

מנגנון תשומת הלב לפני עידן הטרנספורמרים

לאחר שראינו כיצד רשתות איטרטיביות כגון RNN/LSTM ממדלות את התלויות בקלט סידרתי, נציג כעת את המנגנון המהווה את ליבה 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 התפתח במכוונות לומדות?

מהו מנגנון תשומת הלב ?

לפני שנגדיר מהי תשומת לב במכוונות לומדות, נבחן כיצד היא באה לידי ביטוי בקוגניציה האנושית. תשומת לב הינה התמקדות סלקטיבית בחלקים הרלוונטיים ביותר של מידע שאנו חווים וסינון חלקים בעלי חשיבות פחותה. תשומת הלב מאפשרת לנו למקד את המשאבים העומדים לרשותינו בצורה יעילה בכל סיטואציה, ובכך אנו יכולים להתמצא ולהבין טוב יותר את מלוא האינפורמציה הזמינה לנו. מנגנון תשומת הלב הוא יכולת מולדת של בני אדם, אולם היא גם כן יכולת נלמדת הניתנת לשיפור במהלך חיינו, (לדוגמאה [מיינדפולנס](#)).

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 מקדים וסינון המידע על ידי המוח.
3. בחירת החלקים החשובים ביותר של המידע בהתאם למידע קודם וההקשר הנוכחי.
4. המידע שנבחר כרלוונטי עובר עיבוד, ובסיום נשמר בזיכרון.

מנגנון תשומת הלב למעשה מגן עלינו בתור בני אדם, מכיוון שהוא מאפשר לנו **להתעלם** ממידע שאינו חיוני לנו (המקיף אותנו הרבה יותר ממידע 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 שמדבר. המוח שלנו מסייע לנו למקד את תשומת הלב שלנו באדם זה ולהתעלם משאר הקולות שהופכים לרעש רקע.

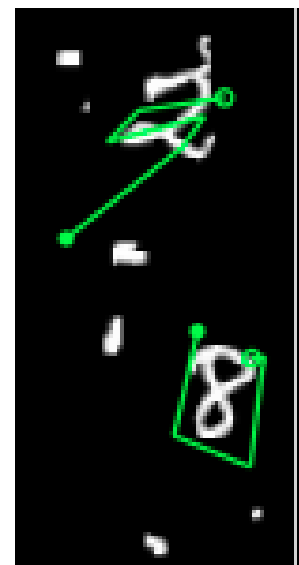
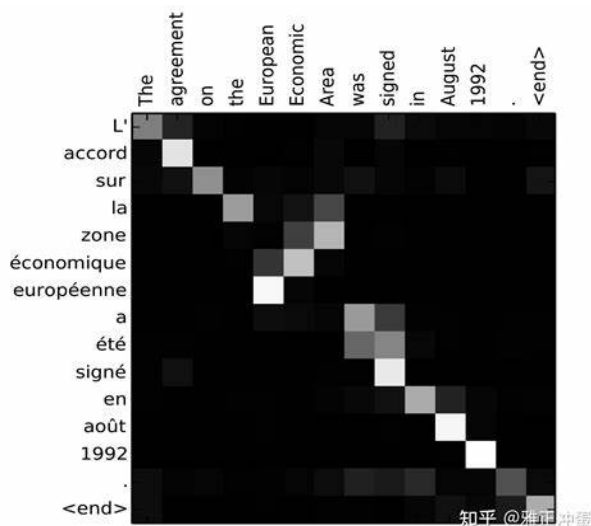
הנקודה המזקקת את העיקרון של תשומת הלב הינה היכולת לשערך מידע (information assessment). במילים אחרות אנו **לומדים להעניק משקל** לכל פיסת מידע, ביחס לכל פיסת מידע אחרת, בהתבסס על החשיבות שלה, והקשר שלה עם מידע שאנו כבר יודעים (הנאגר בזכרון), או חווים ברגע נתון.

