למידת מכונה תרגיל 1

איתמר שכטר, $itamar.sh$ 315092759

חלק ראשון $Diffusion - Models$**:**

$$Unconditional - Model$$

אימון המודל: אימנתי מודל עם שלוש שכבות fc שביניהם יש $leakyRelu$.
השכבה הראשונה מקבלת 2 מימדים לכל נקודה וערך של זמן אז היא בגודל 3 והיא ממופה לווקטור בגודל 1024.

השכבה השנייה ממפה מ1024 1024 (כאן רוב הפרמטרים)

השכבה השלישית מחזירה חזרה לפרדיקציה של 2 מימדים.

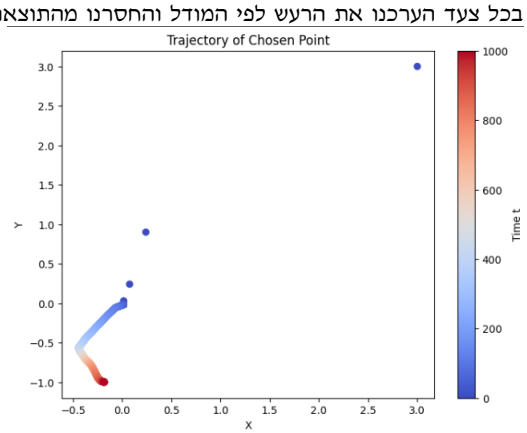
אלה ההיפר פרמטרים שבחרתי למודל: $T = 1000$, $learning\_rate = 0.001$ ואופטימייזר אדאם.

הדאטא סט הוא כמו שנדרש - ריבוע 1 ו־1 על ראשית הצירים. בחרתי 1000 נקודות שיהיה קל ללמוד.

האימון נעשה באמצעות ה$forward$ שהתבקשנו לעשות עם הגרלת זמן ורעש והפעלת לוס על הרעש שהמודל העריך.

שאלה ראשונה - מעקב אחרי נקודה: אוקיי, הנה התוצאה של מעקב אחרי נקודה שהתחילה ב$(3,3)$ והגיעה עד: $(-0.2,-1)$ במשך 1000 צעדים.

בכל צעד הערכנו את הרעש לפי המודל והחסרנו מהתוצאה הנוכחית.

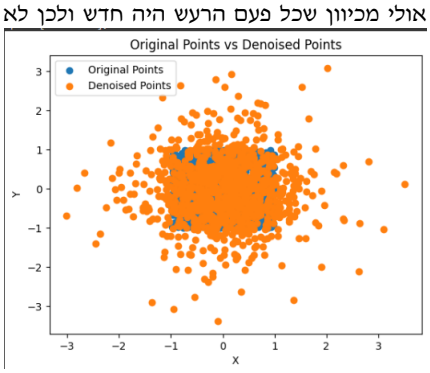


שאלה שנייה - לוס באימון:

הנה התוצאה:

נראה שהמודל התקשה להתכנס לנקודת לוס נמוכה אף על פי שהוא כן הצליחה לאמן מודל $dinoiser$ טוב.

