

# מנגנון תשומת הלב לפני עידן הטרנספורמרים

לאחר שראינו כיצד רשתות איטרטיביות כנו LSTM/RNN מודללות את התלויות בקלט סידרתי, נציג כעת את המנגנון המהווה את לבה של ארכיטקטורת הטרנספורמרים: מנגנון תשומת הלב (Attention mechanism). במהלך פרק זה נענה על השאלות הבאות:

- מהו מנגנון תשומת הלב?
- מדוע הוא נדרש?
- כיצד מנגנון תשומת הלב הפותח במכונות לומדות?

## מהו מנגנון תשומת הלב?

לפנינו שנדריך מהי תשומת לב במכונות לומדות, נבחן כיצד היא באה לידי ביטוי בكونבנצייה האנושית. תשומת לב הינה התמקדות סלקטיבית בחלקים הרלוונטיים ביותר של מידע שאנו חווים וסינון חלקים בעלי חשיבות פחותה. תשומת הלב מאפשרת לנו למקדד את המשאים העומדים לרשותינו בצורה עיליה בכל סיטואציה, ובכך אנו יכולים להתמצא ולהבין טוב יותר את מלאה האינפורמציה הזמין לנו. מנגנון תשומת הלב הוא יכולת מולצת של בני אדם, אולם היא גם כן יכולה נלמדת הניתנת לשיפור במהלך חיינו, (לדוגמה [מיינדפולנינג](#)).

האופן המופשט שבו מנגנון תשומת הלב פועל במוחנו הינו:

1. קבלת קלט על ידי הסנסורים החושיים שלנו (כגון מערכת ראייה, מערכת שמע וחוש הריח).
2. עיבוד מוקדים וסינון המידע על ידי המוח.
3. בחירת החלקים החשובים ביותר של המידע בהתאם למידע קודם והקשר הנוכחי.
4. המידע שנבחר כרלוונטי עובר עיבוד, ובסיום נשמר בזיכרון.

מנגנון תשומת הלב למעשה מנע علينا בתור בני אדם, מכיוון שהואאפשר לנו **להתעלם** ממידע שאינו חיוני לנו (המקיף אותנו הרבה יותר מאשר המידע חיוני). דוגמה לכך היא "מסיבת קוקטייל" (cocktail party) שבה אנו נוכחים במפגש חברתי שבו מספר רב של אנשים מדברים בו-זמנית, ואנו מעוניינים להתמקד באדם אחד במיוחד. המוח שלנו מסייע לנו למקדד את תשומת הלב שלנו באדם זה ולהתעלם משאר הקולות שהופכים לירען רקע.

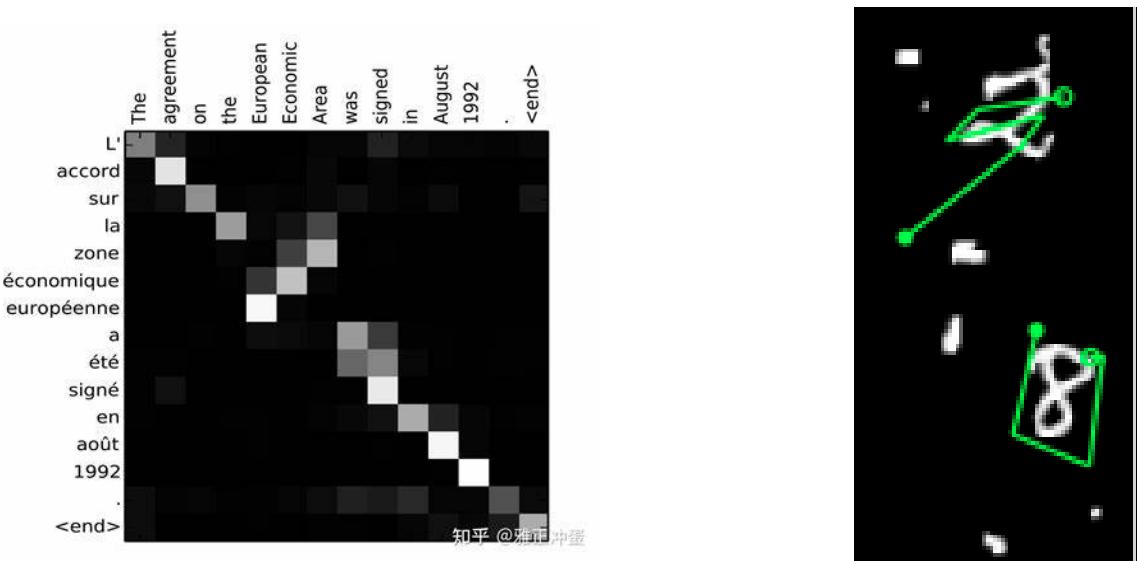
הנקודה המזקקת את העיקרון של תשומת הלב הינה יכולת לשער מידע (information assessment). במלים אחרות אנו **לומדים להעניק משקל** לכל פיסת מידע, ביחס לכל פיסת מידע אחרת, בהתאם להערכתה שלה, והקשר שלה עם מידע שאנו כבר ידועים (הנanager בזיכרון), או חווים ברגע נתון.

