתכנות של קימפול קוד הRTL של רכיב הDPU וצריבתו על הZCU-104. נעזרנו במדריך הבא:

Program/References/PDFs\_Archive/17

**הערה:** יש לשים לב שבעת ביצוע השלבים, אתם פועלים בהתאם לדרישות המקדימות בתחילת המדריך (גרסת petalinux, גרסת מכשיר וכו').

ראשית יש לבצע קימפול של קוד הdesign של חומרת ה-DPU בעזרת כלי Xilinx.

בהמשך, מורידים מפה את שני הקבצים:

<https://www.xilinx.com/support/download/index.html/content/xilinx/en/downloadNav/embedded-design-tools.html>

תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי

תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי

יש לפעול לפי הוראות ההדרכה. נציין כי בכמה מקרים ביצענו התאמות. נפרט:

פקודות בהן נעזרנו (בוצעו התאמות לכרטיס שלנו בהתאם לסוגו):

petalinux-config --get-hw-description=~/Vitis-AI/BootExp/VivadoOutput/prj –-silentconfig

petalinux-build

נדרשנו לשנות את הקונפיגורציה בקובץ הבא:

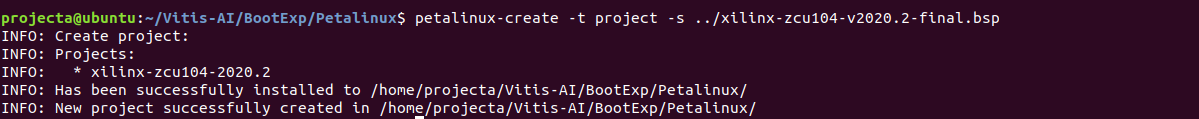
project-spec/meta-user/conf/petalinuxbsp.conf

לערכים הבאים:

תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי

באופן כללי לאחר יצירת הפרויקט:



**מקרים ותגובות:**

במקרה של השגיאה הבאה:

petalinux-create: command not found

יש להסתכל בדיון הבא:

Program/References/PDFs\_Archive/18

במהלך הבניה נתקלנו בשגיאה, פתרנו את השגיאה בעזרת הקישור הבא:

Program/References/PDFs\_Archive/19

העברה לכרטיס זיכרון יש לבצע על פי המדריך הבא, עמוד 61:

Program/References/PDFs\_Archive/20

# צריבת SD-Card

צריבת הכרטיס הינה פעולה הכרחית לטעינה קוד התכן אל הלוח. אנו נעזרנו בתוכנת צריבה אשר נראית etcher ובקובץ מוכן שניתן על ידי Xilinx. יש להיעזר באתחול במדריך הבא:

Program/References/PDFs\_Archive/14

הערה – יש לבצע הרצה של התוכנה etcher כמנהל.

# Setup של כרטיס FPGA

הסבר טכני על חיבור הכבלים:

יש לחבר את הכבלים בצורה הבאה (יש להיעזר בתמונה הראשונה):

הכבל העליון מתחבר לחשמל, הכבל התחתון מתחבר לרשת והכבל הימני מתחבר למחשב. לאחר החיבור יש לבצע הדלקה של הלוח ולחכות כדקה.

תמונה כללית של הלוח לאחר הדלקה (ZCU104):

תמונה שמכילה אלקטרוניקה, מעגל חשמלי

התיאור נוצר באופן אוטומטי

יש לסדר את המתגים כמתואר בתמונה, כאשר כחול מעיד על למעלה ואדום על למטה.

תמונה שמכילה אלקטרוניקה, מעגל חשמלי

התיאור נוצר באופן אוטומטי

יש לוודא כי הנורה המסומנת דלוקה:

תמונה שמכילה אלקטרוניקה, מעגל חשמלי

התיאור נוצר באופן אוטומטי

כעת נחבר את הכרטיס למחשב האישי באמצעות כבל micro-usc. חיבור זה יוצר שלושה חיבורי COM: יש לבחור את הCOM בעל הערך המינימלי (לחפש במנהל ההתקנים).

כדי להתחבר לכרטיס ניעזר בתוכנת Putty שמאפשרת חיבור במגוון פרוטוקולים.