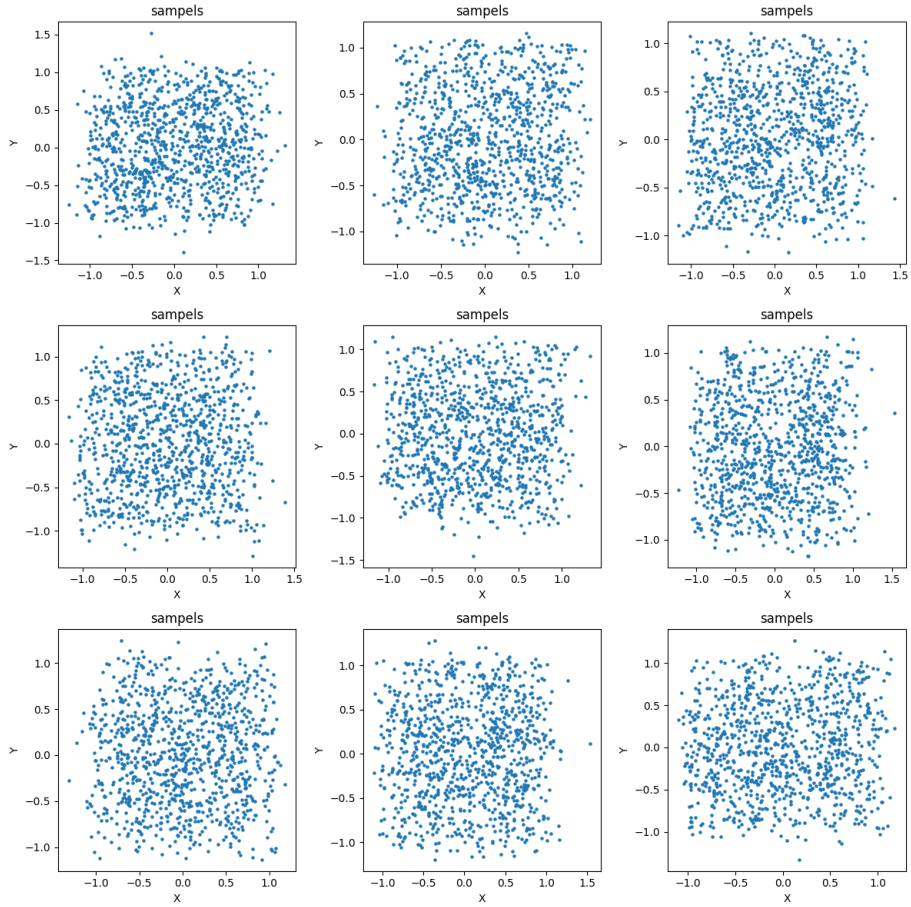
אולי מכיוון שכל פעם הרעש היה חדש ולכן לא הפסיק להקשות על המודל.



והנה איך נראים הנקודות כשמחסירים מהם את הפרדיקציה של המודל (רעש).

לא רואים את הנקודות הכחולות וזה כי אומר שהמודל יחסית מייצג את ההתפלגות המקורית.

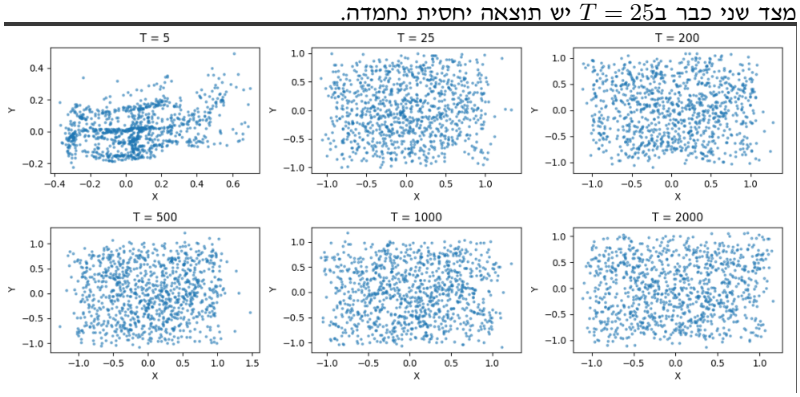
שאלה שלישית - הדפסת דגימות: הנה 9 פעמים 1000 דגימות שנוצרו מתוך $DDim\ sampeling$.



לפעמים הגרף קצת סוטה כי הקואורדינטות שלא משתנות על פי ערכים שבורחים אבל בסך הכל הוא מתלפג די טוב.

