אוניברסיטת תל אביב

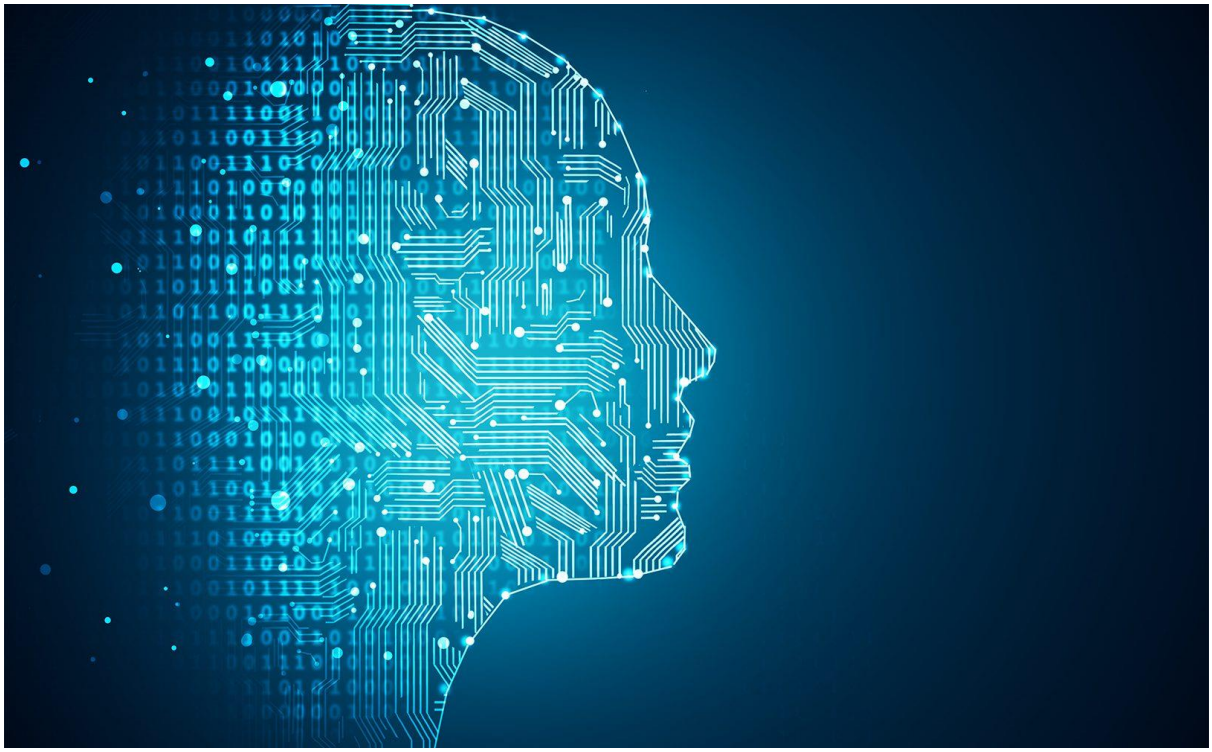
הפקולטה להנדסה

המחלקה להנדסת תעשייה וניהול

מבוא ללמידה עמוקה

מאמר מטלת חלק ב' – קבוצה 11

קישור למטלה בGitHub



מוגש לידי:

ד"ר אייל קולמן

מר שגיב פרסטר

מגישים – קבוצה 11:

אירנה אליושין 328814470

גלעד יקואל 206035222

יובל זיו 207900283

תקציר

המאמר עוסק בבעיית קלסיפיקציית ביקורות לסרטים, הלקוח מסט נתונים שהתקבל מIMDB. סט נתונים זה כולל אלפי רשומות מתויגות של ביקורות חיוביות ושליליות לסרטים שונים. בשימוש בסט נתונים זה, בוצע מחקר השוואה בין ביצועי שני מודלים קיימים לעיבוד שפה טבעית: Bert וGPT2.

כמו כן נבחנו הביצועים של מודל Bert לאחר יישום שתי שיטות כיווץ שונות: Pruning, Knowledge Distillation (KD). נבחנה טיב הקלסיפיקציה של המודל המכווץ בעקבות זאת, ובוצעה השוואת ביצועים.

בסופו של דבר, נראה כי המודל הטוב ביותר הינו מודל הBERT המכווץ בשיטת KD אשר מביא לרמת דיוק גבוהה בבחינת התוצאה לצד הצגת שגיאה נמוכה בסט המבחן. מבין מודלי הBERT והGPT2 לא נראה יתרון חד משמעי של אחד המודלים על האחר במדדי השגיאה והדיוק.