בתוכנה נבחר בחיבור טורי (סריאלי). המסך המתקבל בחיבור הטורי:

תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטייש להגדיר baud rate: 115200 bps

לאחר ההתחברות אפשר לרשום ifconfig ולהתחבר ללוח עם כבל רשת בssh:

(שם משתמש וסיסמא – (root

תמונה שמכילה צילום מסך, צג, מחשב, מחשב נישא

התיאור נוצר באופן אוטומטי

**ניתן להוריד תוכנה בשם – WinSCP – תוכנה אשר משמשת להעברת קבצים מהמחשב הלוקלי למחשב המרוחק עם תצוגה למשתמש.**

לחלופין, ניתן להשתמש בפקודות הבאות:

פקודה להעביר תיקייה מהמחשב הלוקאלי למחשב המרוחק:

scp -r c:\move2fpga\ root@192.168.1.26:~/Vitis-AI/VART/samples/images

במקרה שמתקבלת תשובה לתמונה אחת בלבד יש לפעול לפי האתר הבא:

Program/References/PDFs\_Archive/15

פקודת להעברת קובץ:

scp c:\move2fpga\dpu\_resnet50\_tf\_0.elf root@192.168.1.26:~/Vitis-AI/VART/samples/images

תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטילהלן דוגמה להרצה מוצלחת בFPGA:

קבצי הקוד:

Program/References/PDFs\_Archive/16

root@xilinx-zcu104-2020\_1:~/Vitis-AI/VART/samples/amitalon\_resnet50# ./resnet50 dpu\_resnet50\_tf\_0.elf

WARNING: Logging before InitGoogleLogging() is written to STDERR

I0701 12:39:33.563982 1134 main.cc:288] create running for subgraph: resnet50\_tf\_0

# הקמת רשת, קומפילציה והרצת הרשת Partial Flow For Resent-50 :

## שלב הקוונטיזציה

הדרכה כללית מאתר Xilinx:

Program/References/PDFs\_Archive/6

הורדת validation set (שימושי באופן כללי בשלב זה):

Program/References/PDFs\_Archive/7

הורדת מודל:

Program/References/PDFs\_Archive/8

בחרנו במודל הבא:



זה המודל שהונחנו להשתמש בהחלטה משותפת עם המנחים.

נשים לב כי כעת אנו פועלים על גבי סביבת הUNIX (אלא אם נאמר אחרת).

יצרנו תיקייה בשם imagenet\_images עם אלף תמונות מהvalidation set. יצרנו קובץ פייתון כפי שמוגדר במדריך ויצרנו קובץ imagenet\_calib.txt. הוא מכיל אלף שורות מן הצורה:

תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי

פקודה להעתקת אלף הקבצים הראשונים:

cp `ls | head -1000` ../../Vitis-AI/projecta/imagenet\_images

הפקודה שהרצנו לבצע את תהליך הקוונטיזציה:

vai\_q\_tensorflow quantize --input\_frozen\_graph resnet\_v1\_50\_inference.pb --input\_nodes input --input\_shapes ?,224,224,3 --output\_nodes resnet\_v1\_50/predictions/Reshape\_1 --input\_fn input\_fn.calib\_input --method 1 --gpu 0 --calib\_iter 20 --output\_dir ./quantize\_results

ההודעה בסיום ההרצה:

תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי

שתי תקלות שחשוב להכיר ולהבין:

1. תקלת core dumped – הסיבה לתקלה זו נבעה אצלנו כתוצאה מהקצאה של מקום בכונן שאין בו מקום, לכן חשוב להקצות מקום בכונן עם הרבה מקום, במקרה של המחשב הנוכחי זה היה כונן D.
2. **תקלת FMA Instructions not supported – תקלה אשר התקבלה כתוצאה מעבודה עם virtual box, חשוב לעבוד בVMARE אחרת לא ניתן לעבוד.**

## שלב הקומפילציה

מדריך כללי מאתר Xilinx:

Program/References/PDFs\_Archive/9

הפקודה לביצוע קומפילציה(הרצנו אותה מתוך אותה תיקייה שהרצנו את הפקודה של הquantization):

vai\_c\_tensorflow --frozen\_pb ./quantize\_results/deploy\_model.pb --arch /opt/vitis\_ai/compiler/arch/dpuv2/ZCU104/ZCU104.json --output\_dir model --net\_name resnet50\_tf

*הערה: הפקודה באתר איננה מדויקת, הפקודה שרשמנו כאן היא המדויקת.*

הפלט שקיבלנו:

projecta@ubuntu:/workspace/projecta$ vai\_c\_tensorflow --frozen\_pb ./quantize\_results/deploy\_model.pb --arch /opt/vitis\_ai/compiler/arch/dpuv2/ZCU104/ZCU104.json --output\_dir model --net\_name resnet50\_tf

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* VITIS\_AI Compilation - Xilinx Inc.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

[VAI\_C][Warning] layer [resnet\_v1\_50\_SpatialSqueeze] (type: Squeeze) is not supported in DPU, deploy it in CPU instead.