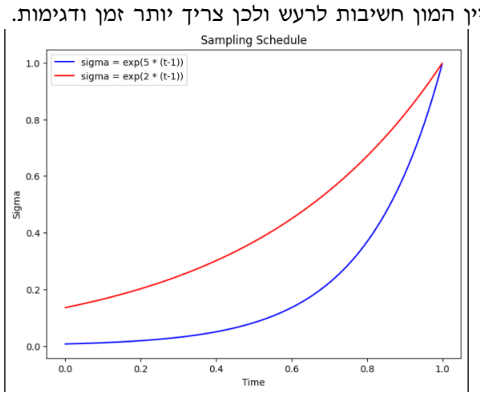
שאלה רביעית - דגימות ביחס לזמן: T מייצג את הזמן שעליו המודל צריך להחזיר פרדיקציה. קיבלנו פיזור טוב יותר ומדוייק יותר ככל שמגדילים את T.

מצד שני כבר ב$T = 25$ יש תוצאה יחסית נחמדה.



שאלה חמישית - שינוי $scheduler$**:** בחרתי לשנות את המקדם שדואג לדעיכה של הרעש לפי הזמן ככה שהוא ידעך יותר לאט, זה מייצר התחשבות גדולה יותר ברעש ומקשה יותר על המודל להתאמן, ככה המודל מכליל טוב יותר, זה מאפשר במודלים מורכבים יותר שיכולים להתגבר על הקושי הזה, או עם מספיק זמן ריצה ודגימות, לקבל תוצאות מדויקות יותר.

חיסרון כאן זה שיהיה לו קשה יותר להתכנס לקראת הסוף לתוצאה מדוייקת כי ה$scheduler$ ותן עדיין המון חשיבות לרעש ולכן צריך יותר זמן ודגימות.

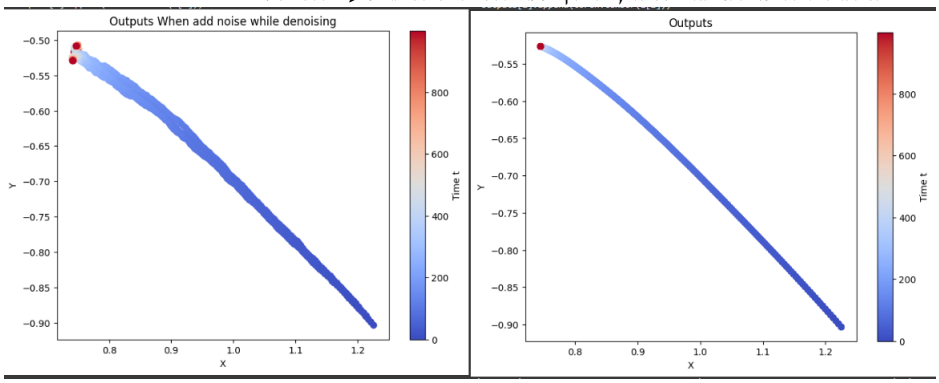


כלל אצבע שלי הוא $scheduler$ שנותן משמעות לזמן ככה שרעש יתחיל ברמות גבוהות כדי לעודד הכללה טובה ולהפחית בהדרגה את הרעש ככל שתהליך הדגימה מתקדם כדי להבטיח התכנסות לקראת תוצאה מדוייקת.

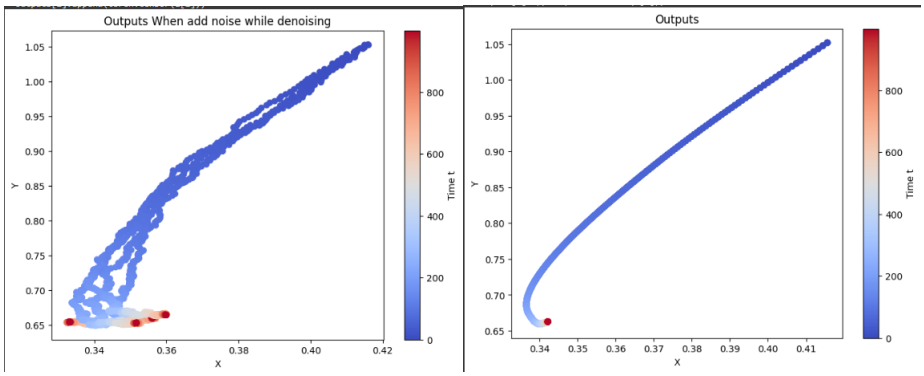
לכן אנחנו משתמשים פה בפונקצייה מעריכית שמתחילה ב1 ואז הערך של $\sigma(t)$ הוא 1 כלומר הרבה רעש, ועם הזמן היא דועכת ככה ש$\sigma(t)$ נהיה מקדם בין 0 ל1 ששואף ל0 ומקטין את הרעש.