עוד נראה כי מודל הBERT המקורי מציג שגיאה המטפסת לאורך המבחן, וכי יישום שיטת כיווץ הPruning עבורו מעצימה את קצב שגיאה זה.

המודלים המכווצים בשיטת הPruning מציגים overfitting כאשר תוצאותיהם נבחנות מול תוצאות סט האימון.

1. הקדמה

עיבוד השפה הטבעית (NLP) הינו מושג המתייחס לאפשר למכונות (מחשבים) להבין ולייצר שפה אנושית. תחום זה קיים עשרות שנים, אך תפס תאוצה באחרונה בזכות עלייה בכוח המחשוב ועליית כמות הנתונים ברחבי הרשת.

קיימות מספר מטלות אשר הוכיחו את עצמן בתחום המודלים המאומנים, ביניהם תרגום, זיהוי פרט בשפה (תאריך, שם, עיר ועוד), סיכום טקסטים שונים, זיהוי דיבור וניתוח רגשי (כלומר, הבנה מתוך הנקרא האם מדובר בטקסט חיובי או שלילי לדוגמה).

על אף שמאמר זה עוסק במטלה הדומה במשמעותה לניתוח רגשי שמוזכר לעיל, היא לא זהה לחלוטין. אוסף הנתונים שנלקח לבדיקה הינו אוסף מתויג מהאתר IMDB (אתר פופולרי לאיסוף והצגת ביקורות על סרטים וסדרות ברחבי העולם) כאשר 0 מצביע על ביקורת שלילית ו-1 מצביע על ביקורת חיובית. המטרה הינה ביצוע קלסיפיקציה בין שני סוגי הביקורות בצורה המיטבית (מזעור הטעויות בסיווג).

לצורך כך, הוחלט לבחון שני מודלים מאומנים אשר נלקחו מהספרייה המשותפת hugging face transformers : Bert ו-GPT2.

מודלים אלו הינם מודלים מפורסמים, ואלו שנלקחו הינם מודלים מאומנים מראש. מודלים אלו קלים ליישום, וכיוון שכבר אומנו בעבר – ניתן לעשות בהם שימוש לקבלת תוצאות טובות לסיווג הבעיה הנקודתית עמה אנו מתמודדים, והם מאפשרים יישום זה בזמן אימון קצר ביחס למודלים הנבנים מאפס לצד דיוק גבוה בסיווג.

2. הצגת המודלים

המודלים הנבחנים במאמר זה הינם מודלים ללמידה עמוקה המשתמשים בשיטת Transformers : מודלים אשר במקום להתקדם על פני שכבות הרשת בצורה סדרתית (כמו מודלים מסוג RNN למשל), הם מתקדמים בצורה מקבילית. הארכיטקטורה של מודלים אלו משתמשת בטכניקת ה"attention" המאפשרת להם להתייחס באופן סלקטיבי לחלק מהקלט בלבד. העיבוד נעשה בו זמנית לאלמנטים שונים בשכבת הקלט (מילים, חלקי משפט) בצורה מקבילית, וכך המודל מסוגל לעבד את כל המילים בשכבה זו ולמדל את השפעתן אחת על השנייה ללא קשר למרחק ביניהן (״מרחק רצף״). כמו כן, העיבוד נעשה מהר הרבה יותר, ומאפשר ניצול מיטבי של המעבד לעומת השיטות הישנות של עיבוד טורי.

עם זאת, מאחר ומודלים אלו גדולים מאוד ומכילים סט פרמטרים רחב (אשר עשוי להגיע לממדי מאות מיליונים ואף מיליארדים של פרמטרים), קיימות שיטות כיווץ למודלים אלו המסננות בצורה חלקית את הפרמטרים במודל מבלי לפגוע יותר מדי בביצועיו (ועל כן מדובר בכיווץ, זהו כיווץ פיזי של נפח המודל).

במחקר זה נבחנו שני מודלים המאופיינים בשיטת הTransformers- מודל GPT2 ומודל BERT. מודלים אלו נלקחו לאחר אימון מראש (pre-trained models) וביצועיהם הושוו על סט הנתונים.

כמו כן, נבחנו הביצועים של אחד המודלים (מודל BERT) לאחר יישום שתי שיטות כיווץ שונות : Pruning וDistillation. הביצועים של מודל הBERT לאחר יישום שיטות הכיווץ הושוו לאלו של המודלים המקוריים (כהשוואה כוללת) ונבחנו תוצאות הסיווג של כל שיטה.

2.1. מודל GPT2

מודל זה הוצג לראשונה ב2019 במאמר "Language Models are Unsupervised Multitask Learners". המודל אומן על מאגר של כשמונה מיליון דפי אינטרנט (במשקל כולל של כ-40GB), והוא מסוגל לכתוב טקסטים בעצמו. המודל אומן לבצע משימות של הבנת הנקרא כגון סיכום טקסטים ותרגום מענה על שאלות, בנוסף לביצוע סיווג מידע לפי הנוסח שלו (כפי שאנו מציגים במאמר זה).