אז מהו החידוש שמנגן תשומת הלב מביא אליו כאשר הוא משלב בראשתות נוירונים? התשובה לכך היא שמנגן תשומת הלב הוא **פונקציה הנלמדת** כחלק מתהיליך אימון הרשת, שמרתתה לחשב מהו המשקל היחסי של כל יחידת DATA בהינתן הקשר ומידע שנאגר עד אותה הרגע. מנגנון תשומת הלב יוצר ייצוג וקטורי רציף תלוי-הקשר (contextualized representation) עבור יחידת המידע (כגון טוקן או מילה), ותוכאו הינה פונקציה רציפה שנייתן לנוזר אותה ביחס לפרמטרים של מנגנון תשומת הלב (כלומר soft attention). מנגנון תשומת הלב משקן יחידות DATA במרחב וקטורי כאשר ייצוג של יחידת DATA במרחב זה מתחשב בעוצמת הקשר **הרציפה (לא דיסקרטית)** בין הבין היחידות DATA. מאוחר והמשמעות של יחידות DATA הינה פונקציה גזירה ביחס לפרמטרים של מנגנון תשומת הלב (וגם ביחס לייצוג הקלט של מנגנון זה) אנו יכולים לאמן את הרשות לייצג את הקשרים האמיתיים בין החלקים השונים של הטקסט.

איור 1 (שמאל) המציג את מפת תשומת הלב של משפט באנגלית בצרפתית, ממחיש את עקרון הרציפות של פונקציית תשומת הלב. האיור מציג משקל של הטוקנים בסדרה הראשונה ביחס לכל הטוקנים בסדרה השנייה (מנגנון זה נקרא תשומת הלב המוצלבת אשר נרחיב עלי' בהמשך). איור זה לquoד מהמאמר שהציג את מנגנון תשומת הלב לראשונה עבור משימות שפה טבעיות. נדון במאמר זה בפרק העוקב בתשובה לשאלת "יכן מנגנון תשומת הלב התפתח במכונות לומדות?".

אולם, לא כל מנגנון תשומת הלב נולדו שווים. [מאמר זה](#), שהציג את השימוש הראשוני במנגנון תשומת הלב (במאמר הופיעה לראשונה המונח attention), עברו סיווג תമונות, השתמש במנגנון תשומת הלב דיסקרטי (hard attention). אופן פעולה הרשות דומה למשחק חיבור נקודות באמצעות קוים היוצרים צורה. הרשות מחפשת את הפיקסלים המקיימים את את האוזור הרלוונטי ביוטר בתמונה (איור 1 מימין), כך שאם נחבר אותו בקוים ישירים, נקבל תמונה חדשה המכילה את המידע החשוב ביוטר הדרוש לסיווג התמונה. בכל הפעלה של הרשות, מנגנון תשומת הלב בוחר פיקסל חדש, ומוסיף אותו לפיקסלים שנבחר באוטרציות הקווות, כאשר בחירת הפיקסלים הללו יוצרת בסופו של דבר את האוזור התוחום. מכיוון שפונקציית תשומת הלב הדיסקרטית אינה גירה (מקבלת ערכים דיסקרטיים בלבד) לא ניתן למתב (optimize) אותה עם שיטות משפחת מורד הנגדיאנט (gradient descent). עקב כך המאמר עשה שימוש בשיטת אימון משטר מבוסס החלטות (policy loss) השואלה מתוך עולם הלמידה מבוססת החיזוקים (reinforcement learning), שלא דורשת גירות של פונקציית מטרה, על מנת לאמן את מנגנון תשומת הלב.

בעולם ניתוח השפה, מנגנון תשומת הלב הדיסקרטי, לימד למצאו את הטוקנים החשובים ביותר (בהתאם לשיממה) לבנייה של וקטור הייצוג עבור טוקן נתון. לדוגמא, נניח והמטרה של הרשות היא למצוא את האובייקט החשוב ביותר במשפט "The cat is playing with the toy, it is soft". הרשות תיתן את המשקל המקסימלי למילה "cat" ומשקל נמוך לשאר המילים.