יכולתי לבחור מקדם שונה שיהיה לדוגמא 01 ואז $\sigma(t) = exp(10(t-1))$ ואז הפעולה ההפוכה הייתה קוראת והמודל היה מגיע להתכנסות מהר יותר אבל פחות מכליל ותופס את המרחב בהתחלה.

שאלה שישית: יש כאן 10 דגימות שהתכנסו באותו מסלול לאותה נקודה ב1000 צעדים.
השינוי שהצעתי הייתה הוספה קטנה של רעש לדגימת המקור כדי ליצור סטייה קלה.
זו התוצאה ל4 דגימות שונות, מימין ללא השינוי ומשמאל עם השינוי:



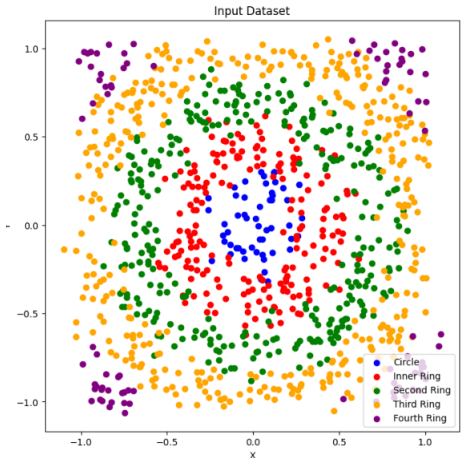
הרצתי עוד פעם כדי לראות אם הם באמת מתכנסים שוב כל כך קרוב אחד לשני וראיתי שלא תמיד זה ממש צמוד, לפעמים יש גם כאלה:



והיו לי ריצות שזה גם בורח ל4 מקומות שונים אחד מהשני.

$$Conditional - Model$$

שאלה ראשונה: כדי שיהיה מעניין בחרתי 5 מחלקות שהן המרחק מהראשית בגדלים קבועים, אפשר לראות שהמחלקה האחרונה היא רק הפינות.



שאלה שנייה:

בניתי עם אותה ארכיטקטורה של שלוש שכבות fc כמו בחלק הקודם. הוספתי ארגומנט של $class\,label$ שנכנס לשכבת $embedding$ ופולט וקטור באורך 46 שייצג את המחלקה והוא, בנוסף לנקודה עצמה ולמימד הזמן ייכנסו לשכבת הfcהראשונה. (שהיא מקבלת וקטור באורך 76 ולא 3 כבר בחלק הקודם)

בעצם כל חישוב של $p(x)$ מקבל זווית חדשה שהוא חישוב של $p(x|c)$.

זה מכריח את המשוואות של dx לקחת בחשבון ב$p(x,t)$ את c. $(4-7)$

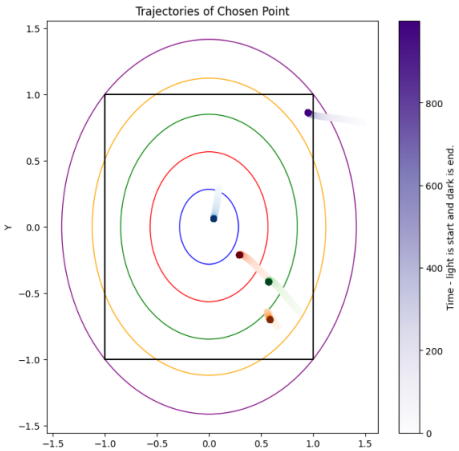
בנוסף המשוואה לחישוב $\hat{x_0}(8)$ משתנה בהתאמה כי המודל מקבל גם את c כחלק מהדגימה.