המודל שמשתמש בטוקנייזר (Tokenizer, מערכת המשמשת לקידוד המילים בטקסט עבור המודל) הנקרא BPE – Byte Pair Enconding אשר משלב בין גישת הצגת כל מילה כטוקן נפרד לבין הצגת כל תו כטוקן נפרד.
גודל אוצר המילים של הBPE הינו 50 אלף טוקנים.

מודל הGPT2 קיים בארבעה גדלים (small, medium, large, XL) כאשר כל גודל מייצג כפולה של 12 שכבות בהתאמה (מ12 שכבות במודל הקטן ועד 48 במודל הגדול ביותר).

במחקר זה נעשה שימוש ביישום המודל הקטן, המכיל כ117 מיליון פרמטרים.

2.2. מודל BERT

המודל השני שנבחר ליישום במחקר הינו מודל BERT (Bidierctional Encoder Representations from Transformers).
מודל זה הוצג לראשונה ב2018 על ידי Google AI במאמר "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding".

מודל הBERT הביא לקפיצת מדרגה משמעותית בעולם עיבוד השפה הטבעית ובעקבותיו יצאו גרסאות נוספות של מודל זה הנשענות על הרעיונות המרכזיים שמהם הוא מבוסס.
החדשנות שהביא עמו היה שהמודל מסוגל להתמודד עם כמה מטלות, מבלי להגדיר אותן בהכרח מראש. דהיינו, למודל לא משנה אם המטלה הינה ניתוח סנטימנטלי של משפטים, השוואה ביניהם, מציאת דמיון או סתירה בין צמד משפטים, חיזוי המשכי משפטים ועוד.
המודל המאומן מראש יודע להבין אנגלית טוב מאוד, ועל מנת להתמודד עם משימה מסוימת ידרש אימון קצר עבור אותה המשימה.

המודל משתמש במקודד (encoder) המקבל קלט משפטים בשני כיוונים- כלומר מההתחלה לסוף ומהסוף להתחלה (ומכך שמו, Bidirectional Encoder).
מודל זה קיים גם בגדלים שונים (base, large, xlarge) כאשר כל גודל מכיל מספר פרמטרים שונה.
במחקר זה נעשה שימוש בBERT Base, המכיל 110 מיליון פרמטרים, ו12 שכבות של Attention (מקודדים דו-כיווניים). מימד הembeddings הינו 768.

גודל אוצר המילים של הEmbedding הינו 30 אלף מילים, והמודל אומן על ויקיפדיה (אשר הכילה בזמנו כ2.5 מיליארד טוקנים) ועל BookCorpus (אשר הכיל כ800 מיליון טוקנים).

2.3. שיטת כיווץ Pruning

שיטת כיווץ זו הינה טכניקה להקטנת גודל הרשת על ידי קיצוץ משקולות שלא תורמות מידע ואינן מסייעות בביצוע המשימה באופן משמעותי.

השיטה מיושמת באחת משתי תצורות. בתצורת הכיווץ הלא מובנה נבחרים x% מהפרמטרים במודל (או בכל שכבה) אשר הכי פחות משפיעים על ביצועיו והם מוסרים מהמודל. בתצורת הכיווץ המובנה, נבחרת בשיטתיות קבוצת פרמטרים שלמה (אשר פחות משפיעה על הביצועים) והיא מוסרת מהמודל.

במחקר זה נבחר לעשות שימוש בשיטת ה global unstructured one-shot pruning, כאשר רכיב הone-shot משמעו ביצוע קיצוץ פעם אחת בכל הרשת (global), במקום לבצע את השיטה בצורה איטרטיבית.

נבחנו שני היקפי הסרה של שיטת כיווץ זו:
הסרת 10% מהפרמטרים במודל והסרת 20% מהם. גם תוצאותיהם הושוו.
ביישום השיטה תרומת הפרמטרים נבחנה לפי מדד L1.

המודל המקורי אשר עליו יושמה שיטת הכיווץ הינו מודל הBERT Base אשר אוזכר לעיל.

2.4. שיטת כיווץ Knowledge Distillation

הרעיון העומד בבסיס שיטת כיווץ זו הינו שימוש במודל "תלמיד" אשר לומד ממודל "המורה" שלו. דהיינו, נוצרת ארכיטקטורה רזה יותר של המודל המקורי, אשר לא לומדת את המידע עצמו אלא לומדת את תשובות המודל לקלטים המוצגים לו בלבד (ללא ידע חיצוני על הקלט, עיקר ההתייחסות בלמידה היא לתגובה מודל המורה).