אז מהו החידוש שמנגנון תשומת הלב מביא איתו כאשר הוא משולב ברשתות נוירונים? התשובה לכך היא שמנגנון תשומת הלב הוא **פונקציה הנלמדת** כחלק מתהליך אימון הרשת, שמטרתה לחשב מהו המשקל היחסי של כל יחידת דאטה בהינתן הקשר ומידע שנאגר עד לאותו הרגע. מנגנון תשומת הלב יוצר ייצוג וקטורי רציף תלוי-הקשר (contextualized representation) עבור יחידת המידע (כגון טוקן או מילה), ותוצאתו הינה פונקציה רציפה שניתן לגזור אותה ביחס לפרמטרים של מנגנון תשומת הלב (כלומר soft attention). מנגנון תשומת הלב משכן יחידות דאטה במרחב וקטורי כאשר ייצוג של יחידת דאטה במרחב זה מתחשבת בעוצמת הקשר **הרציפה (לא דיסקרטית)** בינה לבין שאר היחידות הדאטה. מאחר והמשקול של יחידות דאטה הינה פונקציה גזירה ביחס לפרמטרים של מנגנון תשומת הלב (וגם ביחס לייצוג הקלט של מנגנון זה) אנו יכולים לאמן את הרשת לייצג את הקשרים האמיתיים בין בין החלקים השונים של הטקסט.

איור 1 (שמאל) המציג את מפת תשומת הלב של משפט באנגלית ביחס לתרגומו בצרפתית, ממחיש את עקרון הרציפות של פונקציית תשומת הלב. האיור מציג משקול של הטוקנים בסדרה הראשונה ביחס לכל הטוקנים בסדרה השנייה (מנגנון זה נקרא תשומת הלב המוצלבת אשר נרחיב עליו בהמשך). איור זה לקוח [מהמאמר](#) שהציג את מנגנון תשומת הלב לראשונה עבור משימות שפה טבעית. נדון במאמר זה בפרק העוסק בתשובה לשאלה "כיצד מנגנון תשומת הלב התפתח במכונות לומדות?".

אולם, לא כל מנגנוני תשומת הלב נולדו שווים. [מאמר זה](#), שהציג את השימוש הראשון במנגנון תשומת הלב (במאמר הופיעה לראשונה המונח attention), עבור סיווג תמונות, השתמש במנגנון תשומת הלב דיסקרטי (hard attention). אופן פעולת הרשת דומה למשחק חיבור נקודות באמצעות קווים היוצרות צורה. הרשת מחפשת את הפיקסלים המקיפים את את האזור הרלוונטי ביותר בתמונה (איור 1 מימין), כך שאם נחבר אותם בקווים ישירים, נקבל תמונה חדשה המכילה את המידע החשוב ביותר הדרוש לסיווג התמונה. בכל הפעלה של הרשת, מנגנון תשומת הלב בוחר פיקסל חדש, ומוסיף אותו לפיקסלים שבחר באיטרציות הקודמות, כאשר בחירת הפיקסלים הללו יוצרת בסופו של דבר את האזור התחום. מכיוון שפונקציית תשומת הלב הדיסקרטית איננה גזירה (ומקבלת ערכים דיסקרטיים בלבד) לא ניתן למטב (optimize) אותה עם שיטות ממשפחת מורד הגרדיאנט (gradient descent). עקב כך המאמר עשה שימוש בשיטת אימון משטר מבוסס החלטות (on policy) השאולה מתוך עולם הלמידה מבוססת החיזוקים (reinforcement learning), שלא דורשת גזירות של פונקציית מטרה, על מנת לאמן את מנגנון תשומת הלב.

בעולם ניתוח השפה, מנגנון תשומת הלב הדיסקרטי, ילמד למצוא את הטוקנים החשובים ביותר (בהתאם למשימה) לבנייה של וקטור הייצוג עבור טוקן נתון. לדוגמא, נניח והמטרה של הרשת היא למצוא את האובייקט החשוב ביותר במשפט "The cat is playing with the toy, it is soft". הרשת 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 באופן יזום. לעומת זאת, ישנו מנגנון תשומת הלב נוסף הנקרא תשומת לב "לא

מפורשת" (implicit attention). מנגנון זה הינו "תוצר לא מכוון" של רשתות עמוקות. רשתות אלו נוטות להתמקד בחלקים מסוימים של המידע ולהתעלם מאחרים. לדוגמא, בסיווג של מנח גוף מתוך סרטונים (pose estimation), תהייה לרשת הנטייה להתמקד באזורים שבהם מופיעים חלקי גוף ולהתעלם מאזורים בהם לא מופיע אדם בכלל (רקע או אובייקטים דוממים). ניתן להמחיש באופן ויזואלי את תשומת לב המרומזת על ידי איור 2, המגיע [מההרצאה הבאה](#). האיור ממחיש את מיקוד הרשת בעת ניתוח תמונה עבור אימון סוכן במשחק. הסוכן לומד לנסוע בתוך השביל ולהימנע ממפגעים באמצעות למידה מבוססת חיזוקים. החלקים הבוהקים באדום מייצגים את האזורים בהם הרשת מתמקדת על מנת לקבל את ההחלטה הבאה. מכיוון שהתקדמות בשביל מובילה לעלייה בנקודות הרשת מתמקדת באופק, ובלוח התוצאה שמציג את הניקוד. **מעתה, בכל מקום בטקסט שנתייחס לתשומת לב, נתכוון לתשומת לב מפורשת.**