איור 1 - בחירת הפיקסלים החשובים ביותר בתמונה (ימין) מפת תשומת הלב רציפה בתרגום משפט מאנגלית לצרפתית (שמאל)

### תשומת לב "לא מפורשת" לעומת תשומת לב "מפורשת"

המנגנון שדנו בו עד עתה נקרא תשומת לב מפורשת (explicit attention), שבו אנו משערכים קשרים בין יחידות מידע שונות באותו יום. לעומת זאת, ישנו מנגנון תשומת הלב נוסף הנקרא תשומת לב "לא

מפורשת" (implicit attention). מנגנון זה הינו "תוצר לא מכון" של רשותות עמוקות. רשותות אלו נוטות להתמקדש בחלקים מסוימים של המידע ולהתעלם מאחרים. לדוגמה, בסיוג של מנח גוף מתוך סרטונים (pose estimation), תהיה לרשות הנטיה להתמקדש באזורי שביהם מופיעים חלקי גוף ולהתעלם מאזורי בהם לא מופיע אדם בכלל (ורק או אובייקטים דומים). ניתן להמחיש באופן ניטוח תמונה המרמזת על ידי איור 2, המגייס מההרצאה הבאה. האירור ממחיש את מיקוד הרשות בעת ניתוח תמונה עבור אימון סוכן במשחק. הסוכן לומד לנסוע בתוך השביל ולהימנע ממפגעים באמצעות מנגנון שמתמקד על מנת לקבל את החלטה הבאה. מכיוון שהתקדמות בשביל מובילה לעלייה בנזודות הרשות מתמקדת באופק, ובולוק התוצאה שמצוין את הניקוד. **מעתה, בכל מקום בטקסט שנתייחס לתשומת לב, נתקoon לתשומת לב מפורשת.**



איור 2 - תשומת לב מרמזת ברשות מבוססת חיזוקים

## מדוע מנגנון תשומת הלב נדרש?

כפי שהסבירנו בפרקם הקדמים, ארכיטקטורות איטרטיביות סבלו מבעיה מרכזית המשותפת לכלן, והיא רכבי זיכרון הקבועים בגודלם, ומאידן, קלט בעל אורך משתנה. כתוצאה לכך אין נאלצים לפחות משפתיים באורכים שונים לוקטור בגודל קבוע. ולכן, בעת קידוד קלטים העולים על גודל מסוים, נתקנס לבעיית מידול של תלויות ארכות הטווות. עיקר הבעיה בא לידי ביטוי בכך שלא ניתן להשתמש בכל יחידות הקלט באופן מפורש לבניה של וקטור ההקשר אליו עברו ייחידת DATA. במילים פשטות, מטרתנו לאפשר לוקטור המקיים את הקלט לגשת לכל חלקיו (של הקלט) במקביל בעת בניית הייצוג.

על מנת להמחיש את הנושא, נסתכל על הפסקה הבאה:

**"צח, מהנדס תוכנה,** עבד מהבית בשנתיים האחרונות. הוא מתגורר בישוב קטן בצד הארץ עם אשתו ושני ילדים. היישוב שקט ורנגן ויש בו תחוות קהילתיות. צח נהנה לבളות עם משפחתו ולצאת לטווילים ארוכים בעיר הסמוכה. הוא מעירין את הנגישות **עובדת מרחוק** מצעה לו. בשבוע שעבור גילה צח כי החברה שלו מתכוננת לישם **מדיניות חדשה** שתחייב את כל העובדים לעבוד מהמשרד. צח שוקל כעת האם **לעבור לנגור בקרבת המשרד** שנמצא במרקם הארץ או **לחפש עבודה חדשה**".

מודל ניטוח וסיכון טקסט המבוסס RNN או LSTM עשוי להתקשות להבין מהו המידע החינוי בטקסט זה וליצור סיכון תמציתי איקוטי. מכיוון שאופן עיבוד הקלט הוא טוקן-אחרי-טוקן, תוצאה אפשרית של מודל איטרטיבי יכולה להיות:

**"צח, מהנדס תוכנה, שעובך מרוחק, עובר längor ליד המשרד, או מחפש עבודה חדשה, בוגל שניי" במדיניות החברה"**

למרות שהרשות אכן "דחסה" את כל המידע החשוב, עדין חסירה נהירות (קוהרנטיות) בפolut.

לעומת זאת, רשות המקיימת את התנאים הבאים:

- בעלת מנגנון תשומת הלב.
- מייצרת וקטור הקשר גדול מספק.

יכולת לספק את התמצאות הבא:

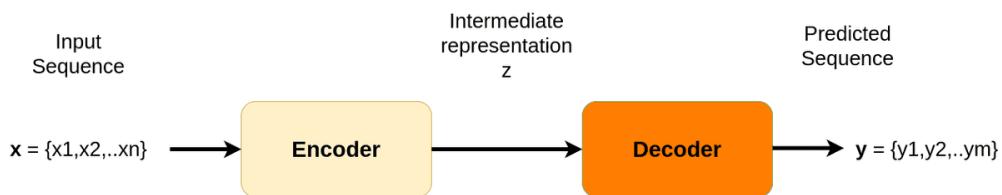
**"צח, מהנדס תוכנה שעובך מרוחק, שוקל האם לעבוּר לעיר הקרובה לעבודתו במשרד או למצוא משרה חדשה המאפשרת עבודה מרוחק בשל מדיניות החדש בחברה"**