במחקר זה נעשה שימוש במודל מכווץ מאומן מראש הנקרא Legal-BERT. זהו מודל המסייע במחקר NLP בתחום המשפטנות. מודל זה מהווה גרסה רזה יותר של מודל BERT המקורי והוא מכיל כ33% מהיקף המודל המקורי בלבד. מודל זה אומן בראשיתו על סט נתונים של מסמכים ותיקים משפטיים הרשומים בשפה האנגלית, אשר פורסמו בארצות הברית, אנגליה ואירופה.

לאחר טעינת שני המודלים (מודל BERT המקורי ומודל התלמיד), נעשה אימון למודל התלמיד על בסיס תשובות המודל המקורי ונבחנה תוצאתו.

3. אימון המודלים

ראשית נדגמו 4,000 רשומות רנדומלית מכלל סט הנתונים של הביקורות. לאחר הצמצום, חולקו הרשומות ל80% מהביקורות לטובת סט האימון (3,200 ביקורות) ו20% נוספים לטובת סט הולידציה (800 ביקורות).

לאחר מכן הוגדרה פונקציית אימון מראש (הנקראת train_model) ונעשה בה שימוש על כל אחד מהמודלים.

כלל המודלים לאימון נלקחו מספריית HuggingFace.

עבור המודלים GPT2 וBERT לולאות האימון היו בגודל של 5 אפוקים, ובגודל batch של 16. מספר הטוקנים בכל דגימה בכניסה, אשר מיוצג על ידי אורך הקלט המקסימלי לטקסט, הינו 128 טוקנים (נעשה שימוש בpadding במידת הצורך).

ללולאות אימון אלו הוגדרה שיטת האופטימיזציה ADAMW כאשר ההיפר-פרמטרים שהוגדרו הינם קצב למידה Learning Rate = $1 * 10^{-5} = 0.00001$, אפסילון בגודל eps=$1 * 10^{-8}$, ודיכוי משקולות בגודל weight decay = 0.01. פונקציית הloss שהוגדרה להליך זה הינה MSE.

מודל הGPT2 שאומן הינו המודל בגרסה הקטנה (GPT2 small), המכיל 117 מיליון פרמטרים. גודל הEmbedding הינו 768, וגודל המילון של אוצר המילים אשר עליו אומן המודל מראש הינו 50 אלף טוקנים.

מודל הBERT שאומן הינו המודל בגרסה הקטנה (BERT-base-uncased), המכיל 110 מיליון פרמטרים. גודל הEmbedding הינו 768, וגודל המילון של אוצר המילים אשר עליו אומן המודל מראש הינו 30 אלף טוקנים.

מודל הכיווץ Pruning נבחן על מודל BERT שצוין לעיל עם יישום שיטת global unstructured one-shot pruning, פעם אחת עבור הורדת 10% מהפרמטרים ופעם נוספת עבור הורדת 20% מהם.

לולאת האימון ששומשה למודל זה, לרבות שיטת האופטימיזציה וההיפר-פרמטרים הנלווים לה (קצב למידה, אפסילון) ופונקציית הloss – כולם זהים לאלו אשר נבחנו למודלים GPT2, BERT. כמו כן, כמות האפוקים וגודלם, גודל הbatch ומספר הטוקנים בכל דגימה בכניסה – גם הם כולם זהים לאלו שצוינו לעיל.

מודל הכיווץ של הKnowledge Distillation נבחן על מודל BERT שצוין לעיל, כאשר נעשה יישום למודל של LEGAL-BERT מספריית HuggingFace.

מודל זה מכיל כשליש מהיקף המודל המקורי (BERT-base) בלבד.

כמות האפוקים שהוגדרה לאימון המודל הינה 13 אפוקים בגודל batch של 8. מספר הטוקנים בכל דגימה בכניסה, אשר מיוצג על ידי אורך הקלט המקסימלי לטקסט, הינו 128 טוקנים (נעשה שימוש בpadding במידת הצורך).

כמות האפוקים שונה באימון מודל זה כיוון שהועלה הצורך לבחון האם אימון נוסף מעבר לזה שהוגדר לשאר המודלים במחקר זה עשוי לשפר או להרע את ביצועיו. היבט זה נבחן גם כן במודל זה, מלבד בחינת ביצועיו של המודל לעומת שאר המודלים.

בשיטת כיווץ זו הרשת הקטנה לומדת הן מהסיווג האמיתי של כל דגימה והן מהתפלגות הסיווג של הרשת הגדולה יותר, כאשר על היחס שולט פרמטר alpha=0.5.

