להלן סיכום מפגש רביעי וחומרים נלווים.

**כללי:**

תחילה אני רוצה לומר תודה רבה על ההזדמנות שקיבלתי והאמון שנתנו בי להדרכת חלק זה של הקורס תחילה לכל שרשרת הניהול האלביטאית שלי (אייל, לאוניד ועודד) ולמנהלי ומובילי המאגד נעמי, אורי ונופר.

תודה רבה ענקית לעינב המקסימה שלאורך כל הדרך הייתה שם לצידי לכל צורך שלי גם שהיא זיהתה משהו וביוזמה שלה הציעה את תמיכתה האדיבה והמקצועית.

אני רוצה לומר לכם הלומדים תודה רבה על החוויה הסופר חיובית שעברתי, על שיתוף הפעולה, המשובים, האינטראקציה.

אני שמחתי לקבל אתמול את המשובים הכנים שבהחלט ישפרו אותי לכל צורך עתידי שיעלה במסגרת כזו או אחרת.

לי היה חשוב מאוד להבין עד כמה בעיות התשתית השפיעו והעיבו על המשוב שבאמת היה ממוקד לאיך אני יכול להשתפר, איך החלק של הקורס הזה יכול להיות מועבר נכון ואפקטיבי יותר, והרושם שאני לוקח איתי זה שהיה לזה השפעה גדולה, כלומר בעולם אופטימאלי ללא בעיות תשתית שהיה מאפשר תרגול רב כפי שתכננתי מראש, התחושות של רוב הלומדים היו חיוביות הרבה יותר.

לכן במקביל לשיפור התשתיות, באופן אישי אתמקד במשובים שהיו מכוונים להדרכתי נטו:

* נושאים  - מה באמת נדרש לחלק זה של הקורס, חשיבה מחדש על הסילבוס וארגונו
* אספקת חומר נלווה, דוגמאות קוד, מצגות וכדומה מבעוד מועד טרם שימוש בהם בזמן ההדרכה
* תיעוד נרחב יותר שנלווה לקוד המסופק

אשמח שתשלחו לי דעתכם אם פספסתי משהו.

**פרויקט:**

אנו נכנסים עכשיו לחודש של פרויקט ומועד ההגשה הוא עד לסוף החודש.

כרגע יש לנו עוד בעיות טכניות עם המכונות הווירטואליות וברגע שמשהו ישתנה נעדכן מיידית.

אני מזכיר שלאור הבעיות יש אפשרות להגיש שלישיות, יש אפשרות להגיש ב-C++\Python, יש אפשרות להגיש גם במתכונת של Static review ולאו דווקא משהו שרץ ועובד וגם אם מישהו מעוניין אז בשיחת Teams תוך כדי הצגה.

באחריות שלכם לקחת את הפרויקט כאמצעי לכל מטרה שתרצו להשיג באמצעותו.

כפי שהבטחתי אני אשמח לסייע בכל נושא שתמצאו לנכון **גם במסגרת שמעבר לScope של הקורס** ויש לכם את האמצעים להשיג אותי.

**המפגש:**

התחלק ל-3 חלקים עיקריים (לא מוצג לפי הסדר הכרונולוגי:(

1. משוב כללי ושיחה פתוחה על הקורס
2. סקירה טכנית מלאה מחדש של פתרון הפרויקט המסכם מהיבט של DL deployment, אותו הפצתי בסוף המפגש הקודם.
3. הצגת יכולת TensorRT quantization Int8 accuracy באמצעות מנגנון ה-Calibration.
4. הצגת של OpenCV DNN module באמצעות Yolov5 inference

להלן פרטים טכניים נוספים.

·         Quantization:

o        בגדול יש שני סוגים:

§         PTQ – Post Training Quantization

§         QAT – Quantization Awareness Training

o        כל התמיכה של ה-TRT ב-Int8 מתוארת כאן:

§         <https://docs.nvidia.com/deeplearning/tensorrt/developer-guide/index.html#working-with-int8>

§         במסגרת הקורס אנו לומדים על ה-PTQ שה-TRT מספק באמצעות שיטת ה-Calibration:

·         <https://docs.nvidia.com/deeplearning/tensorrt/developer-guide/index.html#enable_int8_c>

§         ניתן להשתמש ב-TRT PTQ גם באמצעות הכלי Torch-TensorRT :

<https://pytorch.org/TensorRT/tutorials/ptq.html>

o        לידע כללי, גם ה-OnnxRuntime מספק מימוש של PTQ משלו וניתן ללמוד על האופן בו ניתן להשתמש בלינק זה:

<https://onnxruntime.ai/docs/performance/quantization.html#onnx-quantization-representation-format>

·         OpenCV DNN:

o        <https://github.com/spmallick/learnopencv>

o        <https://github.com/spmallick/learnopencv/tree/master/Object-Detection-using-YOLOv5-and-OpenCV-DNN-in-CPP-and-Python>

§         Convert\_PyTorch\_models.ipynb – כדי לבחור איזו גרסת yolov5 רוצים להפעיל ולהמיר ל-onnx

§         ראו קובץ מצורף  - OpenCV\_Playground.zip.  
       הקובץ לא מכיל את המודל yoloV5s.onnx, כדי להריץ את הפרויקט יהיה עליכם לייצר אותו עצמאית כפי שהצגתי במהלך המפגש באמצעות הקובץ export.py שנמצא תחת התיקייה yolov5

·         לינקים מעניינים נוספים:

o        Accelerating GPU Applications with NVIDIA Math Libraries - <https://developer.nvidia.com/blog/accelerating-gpu-applications-with-nvidia-math-libraries/>

Demystifying GPU Architectures For Deep Learning - <https://learnopencv.com/demystifying-gpu-architectures-for-deep-learning/?ck_subscriber_id=272192348>