#### נקודות לשיכום הפרק:

- מנגנון תשומת הלב במכונות לומדות הינו פונקציה נלמדת, השואבת השראה מתשומת הלב בكونיציה האנושית, וממשקלה חשיבות של קלט ביחס לקלט אחר.
- מנגנון תשומת הלב שהוא דנים בו נקרא soft attention שמהווה פונקציה רציפה. תוכנה זו נובעת מהעבודה שימוש של יחידת DATA נמוך ביחס לכל יחידות הדטה האחרות.
- ישנן שני סוגי של מנגנון תשומת לב. תשומת לב מפורשת, הינה פונקציית תשומת לב המוממשת באופן יומיום חלק מארQUITקTOT המודול. לעומת זאת, תשומת לב לא מפורשת הינה תוצר של עיבוד DATA על ידי רשותות عمוקות, הלומדות חשיבות של אוזרים מסוימים בקלט ללא הכוונה יזומה.
- רשותות בעלות מנגנון תשומת לב יכולות למדו קשרים מורכבים יותר בקלט וליצור פלט קוהרנטי יותר ביחס לרשותות איטרטיביות.

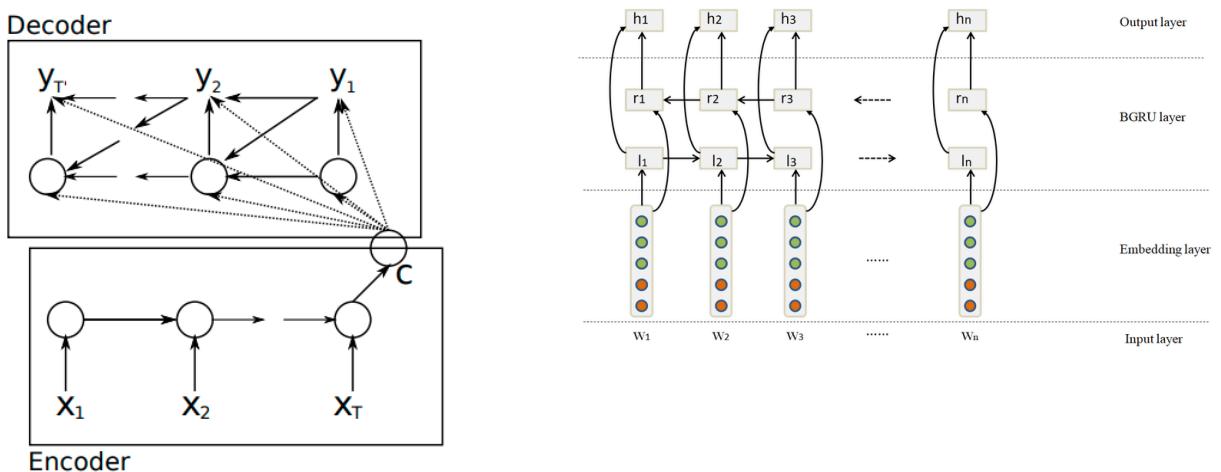
### כיצד מנגנון תשומת הלב התפתח במכונות לומדות?

המאמר הראשו שהציג שימוש במנגנון תשומת הלב עבור משימות של ניטוח שפה טבעית השתמש בארכיטקטורת מיקודד-מפענה (encoder-decoder architecture), המთוארת באייר 3, לטובות תרגום מאנגלית לצרפתית. על מנת להסביר את הצורך במנגנון תשומת הלב, נציג כעט מושג חשוב בתחום עיבוד השפה הטבעית הנקרא "יעשור" (alignment). מושג זה מתאר בהקשר של תרגום, את התאימות בין מילה/מילים משפט המקור לבין מילים בשפט היעד (אייר 1 משמאל מдинם את היישור בין מאנגלית לצרפתית). במקרים אחרים, זהו יציג של "עוצמת הקשר" בין קבוצות של מילים בשפט היעד לבין מילים בשפט המקור.



אייר 3 - הפשטה של ארכיטקטורת מיקודד-מפענה

כאן המקווד מקבל משפט (או קטע של טקסט) בשפה א' כסדרת טוקנים, ומיצג את המידע במשפט כוקטור במינימד חבוי (latent representation). מן הצד השני, המפענה מחלץ מידע הוקטור שהתקבל את המידע הרלונטי ומפיק מתוכו את המשפט בשפה ב'. הרשות מאומנת כמקשה אחת, ולומדת לזכור ולפענה באותו הזמן. במאמר גם המפענה וגם המקווד מושמו על ידי יחידות GRU משורשרות (אולם ניתן להשתמש גם ב-Bi-RNN או LSTM), כאשר המקווד מורכב מושני שרשרים של יחידות GRU כאשר שרשרת הראש שמשאריכטקטורת GRU זו כיוונית מושני שרשרים של יחידות GRU משמש להעברת הקלט מהסוף להעברת הקלט מתחילה (טופו), והשרשור השני של יחידות GRU משמש להעברת הקלט מוקף על הקלט, ככלומר גם מתחילה ו גם מסוף, שכן כל המידע (משפט המקורי) נתן לנו בזמן הפעלת מוקף על הקלט, לעומת זאת הפלט בפענה נוצר באופן אוטורגרטיבי (מילה אחרת מילה, המודל לאחר האימון inference). לעומת זאת הפלט בפענה נוצר באופן אוטורגרטיבי (מילה אחרת מילה, כאשר פלט נכון הופך לקלט עתידי לאחר שנוצר) בזמן ההסקה, שהופך את השימוש ברשות דו-כיוונית בפענה למהר משולל היגיון.