איור 2 - תשומת לב מרומזת ברשת מבוססת חיזוקים

מדוע מנגנון תשומת הלב נדרש?

כפי שהסברנו בפרקים הקודמים, ארכיטקטורות איטרטיביות סבלו מבעיה מרכזית המשותפת לכולן, והיא רכיבי זיכרון הקבועים בגודלם, ומאידך, קלט בעל אורך משתנה. כתוצאה מכך אנו נאלצים לקודד משפטים באורכים שונים ולקטור בגודל קבוע. ולכן, בעת קידוד קלטים העולים על גודל מסוים, נתכנס לבעיית מידול של תלויות ארוכות הטווח. עיקר הבעיה בא לידי ביטוי בכך שלא ניתן להשתמש בכל יחידות הקלט באופן מפורש לבניה של וקטור ההקשר h_t עבור יחידת דאטה i . במילים פשוטות, מטרננו לאפשר ולקטור המקודד את הקלט לגשת לכל חלקיו (של הקלט) במקביל בעת בניית הייצוג.

על מנת להמחיש את הנושא, נסתכל על הפסקה הבאה:

"צח, מהנדס תוכנה, עובד מהבית בשנתיים האחרונות. הוא מתגורר ביישוב קטן בצפון הארץ עם אשתו ושני ילדיו. היישוב שקט ורגוע ויש בו תחושת קהילתיות. צח נהנה לבלות עם משפחתו ולצאת לטיולים ארוכים ביער הסמוך. הוא מעריך את הגמישות שעבודה מרחוק מציעה לו. בשבוע שעבר גילה צח כי החברה שלו מתכננת ליישם מדיניות חדשה שתחייב את כל העובדים לעבוד מהמשרד. צח שוקל כעת האם לעבור לגור בקרבת המשרד שנמצא במרכז הארץ או לחפש עבודה חדשה".

מודל ניתוח וסיכום טקסט המבוסס RNN או LSTM עשוי להתקשות להבין מהו המידע החיוני בטקסט זה וליצור סיכום תמציתי איכותי. מכיוון שאופן עיבוד הקלט הוא טוקן-אחריי-טוקן, תוצאה אפשרית של מודל איטרטיבי יכולה להיות:

"צח, מהנדס תוכנה, שעובד מרחוק, עובר לגור ליד המשרד, או מחפש עבודה חדשה, בגלל שינוי במדיניות החברה"

למרות שהרשת אכן "דחסה" את כל המידע החשוב, עדיין חסרה נהירות (קוהרנטיות) בפלט.

לעומת זאת, רשת המקיימת את התנאים הבאים:

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

יכולה לספק את התמצות הבא:

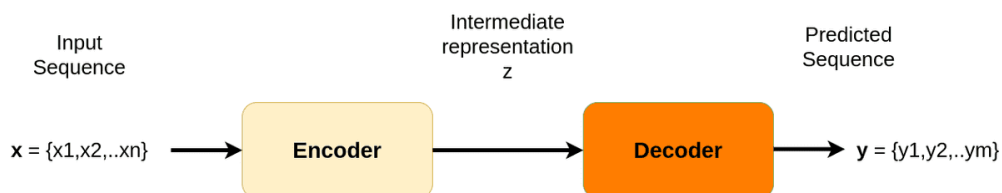
"צח, מהנדס תוכנה שעובד מרחוק, שוקל האם לעבור לעיר הקרובה לעבודתו במשרד או למצוא משרה חדשה המאפשרת עבודה מרחוק בשל מדיניות החדשה בחברה"

נקודות לסיכום הפרק:

- מנגנון תשומת הלב במכונות לומדות הינו פונקציה נלמדת, השואבת השראה מתשומת הלב בקוגניציה האנושית, וממשקלת חשיבות של קלט ביחס לקלט אחר.
- מנגנון תשומת הלב שאנו דנים בו נקרא soft attention שמהווה פונקציה רציפה. תכונה זו נובעת מהעובדה שמשקול של יחידת דאטה נמדד ביחס לכל יחידות הדאטה האחרות.
- ישנן שני סוגים של מנגנוני תשומת לב. תשומת לב מפורשת, הינה פונקציית תשומת לב הממומשת באופן יזום כחלק מארכיטקטות המודל. לעומת זאת, תשומת לב לא מפורשת הינה תוצר של עיבוד דאטה על ידי רשתות עמוקות, הלומדות חשיבות של אזורים מסוימים בקלט ללא הכוונה יזומה.
- רשתות בעלות מנגנון תשומת לב יכולות ללמוד קשרים מורכבים יותר בקלט ולייצר פלט קוהרנטי יותר ביחס לרשתות איטרטיביות.

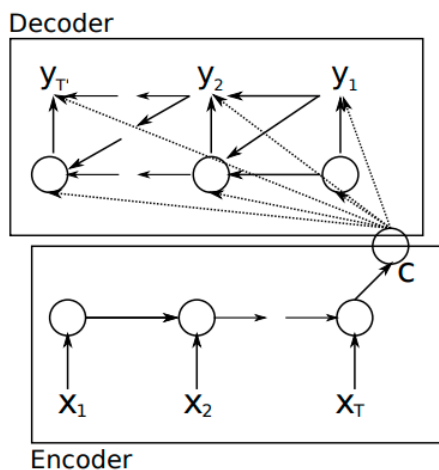
כיצד מנגנון תשומת הלב התפתח במכונות לומדות ?

[המאמר הראשון](#) שהציג שימוש במנגנון תשומת הלב עבור משימות של ניתוח שפה טבעית השתמש בארכיטקטורת מקודד-מפענח (encoder-decoder architecture), המתוארת באיור 3, לטובת תרגום מאנגלית לצרפתית. על מנת להסביר את הצורך במנגנון תשומת הלב, נציג כעת מושג חשוב בתחום עיבוד השפה הטבעית הנקרא "יישור" (alignment). מושג זה מתאר בהקשר של תרגום, את התאימות בין מילה/מילים משפת המקור לבין מילים בשפת היעד (איור 1 משמאל מדגים את היישור בין מאנגלית לצרפתית). במילים אחרות, זהו ייצוג של "עוצמת הקשר" בין קבוצות של מילים בשפת היעד לבין מילים בשפת המקור.

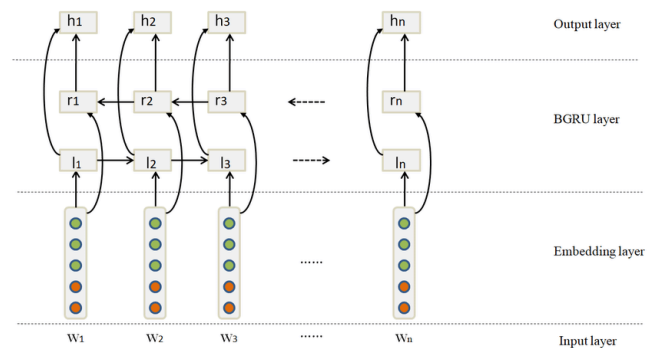


איור 3 - הפשטה של ארכיטקטורת מקודד-מפענח

כאן המקודד מקבל משפט (או קטע של טקסט) בשפה א' כסדרת טוקנים, ומייצג את המידע במשפט כוקטור במימד חבוי (latent representation). מן הצד השני, המפענח מחליץ מתוך הוקטור שהתקבל את המידע הרלוונטי ומפיק מתוכו את המשפט בשפה ב'. הרשת מאומנת כמקשה אחת, ולומדת לקודד ולפענח באותו הזמן. במאמר גם המפענח וגם המקודד מומשו על ידי יחידות GRU משורשרות (אולם ניתן להשתמש גם ב-RNN או LSTM), כאשר המקודד מורכב מיחידות GRU דו-כיווניות ([bidirectional GRU](#)). ארכיטקטורת GRU דו-כיוונית מורכבת משני שרשרים של יחידות GRU כאשר השרשור הראשון משמש להעברת הקלט מתחילתו לסופו, והשרשור השני של יחידות ה-GRU משמש להעברת הקלט מהסוף להתחלה (ראה איור 4). הסיבה לשימוש בארכיטקטורה דו-כיוונית נובע מכך שאנו מעוניינים לקבל מידע מקיף על הקלט, כלומר גם מתחילתו וגם מסופו, שכן כל המידע (משפט המקור) נתון לנו בזמן הפעלת המודל לאחר האימון (inference). לעומת זאת הפלט במפענח נוצר באופן אוטורגרסיבי (מילה אחרי מילה, כאשר פלט נוכחי הופך לקלט עתידי לאחר שנוצר) בזמן ההסקה, שהופך את השימוש ברשת דו-כיוונית במפענח למהלך משולל היגיון.