[VAI\_C][Warning] layer [resnet\_v1\_50\_predictions\_Softmax] (type: Softmax) is not supported in DPU, deploy it in CPU instead.

Kernel topology "resnet50\_tf\_kernel\_graph.jpg" for network "resnet50\_tf"

kernel list info for network "resnet50\_tf"

Kernel ID : Name

0 : resnet50\_tf\_0

1 : resnet50\_tf\_1

Kernel Name : resnet50\_tf\_0

--------------------------------------------------------------------------------

Kernel Type : DPUKernel

Code Size : 0.59MB

Param Size : 24.35MB

Workload MACs : 6964.51MOPS

IO Memory Space : 2.25MB

Mean Value : 0, 0, 0,

Total Tensor Count : 59

Boundary Input Tensor(s) (H\*W\*C)

input:0(0) : 224\*224\*3

Boundary Output Tensor(s) (H\*W\*C)

resnet\_v1\_50\_logits\_Conv2D:0(0) : 1\*1\*1000

Total Node Count : 58

Input Node(s) (H\*W\*C)

resnet\_v1\_50\_conv1\_Conv2D(0) : 224\*224\*3

Output Node(s) (H\*W\*C)

resnet\_v1\_50\_logits\_Conv2D(0) : 1\*1\*1000

Kernel Name : resnet50\_tf\_1

--------------------------------------------------------------------------------

Kernel Type : CPUKernel

Boundary Input Tensor(s) (H\*W\*C)

resnet\_v1\_50\_SpatialSqueeze:0(0) : 1\*1\*1000

Boundary Output Tensor(s) (H\*W\*C)

resnet\_v1\_50\_predictions\_Softmax:0(0) : 1\*1\*1000

Input Node(s) (H\*W\*C)

resnet\_v1\_50\_SpatialSqueeze : 1\*1\*1000

Output Node(s) (H\*W\*C)

resnet\_v1\_50\_predictions\_Softmax : 1\*1\*1000

נציין כי קיבלנו תיקיית מודל ולהלן הפירוט שלה:

תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי

# הקמת רשת, קומפילציה, אופטימיזציה וסיווג: Full Flow For Mnist

**בשלב זה אנו יוצרים מודל בצורה מלאה, החל מכתיבתו בTF ועד הרצת שלב הסיווג על גבי הכרטיס.**

התבססנו על המדריך הבא:

Program/References/PDFs\_Archive/10

מיקום בגיט: References/Vitis-AI-Tutorials-MNIST-Classification-TensorFlow.zip

מצוין באדום הנקודות בהן השתמשנו במדריך.

ראשית, כתבנו קוד בcolab בשפת python בעזרת ספריות tensorflow,keras.

הקוד הסופי מצורף (נעזרנו במודל של המדריך):

מיקום בגיט: References/Start.ipynb

לאחר מכן הרצנו קוד על הטנזור וקיבלנו גרף, לאחר מכן ביצענו "הקפאה" של הגרף על מנת שנוכל לעבוד איתו במכונה הוירטואלית.

Program/References/PDFs\_Archive/11

הערות:

* לא ניתן להעזר בגרסה tensorflow 2 ומעלה (דיון בפורום בנוגע להתאמה בין גרסאות של טנזור):

Program/References/PDFs\_Archive/12

* לא ניתן לעבוד עם שכבות FC, יש לעבוד עם שכבות קונבולוציה.

Program/References/PDFs\_Archive/13

* נוצרה בעיה כאשר נעזרנו בשכבת dense-softmax, המרנו את השכבה לשכבת סיגמואיד והגענו לתוצאה מעולה.

לאחר מכן יצרנו קבצי קליברציה:

מיקום בגיט: References/trainimages\_calib\_original.zip

סקריפטים ליצירת התמונות:

מיקום בגיט: References/4\_quant.zip

## שלב הקוונטיזציה

בשלב הקוונטיזציה נעזרנו בסקריפט (הועתק מהמדריך): image\_input\_fn.py

תוך כדי שהגדרנו בסקריפט מחדש את השם של ה) Input nodeנלקח מהלוגים של הקפאת המודל בקולאב). השורה מסומנת בכתום.

תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי

פקודת להרצת תהליך הקוונטיזציה:

vai\_q\_tensorflow quantize --input\_frozen\_graph frozen\_graph.pb --input\_nodes conv2d\_input\_1 --input\_shapes ?,28,28,1 --output\_nodes flatten\_1/Reshape --input\_fn image\_input\_fn.calib\_input --method 1 --gpu 0 --calib\_iter 10 --output\_dir ./quantize\_results

## שלב הקומפילציה

פקודה להרצת תהליך הקומפילציה:

vai\_c\_tensorflow --frozen\_pb ./quantize\_results/deploy\_model.pb --arch /opt/vitis\_ai/compiler/arch/dpuv2/ZCU104/ZCU104.json --output\_dir model --net\_name mnist\_tf

הפלט לאחר קומפילציה:

INFO: Output:

quantize\_eval\_model: ./quantize\_results/quantize\_eval\_model.pb

deploy\_model: ./quantize\_results/deploy\_model.pb

(vitis-ai-tensorflow) projecta@ubuntu:/workspace/projecta-mnist$ vai\_c\_tensorflow --frozen\_pb ./quantize\_results/deploy\_model.pb --arch /opt/vitis\_ai/compiler/arch/dpuv2/ZCU104/ZCU104.json --output\_dir model --net\_name mnist\_tf

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* VITIS\_AI Compilation - Xilinx Inc.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

[VAI\_C][Warning] layer [activation\_1\_Sigmoid] (type: Sigmoid) is not supported in DPU, deploy it in CPU instead.

Kernel topology "mnist\_tf\_kernel\_graph.jpg" for network "mnist\_tf"

kernel list info for network "mnist\_tf"

Kernel ID : Name

0 : mnist\_tf\_0

1 : mnist\_tf\_1

Kernel Name : mnist\_tf\_0

--------------------------------------------------------------------------------

Kernel Type : DPUKernel

Code Size : 3.46KB

Param Size : 0.08MB

Workload MACs : 3.08MOPS

IO Memory Space : 5.86KB

Mean Value : 0, 0, 0,

Total Tensor Count : 5

Boundary Input Tensor(s) (H\*W\*C)

conv2d\_input\_1:0(0) : 28\*28\*1

Boundary Output Tensor(s) (H\*W\*C)

conv2d\_3\_1\_Conv2D:0(0) : 1\*1\*10

Total Node Count : 4

Input Node(s) (H\*W\*C)

conv2d\_4\_Conv2D(0) : 28\*28\*1

Output Node(s) (H\*W\*C)

conv2d\_3\_1\_Conv2D(0) : 1\*1\*10

Kernel Name : mnist\_tf\_1

--------------------------------------------------------------------------------

Kernel Type : CPUKernel

Boundary Input Tensor(s) (H\*W\*C)

activation\_1\_Sigmoid:0(0) : 1\*1\*10

Boundary Output Tensor(s) (H\*W\*C)

activation\_1\_Sigmoid:0(0) : 1\*1\*10

Input Node(s) (H\*W\*C)

activation\_1\_Sigmoid : 1\*1\*10

Output Node(s) (H\*W\*C)

activation\_1\_Sigmoid : 1\*1\*10

## שלב הסיווג/הרצה על הboard

יצרנו תיקייה חדשה על גבי הרכיב.

העברנו את הקובץ הבא (מועתק מהמדריך):

מיקום בגיט: References/ app\_mt.py

בנוסף העברנו תמונות לסיווג ואת קובץ הELF של הקומפילציה.

תמונה שמכילה שולחן

התיאור נוצר באופן אוטומטי

כעת ביצענו את הרצת הtest set:

python3 app\_mt.py -m dpu\_mnist\_tf\_0.elf

לשינוי threads להוסיף -t

תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי

בוצע על מס' threads שונה, 1 ו-6.