איור 5 - ארכיטקטורת מוקוד מפענה

איור 4 - GRU דו-כיוונית

בארכיטקטורות מסווג מוקוד-מפענה שקדמו למאמר, הקלט שהפענה מקבל בכל איטרציה (יצירת יחידת>Data (הינה המצב הפנימי  $h_i$  והפלט  $y_{i+1}$ ) מהאיטרציה הקודמת (של המפענה). על מנת לחבר את המידע שהמקוד למד מהקלט, המפענה מקבל בסיסף את וקטור המוצא של המקווד שנקרו ל- $C$ . בארכיטקטורות מוקוד-מפענה בסיסית,  $C$  הינו שרשור המצביעים הפנימיים שחושבו באיטרציה الأخيرة מכל אחת מהרטשות המרכיבות את הרשות הדו-כיוונית במקוד ( $[h_i \leftarrow h_i; h_i \rightarrow C]$ ) שכן הוא מכיל מידע על כל הקלט. איור 5 מציג את הארכיטקטורה שהסבירנו בפסקה זאת.

### כיצד בא ידי ביטוי מנגן תשומת הלב במאמר?

כפי שהזכרנו קודם לכן, השימוש בתשומת הלב מועד לפתרור את בעיית היישור בין הקלט לפלט. עקב אחרי של הארכיטקטורות שקדמו לו של המאמר, הייתה שימוש במצבים הפנימיים האחרונים של המקווד (BGRU). מכיוון ש מצבים אלו היכלו מידע דחוס על כל הקלט, לא ניתן היה לנדר את התלותות המקומות בין המצביעים הפנימיים של המפענה, לאלו של המקווד. ככלומר, וקטורי ההקשר המתקברים בכניסה למפענה דוחסים את כל המידע מסדרת הקלט של המקווד, ללא התחשבות בקשר שבין יחידות מידע בסדרה א' לייחידת מידע הנבנתה בסדרה ב'. לעומת זאת, בשיטה המוצעת במאמר, וקטור ההקשר שהפענה מייצר בעת בניית יחידת פלט, מקבל את המידע על כל יחידות הדאטה של הקלט שנבנו על ידי המקווד. וקטור ההקשר נוצר סכום משוקל של כל המצביעים הפנימיים של המקווד, כאשר המשקלים מדלים את הקשרים בין כל יחידות הקלט לייחידת פלט.

## מנגנון תשומת הלב: המידול המתמטי

המושג הראשון שהמאמר מגדר הינו **עוצמת היישור** (alignment score), שמייצג את הקשר בין המצב הפנימי  $i$ -ו לבין מצב פנימי  $j$  כלשהו במקודם. על מנת למןוע בלבול, נגידר את המצבים הפנימיים של המפענה  $C-h$  (כפי שהוא מובא במאמר) ואת המצבים הפנימיים במקודם  $N-h$ . האינדקסים  $j$  ו- $i$  מייצגים את המספר הסידורי של יחידות הקלט והפלט  $i-j$ . ערך עבור יחידת פלט  $j$  נגידר וקטור  $e_{ij}$  באופן הבא:

$$(1) e_{ij} = \text{attention}(s_{i-1}, h_j) = v_a^T * \tanh(W * [s_{i-1}; h_j]), j = 1, \dots, T$$

משוואה 1 - חישוב עוצמת היישור (מנגנון תשומת הלב)

כאשר:

- $e_{ij}$  - עוצמת יישור לא מנורמלת.
- $h_j$  - המצב הפנימי של יחידה  $j$  של המקודם.
- $s_{i-1}$  - המצב הפנימי מהיחידה  $i-1$  של המפענה.
- $W$  - מטריצת המשקלות של מנגנון תשומת הלב.
- $v_a$  - וקטור המשקל של פונקציית תשומת הלב.
- $T$  - מספר יחידות הדטה במקודם.