איור 5 - ארכיטקטורת מקודד מפענח



איור 4 - GRU דו-כיוונית

בארכיטקטורות מסוג מקודד-מפענח שקדמו למאמר, הקלט שהמפענח מקבל בכל איטרציה (יצירת יחידת דאטה חדשה) הינו המצב הפנימי H_i והפלט Y_{i-1} מהאיטרציה הקודמת (של המפענח). על מנת לחבר את המידע שהמקודד למד מהקלט, המפענח מקבל בנוסף את וקטור המוצא של המקודד שנקרא לו מעתה C . בארכיטקטורות מקודד-מפענח בסיסית, C הינו שרשור המצבים הפנימיים שחושבו באיטרציה האחרונה מכל אחת מהרשתות המרכיבות את הרשת הדו-כיוונית במקודד ($h_i = [\rightarrow h_i; \leftarrow h_i]$) שכן הוא מכיל מידע על כל הקלט. איור 5 מציג את הארכיטקטורה שהסברנו בפסקה זאת.

כיצד בא לידי ביטוי מנגנון תשומת הלב במאמר?

כפי שהזכרנו קודם לכן, השימוש בתשומת הלב מיועד לפתור את בעיית היישור בין הקלט לפלט. עקב אכילס של הארכיטקטורות שקדמו לזו של המאמר, הייתה שימוש במצבים הפנימיים האחרונים של המקודד (הפלטם של האיטרציה האחרונה של משני הכיוונים של BGRU). מכיוון שמצבים אלו הכילו מידע דחוס על כל הקלט, לא ניתן היה למדל את התלויות המקומיות בין המצבים הפנימיים של המפענח, לאלו של המקודד. כלומר, וקטורי ההקשר המתקבלים בכניסת המפענח דוחסים את כל המידע מסדרת הקלט של המקודד, ללא התחשבות בקשר שבין יחידות מידע בסדרה א' ליחידת מידע הנבנת בסדרה ב'. לעומת זאת, בשיטה המוצעת במאמר, וקטור ההקשר שהמפענח מייצר בעת בניית יחידת פלט i , מקבל את המידע על כל יחידות הדאטה של הקלט שנבנו על ידי המקודד. וקטור ההקשר נוצר כסכום משוקלל של כל המצבים הפנימיים של המקודד, כאשר המשקלים ממדלים את הקשרים בין כל יחידות הקלט ליחידת פלט i .

מנגנון תשומת הלב: המידול המתמטי

המושג הראשון שהמאמר מגדיר הינו **עוצמת היישור** (alignment score), שמייצג את הקשר בין המצב הפנימי $i-1$ במפענח לבין מצב פנימי j כלשהו במקודד. על מנת למנוע בלבול, נגדיר את המצבים הפנימיים של המפענח כ- s (כפי שהוא מובא במאמר) ואת המצבים הפנימיים במקודד נשאיר כ- h . האינדקסים i, j מייצגים את המספר הסידורי של יחידות הקלט והפלט ה- j, i . כעת עבור יחידת פלט i נגדיר וקטור $e_{ij}, j = 1, \dots, T$ 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 המקודד.
- s_{i-1} - המצב הפנימי מהיחידה $i-1$ של המפענח.
- W - מטריצת המשקלות של מנגנון תשומת הלב.
- v_a - וקטור המשקל של פונקציית תשומת הלב.
- T - מספר יחידות הדאטה במקודד.

כאמור, מנגנון עוצמת היישור הינו פונקציה נלמדת המחשבת את עוצמת הקשר שבין המצב הפנימי של המפענח למצבים הפנימיים של המקודד. מכיוון שהמכפלות בתוך פונקציית הטנגנס ההיפרבולית \tanh יוצרים וקטור בגודל 1x, וערך תשומת הלב בין שתי יחידות דאטה צריך להיות סקלר, המכפלה בוקטור v_a יוצרת סקלר במוצא. נשים לב כי המטריצה W והוקטור v_a הינם פרמטרים הנלמדים (מאומנים) של המודל.