האלגוריתם עצמו של הלימוד לוקח בחשבון שכל דגימה צריכה לחשב את ה$class$ המתאים לה כדי להכניס אותו למודל. או להגריל על פי $class$ נקודות.

האלגוריתם למציאת דגימות $DDim$ וגם מציאת ה$ELBO$ עם הSNR נשארים זהים.

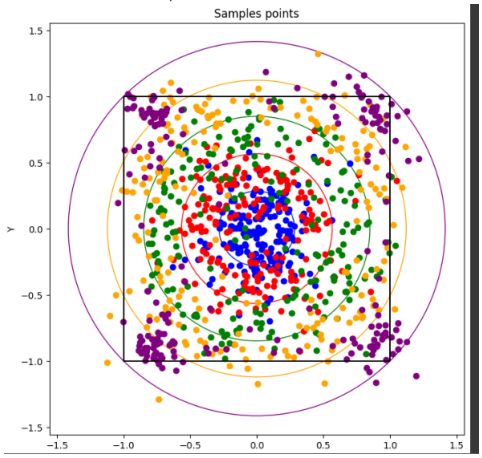
שאלה שלישית:

התחלתי מדגימת נקודות אקראיות מההתפלגות, וקיבעתי את הלמידה של $DDIm$ כשה$class$ שלהן נשאר תמיד אותו $class$.



ניתן לראות בתוצאות שהנקודות עברו ל$class$ המתאים להן לפי הרעש שהעריך המודל שנתנו להן כי הוא יודע בערך איפה ה$class$ צריך להיות.

שאלה רביעית: כמו בשאלה 3, ביצעתי דגימה רק שהפעם ביקשתי שיעשה את זה בסקייל של 1000 דגימות עם יחס שווה לכל מחלקה.

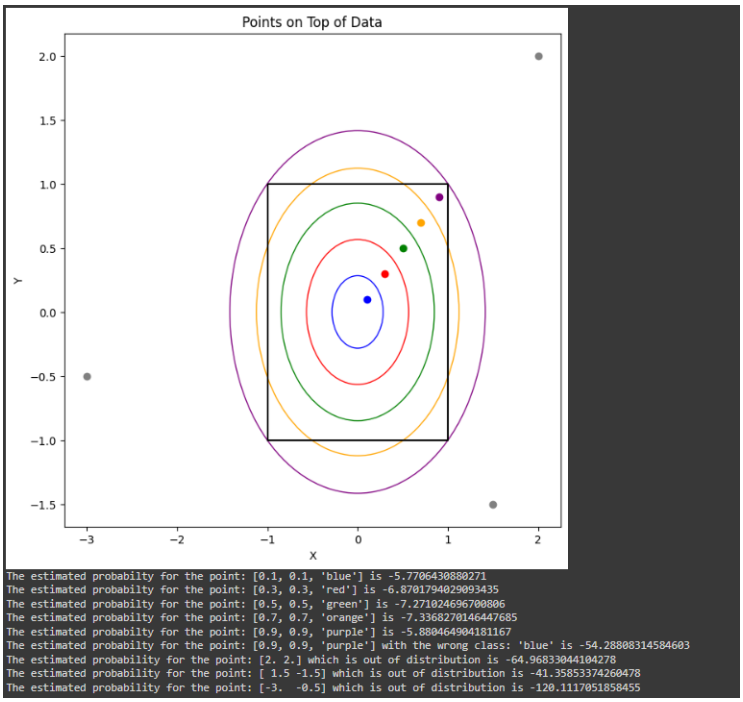


שאלה חמישית: נראה שרוב המחלקות מצליחות להישאר בשלהות הסגול שבקצוות מנסה ממש להתקרב למרכז, והירוק דווקא מנסה לצאת החוצה.

כן נראה שיש התפלגות יחסית טובה אבל יצרתי כאן הטייה לטובת הדגימות שנמצאות בפינות כי שמתי אותם מ=במחלקה נפרדת אבל בפועל הם אמורים לקחת הרבה פחות מקום.