כאמור, מנגנון עוצמת היישור הינו פונקציה נלמדת המכחשת את עוצמת הקשר שבין המצב הפנימי של המפענה למצביעים הפנימיים של המקודם. מכיוון שהמכפלות בתרח פונקציית הטנגנס ההיפרבולית  $\tanh$  יוצרים וקטור בגודל 1 אח, וערך תשומת הלב בין שתי יחידות צריכה להיות סקלר, המכפלה בוקטור  $v_a$  יוצרת סקלר במצאה. נשים לב כי המטריצה  $W$  והוקטור  $v_a$  הינם פרמטרים הנלמדים (מאומנים) של המודל.

המושג השני שהמאמר מגדר הינו **משקל תשומת הלב** (attention weight). מטרת מנגנון תשומת הלב הינה ליצור משקל החשיבות של טוקן אל מול כל טוקן אחר, כפונקציה רציפה ומירה. המשמעות של רציפות בהקשר שאנו מדברים עלי, הינו משקל של עוצמת הקשר  $e_{ij}$  (המקושרת למצביע הפנימי  $i$  מהמקודם) ביחס לכל עוצמות הקשר האחרות. עוצמות אלו מייצגות את הקשר שבין כל שאר המצבים הפנימיים של המקודם ביחס למצביע פנימי הנתון של המפענה. על מנת לעשות זאת, אנו משתמשים בפונקציית softmax המופעלת על עוצמות הקשר. חישוב זה למעשה פותר את בעיית היישור שפתחנו אותה את הפרק, מכיוון שפונקציית softmax תהפוך את תוכנת היישור, שעמדה בפני עצמה, להיות פונקציית צפיפות הסתברות התלויה בכל המצבים הפנימיים של המקודם. אנו מבצעים פעולה זו עבור כל מצב פנימי של המקודם אל מול אותו מצב פנימי של המפענה, ובכך מקבלים את עוצמת הקשר הרציפה שהזכרנו.

$$(2) \alpha_{ij} = \exp(e_{ij}) / (\sum_{k=1}^T \exp(e_{ik}))$$

משוואה 2 - חישוב משקל תשומת הלב עבור זוג יחידות DATA  $i-j$

כאן  $\alpha_{ij}$  הוא מספר יחידות חישוב במקודם כולם אורך מקסימלי של סדרת דטה שניית להכנס בו כמקרה אחד.

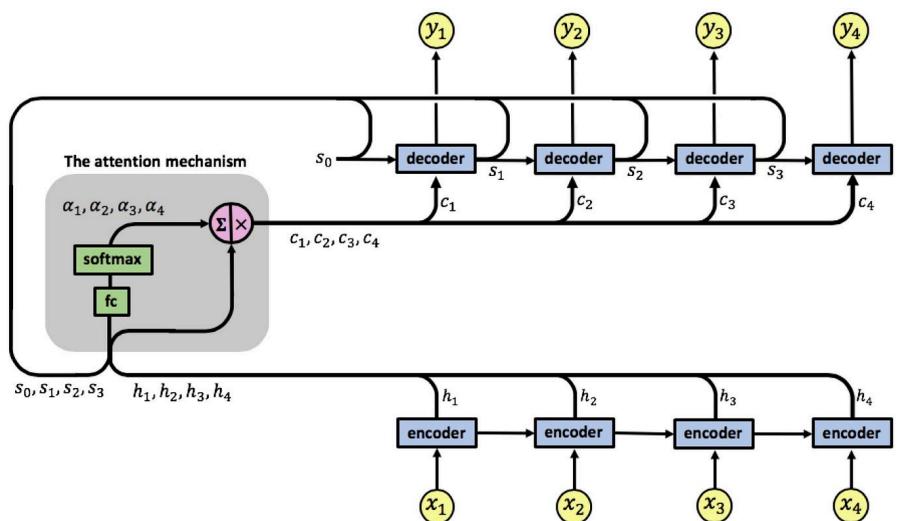
החלק האחרון בהסבר שלנו הוא **בנייה וקטור הקשר דינמי C**. אנו משתמשים במשקלות תשומת הלב שחייבנו, ורכיבים את הוקטור מכפלה של משקלות זו במצב הפנימי של המktorד הקשור אליו. במקרה זה של הוקטור הקשר מאפשר לנו לחתם למפענה את המידע הרלוונטי ביותר ביחס למצב הפנימי הנוכחי שלו. מצבים פנימיים של המktorד בעלי עוצמת קשר נמוכה ביחס למצב הפנימי של המפענה, לא ישפיעו על הוקטור C (שכן אם ערך משקלות תשומת הלב שלהם נמוך, חלקם היחסי בוקטור C יהיה נמוך גם כן).

$$(3) C_i = \sum_{j=1}^{T_x} \alpha_{ij} * h_j$$

משווה 3 - בניית וקטור הקשר הדינמי Ci

הקלט של המפענה הינו שקלול של המצב הפנימי הקודם שלו, וקטור ההקשר, ומוצא הפלט הקיים המונחים לתוך הרשות בדומה לרשות איטרטיביות קודמות שראיתנו (LSTM\RNNS). פונקציית השקלול מרכיבת מפונקציות האקטיבציה הפנימיות של בלוק GRU (בדומה ל LSTM) ישם שעריו שכחה ועדכו. בניית פלט המפענה מתוארת באירור 6.