המושג השני שהמאמר מגדיר הינו **משקולת תשומת הלב** (attention weight). מטרת מנגנון תשומת הלב הינה ליצור משקול חשיבות של טוקן אל מול כל טוקן אחר, כפונקציית רציפה וגזירה. המשמעות של רציפות בהקשר שאנו מדברים עליו, הינו משקול של עוצמת הקשר e_{ij} (המקושרת למצב הפנימי h_i מהמקודד) ביחס לכל עוצמות הקשר האחרות. עוצמות אלו מייצגות את הקשר שבין כל שאר המצבים הפנימיים של המקודד ביחס למצב פנימי הנתון של המפענח. על מנת לעשות זאת, אנו משתמשים בפונקציית softmax המופעלת על עוצמות הקשר. חישוב זה למעשה פותר את בעיית היישור שפתחנו איתה את הפרק, מכיוון שפונקציית ה-softmax תהפוך את תוצאת היישור, שעמדה בפני עצמה, להיות פונקציית צפיפות הסתברות התלוייה בכל המצבים הפנימיים של המקודד. אנו מבצעים פעולה זו עבור כל מצב פנימי של המקודד אל מול אותו מצב פנימי של המפענח, ובכך מקבלים את עוצמת הקשר הרציפה שהזכרנו.

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

משוואה 2 - חישוב משקל תשומת הלב עבור זוג יחידות דאטה i, j

כאן T_x הוא מספר יחידות חישוב במקודד כלומר אורך מקסימלי של סדרת דאטה שניתן להכניס בו כמקשה אחת.

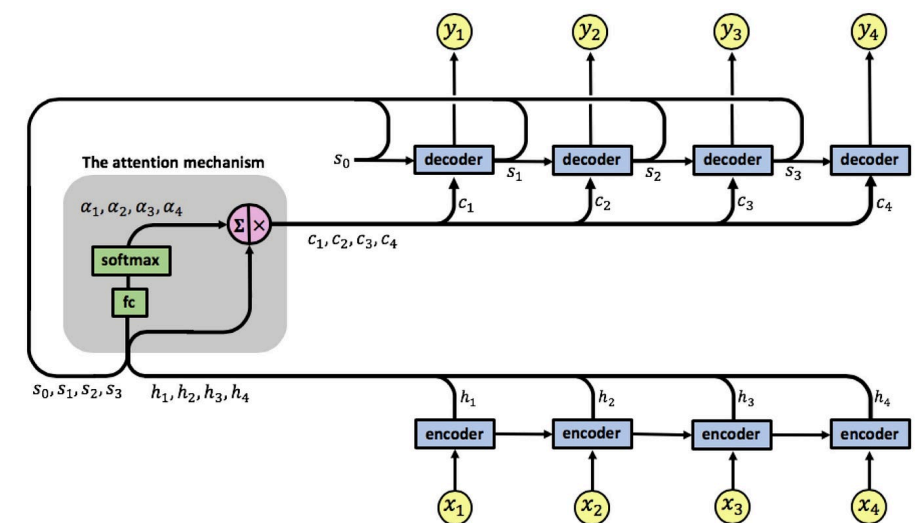
החלק האחרון בהסבר שלנו הוא **בניית וקטור הקשר דינמי C**. אנו משתמשים במשקולות תשומת הלב שחישבנו, ומרכיבים את הוקטור כמכפלה של משקולות זו במצב הפנימי של המקודד המקושר אליה. בנייה זו של הוקטור ההקשר מאפשרת לנו לתת למפענח את המידע הרלוונטי ביותר ביחס למצב הפנימי הנוכחי שלו. מצבים פנימיים של המקודד בעלי עוצמת קשר נמוכה ביחס למצב הפנימי של המפענח, לא ישפיעו על הוקטור C (שכן אם ערך משקולת תשומת הלב שלהם נמוך, חלקם היחסי בוקטור C יהיה נמוך גם כן).

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

משוואה 3 - בניית וקטור הקשר הדינמי C_i

הקלט של המפענח הינו שקלול של המצב הפנימי הקודם שלו, וקטור ההקשר, ומוצא הפלט הקודם המוזנים לתוך הרשת בדומה לרשתות איטרטיביות קודמות שראינו (LSTM/RNN). פונקציית השקלול מורכבת מפונקציות האקטיבציה הפנימיות של בלוק ה-GRU (בדומה ל-LSTM ישנם שערי שכחה ועדכון). בניית פלט המפענח מתוארת באיור 6.