ללולאת האימון הוגדרה שיטת האופטימיזציה ADAMW עם אותם ההיפר-פרמטרים שצוינו לעיל בשיטת האימון למודלים הקודמים.

פונקציית הloss הינה פונקציית הMSE.

המדד לבחינה בכל אחד מלולאות האימונים היו השוואת מדד הAccuracy ($\frac{TP+TN}{N}$) של סט האימון לסט הבחינה לאורך האפוקים של לולאת האימון. מדד זה מספק הסתכלות כללית על פרופורציית הסיווגים הנכונים של המודל לעומת כלל הסיווגים.

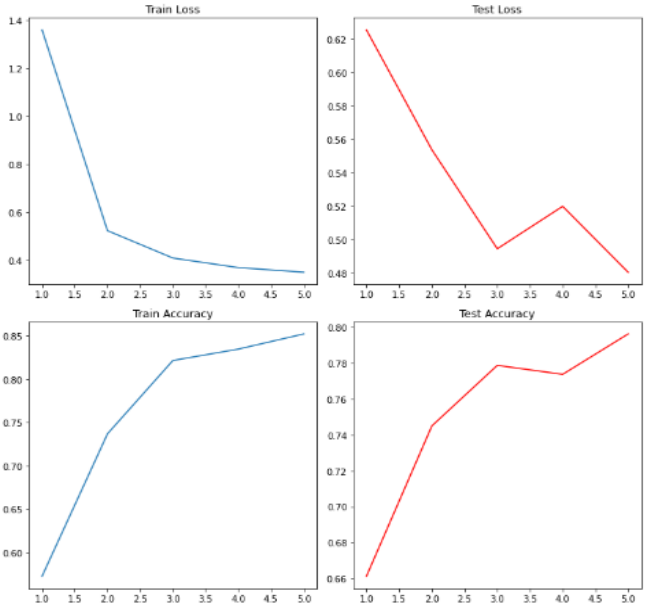
נעשתה השוואה דומה גם לערכים המתקבלים מפונקציית הloss במהלך האימון, על מנת לראות האם לאורך האימון המודלים.

ההסתכלות נעשתה לאורך האימון וצוין הערך המתאים בכל אפוק. כך התאפשר להבין האם המודל משתפר ככל שנמשך האימון וכן מוצג קצב השיפור שלו כפונקצייה של הזמן (לפי האפוקים למעשה).

4. תוצאות

4.1. GPT2

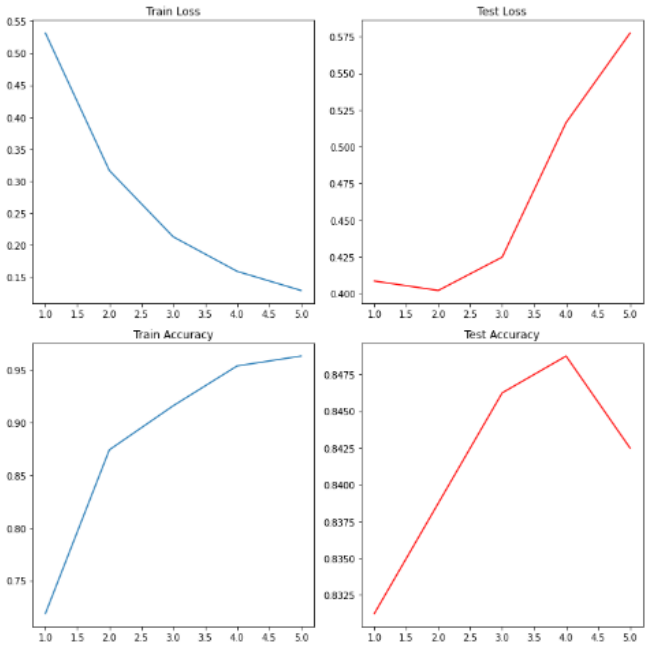
בתרשים 1 ניתן לראות כי מודל GPT2 הצליח לשפר את הדיוק שלו הן בסט האימון והן בסט המבחן.
בסט האימון בוצע שיפור במדד הAccuracy מ0.57 באפוק הראשון ל0.85 באפוק החמישי, וכן פונקציית הloss ירדה בהתאם מ1.4 בקירוב ל0.4 בקירוב.
בסט המבחן ניתן לראות עלייה תואמת כמעט במדויק במדד הAccuracy מ0.66 באפוק הראשון ל0.8 באפוק החמישי, וכן פונקציית הloss ירדה מ0.65 בקירוב ל0.48 בקירוב.
יש לציין כי חלה עלייה קלה בloss של סט המבחן באפוק הרביעי, אך הירידה גדולה יותר מהעלייה באפוק העוקב, כפי שניתן לראות בתרשים 1.