משהו מעניין זה שיש גם משחק עם זה שהמחלקה מוגדרת מחוץ להתפלגות המקורית וזה יפה לראות שהמודל מצליח לא להתבלבל ולהסיט נקודות צהובות וסגולות לשמה.

שאלה שישית:



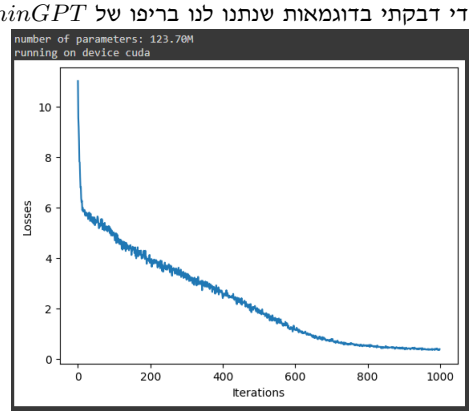
התוצאות טובות לכל הדגימות שאכן באיזור שלהם, וניתן לראות קפיצה אדירה כשמבקשים לחשב נקודה עם $class$ לא נכון. יכול להיות שזה כי ביקשתי מנקודה מהפינות להיות במרכז. אך נקודות שלא בהתפלגות בכלל קיבלו את התוצאות הגרועות ביותר, אפילו יותר מנקודה שבהתפלגות המקורית אך לא ב$class$ שלה בכלל.

חלק שני $Auto-Regressive-Text-Model$

שאלה ראשונה ב3.2.4:

אימנתי את המודל עם הפרמטרים הבאים: $max\_iters = 1000, learning\_rate = 0.0005, block\_size = 64$

די דבקתי בדוגמאות שנתנו לנו בריפו של $minGPT$. אלה התוצאות של הלוס תוך כדי אימון:



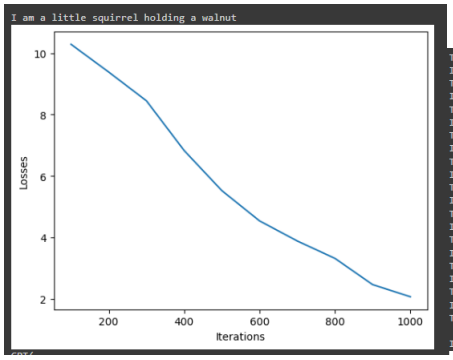
שאלה שנייה - $inversion$:

הוספתי לפונקציית $forward$ של המודל ארגומנט שנקרא $embedded$ שמסמל שהקלט הוא כבר וקטור שכבר עבר שכבת $embedded$.

בתוך הפונקציה יש תנאי האם להפעיל את השכבה על הוקטור או לא על פי אותו ארגומנט ($flag$).

מבחוץ למודל, בשביל האימון, יצרתי וקטור מטריצה שהיא הפרמטר שלנו ואותה למדנו כשאחנו משווים את הפלט של המודל כל פעם ל$tokens$ מהמשפט הנדרש.

והנה התוצאה של 1000 איטרציות:



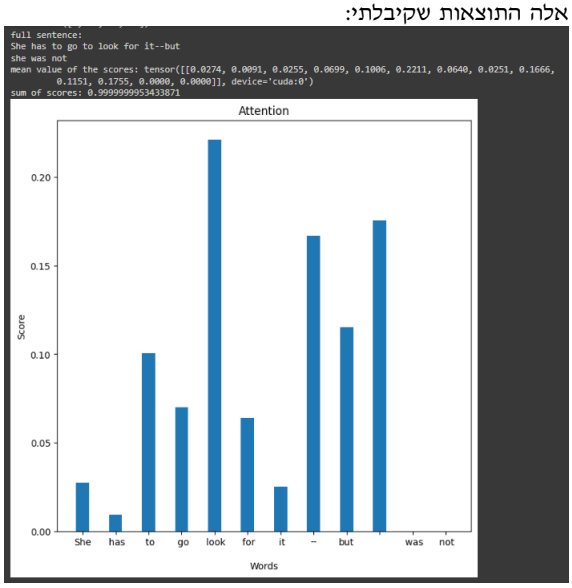

```
The current result is:
I am a little room up like a little
The current result is:
I am a little across a walem from
The current result is:
I am a little squirrel holding a walam
The current result is:
I am a little squirrel holding a walanch
The current result is:
I am a little squirrel holding a walnut
The current result is:
I am a little squirrel holding a walnut
The current result is:
I am a little squirrel holding a walnut
The current result is:
I am a little squirrel holding a walnut
The current result is:
I am a little squirrel holding a walnut
The current result is:
I am a little squirrel holding a walnut
The final result is:

I am a little squirrel holding a walnut
```