אם נחזור לאיור 1, נוכל לראות כי לאורך רוב האלכסון, רק מילה אחת מהפלט מיושרת (קשורה) באופן מובהק למילה בקלט, ולכן המצב הפנימי מהמקודד המתאים יועבר כמעט בשלמותו לווקטור ההקשר C_i (נוסחה 3). לעומת זאת, בחלקים שבהם מילה בפלט תלויה (בהתאם למשקולות תשומת הלב) בכמה מילים מהקלט, וקטור ההקשר יהיה מורכב מסכום משוקלל של מצבים פנימיים של המקודד. איור 6 מציג את הארכיטקטורה כפי שהיא מובאת במאמר בשלמותה.



איור 6 - תיאור הארכיטקטורה בשלמותה.

אז כמה מנגנון תשומת הלב מסייע לנו?

מלבד היכולת לגרום למפענח להתרכז בקלט מסוים בעת החיזוי של המילה הנוכחית, מנגנון תשומת הלב עובד באופן דומה למנגנון skip connection ברשתות עמוקות. אנו מספקים גישה ישירה בין המצבים הפנימיים של המקודד למפענח באמצעות הוקטור הדינמי שאנו יוצרים, ובכך מאפשרים למידע הקיים בהם "לזרום" כמעט ללא שינוי בדומה לאופן פעולת skip connection. וזאת בשונה מארכיטקטורות מקודד-מפענח ללא מנגנון זה שחלק מהמידע אובד בעת יצירת הוקטור ביציאת המקודד. מנגנון זה נותן מענה לשתי החסרונות המרכזיים של הרשתות האיטרטיביות: צוואר הבקבוק של ייצוג סדרות דאטה ארוכות ודעיכת הגרדיאנט. בנוסף, מנגנון תשומת הלב מאפשר "פרשנות" (Interpretability) טובה יותר לרשת.

פרשנות היא מושג המתאר את היכולת שלנו בתור בני אדם להבין את התהליכים המתרחשים בתוך רשתות לומדות. בהקשר של הארכיטקטורות שמשמשות במנגנון תשומת לב, ניתן להבין כיצד המודל מייצר את הפלט באמצעות המשקל שכל יחידת מידע קיבלה, ומתוך כך ללמוד את מגבלות הרשת ולמצוא דרכים לשפר אותה.

מה היו החסרונות של הארכיטקטורה שראינו עד כה?

אף על פי שהארכיטקטורה שהצגנו היוותה התקדמות עצומה בתחום ניתוח שפה טבעית, היו לה מספר מגבלות.

המגבלה הראשונה נבעה מכמות המשאבים שנדרשו לחישוב וקטור ההקשר C , שהוסיף על עומס החישוב הקיים גם כך ברשתות איטרטיביות. כעת נדרשות $O(n * m)$ הפעולות של פונקציית היישור (alignment score) כאשר m מייצג את מספר הטוקנים בקלט ו- n הוא מספר הטוקנים בפלט. דבר זה גרם לזמני אימון והסקה (inference) ארוכים במקודד.

המגבלה השנייה של הרשת הינה ייצוג ההקשר המוגבל שלה (limited context representation). הרשת לומדת את הקשר בין קלט לפלט. אולם, היא אינה לומדת את ההקשרים שיחידות המידע יוצרות אחת עם השנייה (תלויות פנים) בקלט ובפלט. לכן, היא לא יכולה "להבין" סמנטיקה מורכבת כגון סלנג, סרקזם, כפל משמעות, ומערכות יחסים עקיפות החבויות בקלט (שבתור בני אדם אנו מבינים בקלות). דבר זה מוביל לכך שביצועי הרשת פוחתים ביחס ישיר לאורך הסדרה.

כיצד ניתן לפתור בעיה זו?

מנגנון תשומת הלב אותו תיארנו עד כה נקרא תשומת לב מוצלבת (cross attention). שכן, היא (תשומת הלב) מצליבה את הידע מהמקודד עם זה שהתקבל מהמפענח. עתה, נציג מנגנון תשומת לב נוסף, **תשומת לב עצמית (self attention)** שהוצג לראשונה [במאמר זה](#). תשומת לב עצמית מחפשת את עוצמת הקשר בין כל טוקן עם כל טוקן אחר באותה סדרה. נזכיר, שבארכיטקטורות מקודד-מפענח, גם המקודד וגם המפענח מקבלים סדרה במהלך האימון, ולכן ניתן לשבץ את מנגנון תשומת הלב העצמית בכל אחד מהם. בהקשר זה **אחד החידושים המרכזיים של ארכיטקטורת הטרנספורמרים היה השילוב של שני מנגנונים אלו**.