תרשים 1: תוצאות אימון מודל הGPT2

4.2. BERT

בתרשים 2 ניתן לראות כי מודל BERT הצליח לשפר את הדיוק שלו בסט האימון בעיקר (מתחילות טובות והופעות למעולות), אך תוצאותיו בסט המבחן היו פחות טובות.
בסט האימון בוצע שיפור במדד הAccuracy מ0.72 באפוק הראשון ל0.96 באפוק החמישי, וכן פונקציית הloss ירדה בהתאם מ0.55 בקירוב ל0.15 בקירוב.
בסט המבחן ניתן שיפור מזערי במדד הAccuracy מ0.83 באפוק הראשון ל0.84 באפוק החמישי, כאשר ניתן לראות ירידה לעומת האפוק הרביעי. הדבר לא מצביע על שיפור ניכר לאורך האימון, וסימוכין נוסף לכך הינו פונקציית הloss אשר עולה לאורך האפוקים מ0.41 ל0.58 בקירוב.



תרשים 2: תוצאות אימון מודל הBERT

4.3. מודל כיווץ Pruning

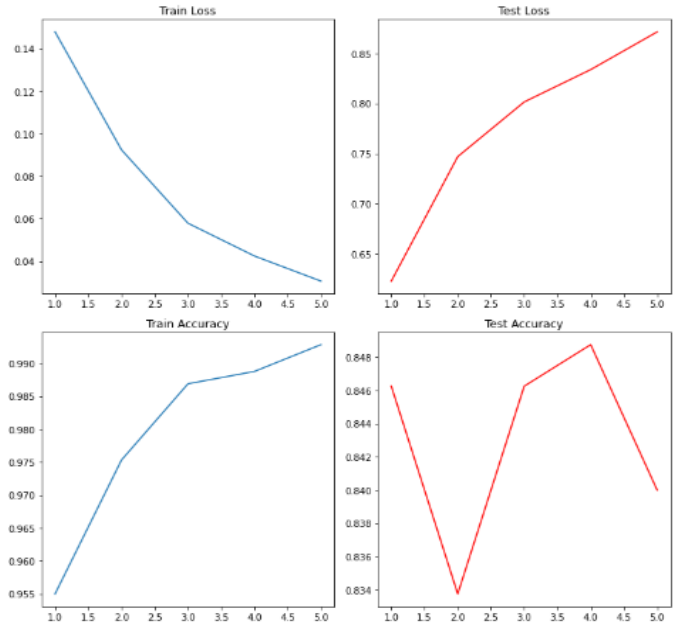
בתרשימים 3 ו4 המייצגים את מודל BERT המקוצץ ב10% וב20% מהפרמטרים בהתאמה, ניתן לראות תופעה דומה לזו שהתקבלה בתוצאות מודל BERT המקורי: התוצאות בסט האימון משתפרות לאורכו, אך בסט המבחן לא ניכר שיפור משמעותי (אם בכלל) והשגיאה אף גדלה לאורך האימון.

כאשר מתבוננים בתוצאות סט האימון של כל אחד מהמודלים המקוצצים, ניתן לראות נקודת פתיחה גבוהה מאוד באפוק הראשון (מדד Accuracy העומד על מינימום של 95%), ושיפורים לכדי 100% דיוק באפוק החמישי. כמו כן השגיאה בסט האימון מתחילה נמוכה מאוד (0.15 לכל היותר בסט הראשון), ושואפת ל0 ככל שמתקדם האימון.

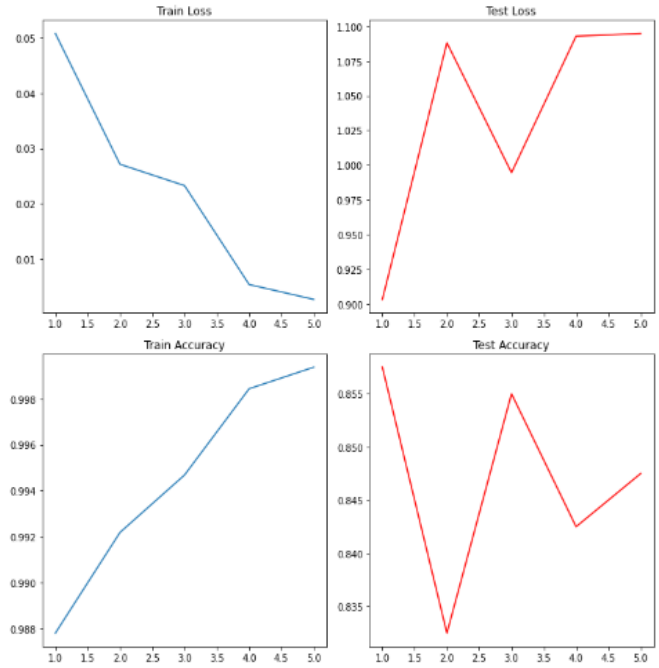
עם זאת, התוצאות בסט המבחן לא מזהירות באותה המידה: מדד הAccuracy גבוה, אך לא

משתנה משמעותית לאורך האימון (סביב 84%).
עם זאת השגיאה לאורך האימון הולכת וגדלה.