אם נחזר לאיור 1, נוכל לראות כי לאורך רוב האלכסון, רק מילה אחת מהפלט מיושרת (קשורה) באופן מובהק למילה בקלט, ולכן המצב הפנימי מהמktorד המתאים יעבור כמעט בשלמותו לווקטור ההקשר Ci (נוסחה 3). לעומת זאת, במקרים שבהם מילה בפלט תלויה (בהתאם למשקלות תשומת הלב) בכמה מילים מהקלט, וקטור ההקשר יהיה מורכב מסכם משקלל של מצבים פנימיים של המktorד. אייר 6 מציג את הארכיטקטורה כפי שהיא מובאת במאמר בשלמותה.



אייר 6 - תיאור הארכיטקטורה בשלמותה.

### או במה מנגן תשומת הלב מסויע לנו?

מלבד יכולת לנגרום למפענה להתרכז בקלט מסוים בעת החיזוי של המילה הנוכחי, מנגן תשומת הלב עובד באופן דומה למנגן skip connection ברשתות עמוקות. אנו מספקים נישה ישירה בין המצביעים הפנימיים של המktorד למפענה באמצעות הוקטור הדינמי שהוא יוצרם, ובכך מאפשרים למידע הקיים בהם "לזרום" כמעט ללא שינוי בהתאם לפעולת skip connection. וזאת בשונה מארQUITקטורות מקודד-מפענה ללא מנגן זה שחלק המידע אובד בעת יצירת הוקטור ביציאת המktorד. מנגן זה נותן מענה לשתי החסכנות המרכזיות של הרשות איטרטיביות: צוואר הבקבוק של ייצוג סדרות דאטה ארוכות ודעתית הנדריאנט. בנוסף, מנגן תשומת הלב מאפשר "פרשנות" (Interpretability) טוביה יותר לרשות.

פרשנות היא מושג המתאר את יכולת שלנו בטור בני אדם להבין את התהליכים המתרחשים בתוך רשות לומדות. בהקשר של הארכיטקטורות שמשתמשות במנגנון תשומת לב, ניתן להבין כיצד המודל מייצר את הפלט באמצעות המשקל שכלי ייחידת מידע קיבלה, ומtower כך ללמידה את מגבלות הרשות ולמצוא דרכי לשפר אותה.

### מה היי החסרונות של הארכיטקטורה שראינו עד כה?

אף על פי שהאררכיטקטורה שהצנו היזומה התקדמות עצומה בתחום ניתוח שפה טبيعית, היי לה מספר מנובלות.

**המגבלה הראשונה** נבעה מכמות המשאים שנדרשו לחישוב וקטור ההקשר C, שהוסיף על עומס החישוב הקיים גם כך בנסיבות איטרטיות. כתע נדרשות ( $m * n$ ) הפעולות של פונקציית היישור alignment) כאשר  $m$  מיצג את מספר הtokנים בקלט ו- $n$  הוא מספר הtokנים בפלט. דבר זה גרם לזמן אימון והסקה (inference) ארוכים במקודם.

**המגבלה השנייה** של הרשות הינה ייצוג ההקשר המוגבל שלה (limited context representation). הרשות לומדת את הקשר בין קלט לפולט. אולם, היא אינה לומדת את ההקשרים שישידות המידע יוצרות אחת עם השניה (تلויות פנים) בקלט ובפלט. לכן, היא לא יכולה "להבין" סמנטיקה מורכבת כגון סLANG, סרקום, כפלו שימושות, ומערכות יחסים עיקריות החוביות בקלט (שבTower בני אדם אנו מבינים בקלות). דבר זה מוביל לכך שביצועי הרשות פוחתים ביחס לשיר לאורך הסדרה.

### כיצד ניתן לפתור בעיה זו?

מנגנון תשומת הלב אותו תיארנו עד כה נקרא תשומת לב מוצלבת (cross attention). Skn, היא (תשומת הלב) מצליבה את המידע מהרקע עם זה שהתקבל מהמפענה. עתה, נציג מנגנון תשומת לב נוסף, **תשומת לב עצמית (self attention)** שהוזג לראשונה במאמר זה. תשומת לב עצמית מחפשת את עצמת הקשר בין כל tokן עם כל tokן אחר באותה סדרה. נזכיר, שבאררכיטקטורות מוקוד-משמעות, גם המוקוד וגם המפענה מקבלים סדרה במהלך האימון, ולכן ניתן לשבע את מנגנון תשומת הלב העצמית בכל אחד מהם. בהקשר זה **אחד החידושים המרכזיים של ארכיטקטורת הטרנספורמרים היה השימוש של שני מנגנונים אלו.**