כבר באיטרציה ה300 המודל הגיע לתוצאה קרובה ואחרי ה500 כבר הגיע לתוצאה הנכונה.
(בשביל יצירת המשפט הוספתי את הארגומנט גם לפונקצייה $generate$)

שאלה שלישית - $attention$:

הוספתי לפונקציית $generate$ של המודל עוד ארגומנט בשם $attention$ שצריך להיות מספר.
שמסמל שאנחנו מבקשים מהמודל להחזיר את ה$attention$ של המילה במקום של ה$attention$.
הארגומנט מחולחל גם ל$forward$ וגם לבלוקים ובתוכם ל$attention$ כדי לשאוב את המידע המתאים.

אלה התוצאות שקיבלתי:

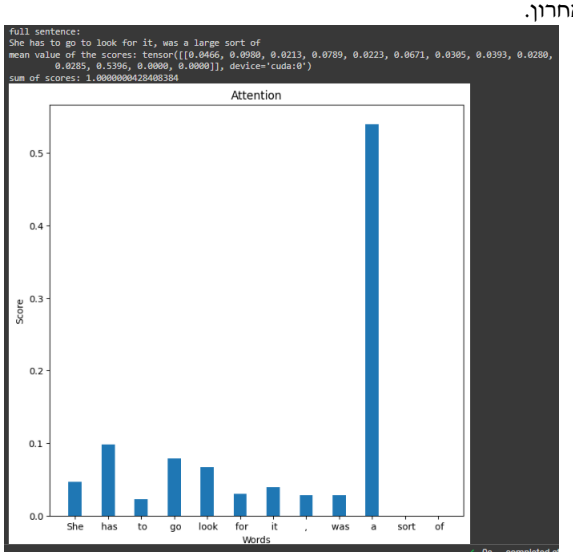


יש כאן חשיבות מעניינת מילים מסוימות.

המשפט שנפלט היה: $She\ has\ to\ go\ to\ look\ for\ it - but\ she\ was\ not$
המילה ה11 הייתה "was" כשה$prefix$ היה $She\ has\ to\ go\ to\ look\ for$.
נדמה שהמילה שהכי השפיעה על המילה "was" הייתה המילה $look$ וגם הרווח והמקף היו משמעותיים.
הסכום של ה$score$ היה 1 כי יש לנו כאן $softmax$ ואנחנו שואלים למה המילה הכי קרובה.

שאלה רביעית - $more\ attention$:

שוב הוספתי ארגומנט, שמבקש אותו דבר רק שייקחו את ה$attention$ של הבלוק הראשון ולא האחרון.



הפעם המשפט היה $She\ has\ to\ go\ to\ look\ for\ it\ ,\ was\ a\ large\ sort\ of$
כשהמילה ה11 הייתה $sort$ והכי הושפעה מהמילה a בבלוק הראשון - מאז אני מעריך שנעשו עוד שינויים שהשפיעו על המידע.

שאלה חמישית :

```
Generated Sentence:
Beginning of a journey is
boots every Christmas.'

  And so
Score: tensor(6.1967, device='cuda:0', grad_fn=<AddBackward0>)
```

נתתי את הפתיחה: $Beginning\ of\ a\ journey\ is$
והמודל השלים: $Beginning\ of\ a\ journey\ is\ boots\ every\ Christmas.'\ And\ so$