## אופטימיזציה וביצועים

בדיקת ביצועים לאחר שינוי פרמטרים**:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **מבוצע על ידי** | **שלב ארכיטקטוני** | **פרמטר שינוי** | **תיקייה** | **ביצועים** |
| Colab | - | - | - | **Accuracy 98.25%** |
| Board | - | - | **Modifications/Quantization/**  **Original** | **Accuracy 98.19%** |
| Board | Quantization – קוונטיזציה | --weight\_bit  Bit width for quantized weight and bias.  הערך הדיפולטי: 8.  שינוי שביצענו: מעבר ל16. | **Modifications/Quantization/**  **Weight\_bit\_16**  INT8 is only supported for DPU deployment. You can use 16 and 32 bits for simulation, but you won't be able to deploy this on DPU hw. | **ביצועים ירודים כתוצאה מחוסר תמיכה של המערכת** |
| Board | Quantization – קוונטיזציה | --activation\_bit  Bit width for quantized activation.  הערך הדיפולטי: 8.  שינוי שביצענו: מעבר ל16. | **Modifications/Quantization/**  **Activation\_bit\_16**  INT8 is only supported for DPU deployment. You can use 16 and 32 bits for simulation, but you won't be able to deploy this on DPU hw. | **ביצועים ירודים כתוצאה מחוסר תמיכה של המערכת** |
| Board | Quantization – קוונטיזציה | --calib\_iter  The iterations of calibration. Total number of images for calibration = calib\_iter \* batch\_size. | **Modifications/Quantization/**  **Calib\_Iter\_100** | **Accuracy 98.22%** |
| Board | Quantization – קוונטיזציה | שתי שיטות על פיהן ניתן לקוונטט. שיטה 0 יוצרת טווח רחב של ערכים קוונטיזציה ושיטה 1 מאפשרת טווח צר יותר. | **Modifications/Quantization/**  **Method\_0** | **Accuracy 98.33%** |
| Board | Inference – סיווג | --threads  Number of threads. Default is 1 | **-** | **Accuracy 98.19%** |

ללא שינויים:

תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי

Method 0

תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי

Calib:

תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי

Activation:

תמונה שמכילה טקסט, צילום מסך, צג, מסך

התיאור נוצר באופן אוטומטי

Weight:

תמונה שמכילה טקסט, צילום מסך, צג, מסך

התיאור נוצר באופן אוטומטי

Threads:

תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי

# טיפים, בעיות ופתרון בעיות

|  |  |  |
| --- | --- | --- |
| **הבעיה** | **מה היא אומרת** | **איך ניתן לתקן** |
| שגיאת core dumped | בתחילה ניסינו ליצור מכונה עם גודל אחסון של כמה עשרות ג'יגה, הגענו במהירות לאחסון מלא. | נתנו למכונה הוירטואלית מרחב אחסון גדול בהרבה של 350 ג'יגה. |
| Virtual box לא תומך בפקודות FMA |  | יצרנו מכונה וירטואלית בVMARE. |
| המכשור לא מסוגל לעבוד עם גרסה tensorflow 2 ומעלה |  | ניתן לבצע המרה בתחילת הקוד בcolab לגרסה ישנה יותר כפי שהראנו בקוד. |
| המכשור לא מסוגל לעבוד עם שכבות FC אחת אחרי השניה. | באופן כללי מתבצעת המרה משכבת FC לשכבת קונבולוציה. כאשר משתמשים בשתי שכבות כאלו אחת אחרי השניה נוצר מצב בו גודל הכניסה לשכבת הקונבולוציה גדול מהגודל המקסימלי ((16x16. | מעבר לארכיטקטורה של שכבות קונבולוציה בלבד כפי שהראנו בקוד שהכנו. |
| שכבת dense softmax נכשלת בשלב הקוונטיזציה | הקוונטייזר לא יודע איך לעבוד עם סוג המשתנה שיוצא משכבה זו (int32) | עברנו לשכבת אקטיביזציה אחרת בשם סיגמואיד. |
| שכבת סיגמואיד לא עובדת על הDPU | מכיוון שהמשתנה היוצא משכבה זו הינו int16 הDPU לא יכול להריץ אותו ולכן הוא מורץ על הCPU. | זוהי אינה בעיה. הDPU לא יכול להריץ את כל סוגי השכבות. אין בעיה שחלק מהקוד יורץ על הCPU. |
| הDPU לא תומך ברוחב ביט שאינו 8 |  | הפתרון הינו כמובן לעבוד ברוחב ביט זה בלבד. |