### אז מדוע תשומת לב עצמית נדרשת מלכתחילה?

כאמור, אחת המוגבלות של הארכיטקטורה הקודמת שהצנו הייתה יכולת מוגבלת לעבד טקסטים מורכבים. מכיוון שהרשות חיפשה **זק את עצמת הקשר** שבין מילה בשפה A' לקבוצת המילים בשפה B' הדרישות לתרגם שלה (או במילים אחרות בנתה את וקטור ההקשר רק מתוך המצביעים הפנימיים של המקודד). אולם, כאשר אנו נינשים למשימת תרגום, אנו צריכים למפות את התလית המורכבות במשפט המקורי על מנת לתרגם בצורה נכונה. על ידי שימוש של מנגנון תשומת הלב העצמית, ניתן לשקל את הקשר שבין מצב פנימי אחד לאחר, בנוסף על בוחנת הקשר שבין מצב פנימי במקודד למפענה. נמחיש בעיה זו באמצעות דוגמא המצינה תרגום מאנגלית לעברית, ונראה כיצד ארכיטקטורת המשלבת את שני מנגנוני תשומת הלב תתרגם את המשפט לעומת ארכיטקטורה עם מנגנון תשומת לב מוצלבת בלבד.

נניח והמשפט אותו אנו מעוניינים לתרגם הוא:

*"Despite the stormy weather causing some delays, the couple, who were accompanied by their close friends, managed to reach the mountaintop and enjoy the breathtaking view."*

אררכיטקטורה משתמשת בשני המנגנונים תתרגם את המשפט באופן הבא:

*"למרות האיזורם שנגרם מזג האוויר הסוער, הזוג, שלויהם אותו חברים קרובים, הצליח להגיע לפסגת ההר ולהנוף מהנוף ועצר נשימה"*

לעומת זאת ארכיטקטורה משתמשת רק במנגנון תשומת הלב המוצלבת עלולה לתרגם את המשפט כך:

"**זהו, שלוו אותו חברים קרובים, הצליח להגיע לפסקת ההר, למרות האחים ומוג האויר הסוער, ולהנוט מהנוּ עוצר הנשימה.**"

למרות שני התרנגולים קוורנטיים ושמרו על כללי תחביר ודקוווק, התרנגול השני נכשל בהבנת הקשר שבין האחים למוג האויר. לעומת זאת, הארכיטקטורה שכנן משתמשת בתשומת לב עצמה הצליחה למצוא קשר זה, ולהביא אותו לידי ביטוי בתרגום.

### **נקודות לסיכום הפרק:**

- ישנם שני סוגים של מנגנון תשומת לב. תשומת לב עצמת בוחנת את עצמת הקשר בין יחידות מידע בתוך הקלט. לעומת זאת, תשומת לב מוצלבת בוחנת את הקשר שבין יחידות מידע בקלט אל מול יחידת מידע בפלט.
- המאמר הראשון שהציג את השימוש במנגנון תשומת לב עבור יישומים של עיבוד שפה טבעית השתמש בארכיטקטורת מקודד-מפענה, ומנגנון תשומת לב שחבר את המפענה למקודד באמצעות וקטור הקשר דינמי.
- וקטור הקשר מתחשב בעוצמת הקשר שבין יחידת פלט הבנייתית קרנגי למפענה (ומייצנת על ידי המצביע הפנימי  $i \in S$ ) לבין כל היחידות של הקלט (המצביעים הפנימיים של המקודד). כאשר היחיד מידע מסוימת במקודד מתבררת כבעל חשבונות גדולה עבור יחידת פלט זו, היא מקבלת משקל גבוה יותר בבניית וקטור הקשר שלה (יחידת פלט).
- אררכיטקטורת הטרנספורמרים הייתה הראשונה לשלב את שני המנגנונים תשומת הלב (מוצלבת עצמית) על מנת לאפשר "העברת" תלוויות מורכבות מהמקודד למפענה הנחוצות לייצור של תרגום איותי.

### **או מה צפוי לנו בפרק הבא?**

עד עכשוי ניתחנו את מנגנון תשומת הלב, כתע אנחנו מתפנים למשימה שלשמה התכונסנו, והיא ניתוח עמוק ומקיף של ארכיטקטורת הטרנספורמרים. מכיוון שהאררכיטקטורה בנויה ממספר ריעונות השוואים אחד בשני, אני וMICHAEL נפרק אותו בשיטת "בבשקות של קופסאות שחורות". בשיטה זו אנו נסתכל על הארכיטקטורה כסט של קופסאות שחורות אחת בתוך השנייה, כאשר כל בבשקה תיזכר רמת אבסטרקציה נוספת של הרשת. בהתאם נסביר את הארכיטקטורה כקובוס שחורה אחת גדולה, בכל שלב נפתחת מכסה המנווע ונציג את מרכיבי השכבה גם כן כקובסאות שחורות עד שנגען לאבני הבניין הבסיסיים של הארכיטקטורה.

**שווה לחכות!**