ניתן לראות כי ככל שהמודל מקבל אחוז קיצוץ גבוה יותר, כך הדיוק בסט האימון כביכול עולה, אך בעת המבחן השגיאה גבוהה יותר.



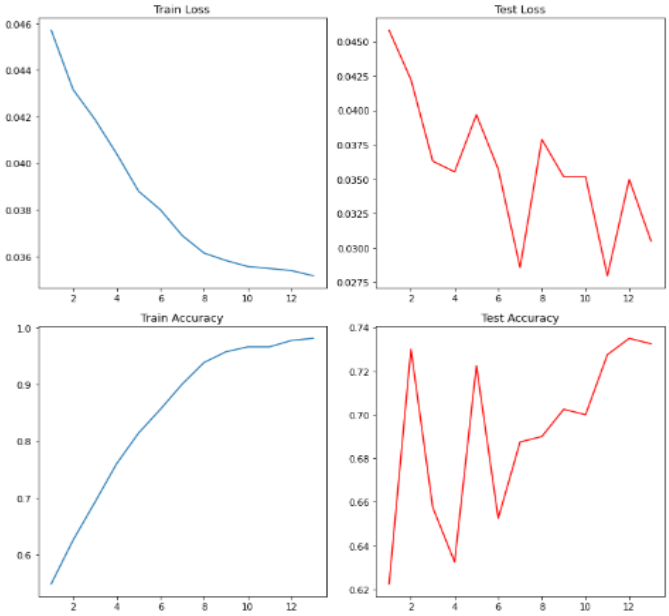
תרשים 3: תוצאות אימון הBERT המקוצץ ב10%



תרשים 4: תוצאות אימון הBERT המקוצץ ב20%

4.4. מודל כיווץ KD

בתרשים 5 ניתן לראות כי מודל הBERT המכווץ בשיטת KD הצליח לשפר את הדיוק שלו הן בסט האימון והן בסט המבחן.
בסט האימון בוצע שיפור במדד הAccuracy מ0.5 באפוק הראשון ל0.97 באפוק האחרון, וכן פונקציית הloss ירדה בהתאם מ0.045 בקירוב ל0.03 בקירוב.
בסט המבחן ניתן שיפור במדד הAccuracy מ0.62 באפוק הראשון ל0.74 באפוק האחרון, וכן פונקציית הloss ירדה מ0.045 בקירוב ל0.03 בקירוב.
יש לציין כי הירידה איננה רצופה לאורך הדגימות בסט המבחן (קיימות עליות בשגיאה במהלך האפוקים), אך לאחר ריצה מספיק ארוכה השגיאה יורדת שוב לעומק נמוך יותר.



תרשים 5: תוצאות אימון הBERT המכווץ בשיטת KD

5. מסקנות ודיון

מהתוצאות שהוצגו לעיל ניתן להפיק מספר מסקנות.

ראשית, ניתן לראות כי בהסתכלות כוללת (בחינת הביצועים עבור סט האימון וסט הולידציה יחדיו), המודל שביצע את עבודת החיזוי בצורה הטובה ביותר הינו מודל הBERT המכווץ בשיטת Knowledge Distillation. מודל זה הצליח להביא ביצועים טובים מאוד בסט האימון (דיוק גבוה מאוד לצד שגיאה כמעט אפסית), יחד עם ביצועים טובים לא פחות בסט המבחן (דיוק מירבי של 0.74 לצד שגיאה כמעט אפסית).

השאלה הנוספת שהועלתה במחקר זה הינה האם כמות גבוהה יותר של אפוקים עשויה לפגוע בביצועי המודל.
כפי שניתן לראות בתרשים 5, על אף שהשגיאות ומדד הדיוק "מזגזגים" לאורך הליכי האימון, המגמה לאורך האימון (והמבחן) מעידים על שיפור לאורך הריצה. כלומר, כמות האפוקים הרבה הייתה טובה למודל זה.
יתכן כי גם עבור המודלים הנוספים הרצת כמות גדולה יותר של אפוקים הייתה משפרת את תהליך הלמידה

המודלים הבאים בתור בביצועים המיטביים הינם מודלי הGPT2 והBERT-base (המודל המקורי). ניתן לראות כי מחד גיסא מודל הGPT2 מעניק שגיאה נמוכה יותר מזו של הBERT בשעת המבחן וכן השיפור שלו ניכר מאוד במדד הדיוק במהלך הרצת האפוקים (הן באימון והן במבחן). מאידך גיסא, למודל הBERT מדד דיוק גבוה יותר מאשר למודל הGPT2 בסט המבחן, על אף שלא חל שיפור מהותי במהלך המבחן.