אז מדוע תשומת לב עצמית נדרשת מלכתחילה?

כאמור, אחת המגבלות של הארכיטקטורה הקודמת שהצגנו הייתה יכולת מוגבלת לעבד טקסטים מורכבים. מכיוון שהרשת חיפשה **רק את עוצמת הקשר** שבין מילה בשפה א' לקבוצת המילים בשפה ב' הדרושות לתרגום שלה (או במילים אחרות בנתה את וקטור ההקשר רק מתוך המצבים הפנימיים של המקודד). אולם, כאשר אנו ניגשים למשימת תרגום, אנו צריכים למפות את התלויות המורכבות במשפט המקור על מנת לתרגם בצורה נכונה. על ידי שילוב של מנגנון תשומת הלב העצמית, ניתן לשקול את הקשר שבין מצב פנימי אחד לאחר, בנוסף על בחינת הקשר שבין מצב פנימי במקודד למפענח. נמחיש בעיה זו באמצעות דוגמא המציגה תרגום מאנגלית לעברית, ונראה כיצד ארכיטקטורה 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."

ארכיטקטורה המשתמשת בשני המנגנונים תתרגם את המשפט באופן הבא:

"למרות האיחורים שגרם מזג האוויר הסוער, הזוג, שליוו אותו חברים קרובים, הצליח להגיע לפסגת ההר ולהנות מהנוף עוצר הנשימה"

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

"הזוג, שליוו אותו חברים קרובים, הצליח להגיע לפסגת ההר, למרות האיחורים ומזג האוויר הסוער, ולהנות מהנוף עוצר הנשימה."

למרות ששני התרגומים קוהרנטיים ושמרו על כללי תחביר ודקדוק, התרגום השני נכשל בהבנת הקשר שבין האיחור למזג האוויר. לעומת זאת, הארכיטקטורה שכן משתמשת בתשומת לב עצמית הצליחה למצוא קשר זה, ולהביא אותו לידי ביטוי בתרגום.

נקודות לסיכום הפרק:

- ישנם שני סוגים של מנגנוני תשומת לב. תשומת לב עצמית בוחנת את עוצמת 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-מפענח, ומנגנון תשומת לב שחיבר את המפענח למקודד באמצעות וקטור הקשר דינמי.
- וקטור ההקשר מתחשב בעוצמת הקשר שבין יחידת פלט הנבנית כרגע במפענח (ומיוצגת על ידי המצב הפנימי S_{i-1}) לבין **כל** היחידות של הקלט (המצבים הפנימיים של המקודד). כאשר יחידת מידע מסוימת במקודד מתבררת כבעלת חשיבות גדולה עבור יחידת פלט זו, היא מקבלת משקל גבוה יותר בבניית וקטור ההקשר שלה (יחידת פלט).
- ארכיטקטורת הטרנספורמרים הייתה הראשונה לשלב את שני המנגנונים תשומת הלב (מוצלבת ועצמית) על מנת לאפשר "העברת" תלויות מורכבת מהמקודד למפענח הנחוצות ליצירה של תרגום איכותי.

אז מה צפוי לנו בפרק הבא?

עד עכשיו ניתחנו את מנגנון תשומת הלב, כעת אנחנו מתפנים למשימה שלשמה התכנסנו, והיא ניתוח מעמיק ומקיף של ארכיטקטורת הטרנספורמרים. מכיוון שהארכיטקטורה בנויה ממספר רעיונות השזורים אחד בשני, אני ומיכאל נפרק אותם בשיטת "בבושקות של קופסאות שחורות". בשיטה זו אנו נסתכל על הארכיטקטורה כסט של קופסאות שחורות אחת בתוך השניה, כאשר כל בבושקה תייצג רמת אבסטרקציה נוספת של הרשת. בהתחלה נסביר את הארכיטקטורה כקופסא שחורה אחת גדולה, בכל שלב נפתח את מכסה המנוע ונציג את מרכיבי השכבה גם כן כקופסאות שחורות עד שנגיע לאבני הבניין הבסיסיים של הארכיטקטורה.

שווה לחכות!