# תובנות כלליות

**מסקנות שעלו בעקבות תהליך הביצועים והאופטימיזציה**

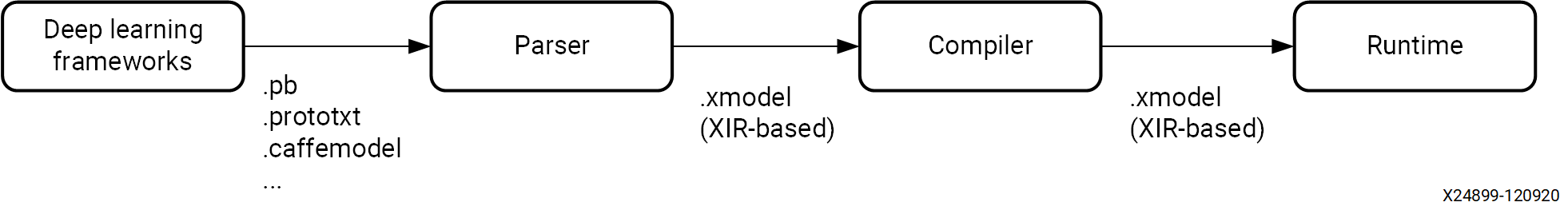
* ביצועים אופטימליים מבחינת סיווג – יש לבצע קוונטיזציה על פי שיטה 0 אשר עובדת עם טווח רחב של ערכים.
* ביצועים אופטימליים מבחינת זמנים – במקרה של עבודה עם הלוח יש לבצע חלוקה למספר חוטים אשר נותן תוצאה אופטימלית. במקרה שלנו זה התקבל עבור 8 חוטים.
* יש לעבוד עם רוחב ביט של 8.
* יש לבצע שינויים בפרמטרים בשלב הקוונטיזציה, בשלב הקומפילציה לא נמצאו פרמטרים רלוונטיים.

**הרצת שכבות על גבי הDPU ועל גבי הCPU**

כפי שניתן לראות בתוצאות הקומפילציה, ישנן שכבות אשר מיועדות לריצה על גבי המעבד (ARM) וחלקן מיועדות לריצה על גבי הDPU. נרצה טיפה להסביר על המניע וכיצד המנגנון הפנימי עובד.

התשתיות בפרויקט מתבססות על פרוטוקול ייצוג XIR - Xilinx Intermediate Representation. זהו פרוטוקול מבית XILINIX של ייצוג גרף לאלגוריתמי AI. הוא נועד להצגה איכותית של רשתות נוירונים ולאחר מכן גם לפרישה נוחה ומהירה על גבי פלטפורמות חומרה שונות (במקרה שלנו DPU על גבי הFPGA).

באופן כללי התהליך נראה כך:



כאשר המודל (רשת הנוירונים הסופית) מכילה פעולות שהDPU אינו יכול לבצע ברמת החומרה, נוצרים Sub-graph לשכבות הספציפיות הללו והם ממופים לביצוע על גבי ה-CPU.   
המידע לגבי איזה שכבות ניתן להריץ על גבי הCPU ואילו לא, מוחלט בשלב הקומפילציה על ידי הקומפיילר. בסופו של דבר, כאשר נעזרים ב- Vitis AI unified high-level APIהמודל שלנו ינסה לרוץ על ה-DPU וכאשר לא יוכל יבצע חלק מהפעולות על גבי הCPU. באופן כללי, הData Flow הוא מן הצורה: ...CPU--DDR--DPU--DDR...

**קימפול ה-RTL לכרטיס (בדיקת היתכנות)**  
  
מתוך המחקר שביצענו, אנו ראינו כי אכן יש אפשרות להצליח לקמפל את קוד הDPU ולהורידו לכרטיס. עם זאת, מדובר בתהליך מורכב, הכולל שלבים רבים של התאמות וכולל בעיות רבות של סנכרון בין גרסאות שונות של תוכנות שונות ושימוש בקבצי תלויות מסוימים. כפי שניתן לראות, המדריך המקורי (מדריך 17) יועד לכרטיס אחר- ZCU-102. אנו מאמינים שבמידה וניתן יהיה להשיג כרטיס זה, עבודת הקימפול והצריבה לכרטיס לאחר מודיפיקציות שונות תתבצע בקלות יותר (אך עדיין מורכבת לאור התלויות המשתנות).

**אחרית דבר**

* במקרים מסוימים, אכן הכרטיס בשילוב סביבת Vitis-AI מצליחים להשיג אחוזי דיוק טובים יותר מאשר מאשר סיווג על גבי מחשב ממוצע וזאת הודות לחומרה הייעודית שלו.
* חברת Xilinx הצליחה להעניק בצורה יחסית פשוטה יכולות של בינה מלאכותית לכל החפץ בכך - גם כאלה בעלי מעט מאוד ניסיון בעולם זה.
* יצירת מודל חדש הינו תהליך מורכב מאוד. לאורך התהליך מהרגע שבו יצרנו מודל חדש ועד להרצתו על גבי הכרטיס, נדרשו מספר רב של התאמות לאורך הדרך.
* לצערנו המידע על תשתיות Xilinx ברשת הינו מוגבל ולעיתים קרובות קשה מאוד להתמודד עם בעיות שצפות במהלך תהליך הפיתוח של מודל חדש.