הדבר עשוי לרמז על כך שמודל הBERT לא מבצע אדפטציה מתאימה לסט הנתונים במהלך הלימוד באופן דומה למודל הGPT2. אינדיקציה נוספת לכך היא ששגיאתו בסט המבחן הולכת ומטפסת לאורך האפוקים בסט זה.
עם זאת, לא נראה בבירור כי אחד המודלים מעפיל על האחר, ועל כן ביצועי שני המודלים הללו (באופן משוקלל) הוערכו בדיון זה בצורה טובה.

אחרונים חביבים בביצועים הינם מודלי הBERT אשר עברו שיטת כיווץ של Pruning בגובה 10% ו20%. כפי שניתן לראות מהמבט של התקדמות האימונים בסט האימון לעומת סט המבחן – המודלים הללו ביצעו תהליך של overfitting ככל שהתקדמה הלמידה.

האינדיקציה הראשונה להליך הoverfitting הינה נקודת הפתיחה הגבוהה של מדד הaccuracy- לאחר מספר מועט מאוד של אפוקים (כבר באפוק השני) ניתן לראות מדד דיוק השואף ל100%, כאשר מדד הדיוק באפוק הראשון מתחיל ב0.95 במודל הכיווץ של 10% וב0.99 במודל של ה-20%.
לאחר 5 אפוקים בלבד שני המודלים מגיעים לרמת דיוק של כמעט 100% בסט האימון. בסט המבחן לא נראה שיפור משמעותי (רמת הדיוק נותרת זהה בקירוב סביב 84%, בשני סוגי המכווצים).
האינדיקציה הנוספת להליך הoverfitting הינה התבוננות בתוצאות סט המבחן של כל אחד מהמודלים במקביל להתקדמותם בסט האימון: השגיאה הולכת ומטפסת ככל שמתקדמים בין האפוקים.

יתכן כי הדבר נובע מכך שסט המבחן מורכב באופן שונה מסט האימון, אך "מספיק דומה" לו כדי להעניק מדד דיוק גבוה יחסית (84%) למודלים המכווצים. המודלים מעניקים "תשובות ששיננו מראש מסט האימון" לסיווגים בסט המבחן ועל כן שגיאתם הולכת וגדלה.

עוד עולה כי ככל שהמודל מאבד אחוז גבוה יותר של פרמטרים (קיצוץ של 20% מהפרמטרים לעומת 10%) – כך הoverfitting הולך וגדל. ניתן לראות זאת כמדד דיוק כמעט מושלם החל מסט האימון הראשון במודל ה20% לעומת מדד דיוק של 0.95 במודל ה10%, וכן שגיאה המטפסת גבוה יותר במודל ה20% (מגיעה ל1.1) לעומת שגיאת מודל ה10% (מגיע "רק" ל0.9).

למעשה, מודלים אלו הינם "גרסה מוקצנת" של מודל הBERT המקורי כפי שהוצג קודם לכן: ביצועים טובים (מאוד) באימון לצד שגיאה גבוהה במבחן אשר הולכת ומטפסת עם התקדמות הסט, וכן מדד דיוק גבוה אך שאינו משתפר בסט זה.

על כל פנים, ביצועי מודלים אלו פחות טובים מאשר מודל BERT המקורי. צפינו כי הביצועים אמנם יפחתו באיכותם, אך מבחן התוצאה מוכיח כי שיטה זו לא עבדה טוב עבור מודל הBERT (כפי שהוכיחה שיטת הכיווץ של הKD למשל).

יתכן כי ביצוע הכיווץ בצורה איטרטיבית (כיווץ x% מהמודל כמה פעמים במספר איטרציות) עבור אחוז נמוך יותר של פרמטרים מקוצצים (5% למשל) היה משפר את ביצועי המודל המכווץ, שכן כך אולי הייתה מתבטאת הדרגתיות מסוימת באופן הקיצוץ.

6. References

6.1. Dr. Eyal Kolman: Advanced Topics in Deep Learning Lectures content
6.2. Models & Academic implementation: *huggingface.co*
6.3. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Lilion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin: *Attention is all you need*
6.4. Mary Phuong, Christoph H. Lampert: *Towards Understanding Knowledge Distillation*
6.5. openAI: *Better Language Models and Their Implications*