



איור 2: בלוק LSTM

כפי שראינו, דעיכת הנדריאנט וכשל בධיסת מידע מונעים מרשת RNN את יכולת התררכז בחלקים החשובים ביותר בסדרה. כתוצאה לכך, הרשות מתנסה לאטר את התלות בין מרכיביה השונים שזו כאמור המטריה העיקרית בעיבוד שפה טبيعית. כיצד ניתן להתגבר על בעיות אלה ברשותות המקובלות ככלט מידע סדרתי?

על מנת שנוכל לענות על שאלת זו, נבחן כיצד בני אדם מתמודדים עם סוג מידע שכזה. נציג הפשטה של הרעיון: בני אדם משתמשים בשני אלמנטים עיקריים: הראשון הוא זיכרון לטווח אורך וקצר, והשני הוא יכולת להפריד בין טפל ועיקר. שני מנגנונים אלו קשורים אחד בשני באופן הפעולה שלהם. כמובן, אנחנו מסוגלים להפריד בין טפל לעיקר בזכות ניסיון עבר והשלכתו על סיטואציה חדשה.

זהו בסיס הרעיון שהרשת הבאה מנסה לישם. [LSTM](#) (ונרגסתה האלטראנטיבית GRU) היא ארכיטקטורה איטרטיבית שבדומה ל-RNN מכילה זיכרון לטווח קצר H , (המצב הפנימי -RNN) ובנוסף, רכיב זיכרון ארוך טווח C_t . בינהן ל-RNN שמתמצחת את המידע שהתקבל עד לנקודת הזמן הנוכחי מיכרזן קצר טווח, ההפרצה לזכרון לטווח קצר ואורך ארוך-ב-LSTM מאפשרת לשמר את מאפייני הקלט המקוריים גם בטווח הארוך וпотורת את בעית צואර הבקבוק של RNN. מבנה הרשות מאפשר לה לשמור את שני סוגי הזיכרונות והמילה החדש בכל איטרציה, ולונתח את המילה החדשה באמצעותם.

ניהול זיכרון מתבצע באמצעות שלושה שערים: שער השכחה, שער הקלט ושער הפלט (איור 2). **שער השכחה** אחראי על הרשות מידע מהזיכרון לטווח הארוך לאחר שזהו התגלה מלבד רלוונטי יותר. **שער הקלט** אחראי על הזנתה המידע לתוך תא הזיכרון. שער זה משתמש במידע המתקבל מהתגלו המקורי הקודם ומקלט חדש, ומחליט מה מהידעת החדש רלוונטי ויש להוסיפו לתא הזיכרון. **שער הפלט** אחראי על חישוב מזיא הרשות $-H$, המהווה למעשה את הזיכרון קצר הטווח.

היתרון של LSTM מتبטא בעיקר במקרים שימושיים ארוכים, לדוגמה, במשפט "She left" תרגום המילה "left" על ידי RNN יהיה "הלה" ולא "הלה" (או "שמאלי") מכיוון שההקשר הוא קרוב. אולם, אם ניקח את המשפט "The goalie was determined to protect the goal" את המילה האחורונה ניתן לתרגם כ"משימה/מטרה" או "שער (כדורגל)", במקרה זה, מילה ההקשר "goalie" (שער) מרוחקת מהמילה שאנו רוצחים לתרגם, ולכן RNN עלולה לטעות בתרגום, במקרה זה יבוא לידי ביטוי הזיכרון ארוך הטווח, שמתבצע במילה "goalie" בתרגום המילה "goal".

מדוע LSTM סובלת פחות מדעיכת הנדריאנט? הסיבה נובעת בכך שרוכב הזיכרון אינו מתעדכן בפלט של פונקציית האקטיבציה כמו ב-RNN, אלא בשימוש במושג שער השכחה ושער הקלט. שער השכחה מסיר חלקים מהקלט על ידי יצירת וקטור של ערכים רציפים במרקוטע של $[0,1]$, כאשר 0 משמעו שכחה מוחלטת, ו-1 מזיאו שימור מוחלט של המידע. הקלט והמצב הפנימי קובעים כמה מידע יש לשוכות, אולם הם אינם קובעים מה יהיה תוכן תא הזיכרון לאחר העדכון. מן הצד השני, הוספה מידע חדש לזכרון אינה מתבצעת כדראסה של התוכן על ידי הקלט, אלא כחיבור שלו. שער הקלט מייצר וקטור שמתווסף למידע הקיים בתא

הזכרון בפעולות חיבור (element-wise addition) ובכך מאפשר את העברת האינפורמציה ברשות ומונע את דעיכת הנדריאנט, מכיוון שפעולה זו פוחות רגישה לשינויים בקלט כפי שמצוין פונקציית האקטיבציה, רגישה אליהם. כתוצאה לכך, הנזירות שאין מושפעות באופן ישיר מפונקציות האקטיבציה ולא מתאפסות, מאפשרות את עדכון המשקלות (בשונה מ-RNN שהנזרות מתאפסות די מהר). בנוסף, מכיוון ששער השכחה מייצר וקטור הקבוע אליו חלקים ברכיב הזיכרון יש לשכך ואילו לשמר, כאשר ערך הוקטור קרוב ל-1, המידע יכול לזרום ברשות מבלי להינזק באופן משמעותי או במילימ אחרות, מתאפשרה שמירת מידע ארוך טווח (מידול מתמטי של נושא זה ניתן למצוא [בקישור](#)). בזכות השימוש ברכיב זיכרון, המצב הפנימי מצטמצם להיות זיכרון לטוויה קצר בלבד והשפעת צוואר הבקבוק שנוצרת ב-RNN פוחתת גם היא.

למרות היתרונות המובהקים של הרשות על פני RNN במשימות בהן ישנו צורך לנתח תלויות ארכוכות טווח, קיימות לה מוגבלות בניתוח שפה טבעית הנובעת ממורכבות הארכיטקטורה ואופן הפעולה האיטרטיבי שלה. הבעיה הראשונה משוטפת לכל הרשותות האיטרטיביות, והוא חסר יכולת להפעיל את הרשות במקביל, דבר הגורר הפעלה איטית של הרשות ואי יכול מספק של מערכי החישוב, בעוד, לא מאפשרת לממד תלויות בין מקטעים שונים של הקלט.

מכיוון שככל תא LSTM מכיל מספר גדול יותר של פרמטרים בהשוואה ל-RNN, אימון והפעלת הרשות דורשים משאבי חישוב (זיכרון וכוח עיבוד) נוספים יותר. בנוסף, על מנת שהרשות תמדו את התלוויות ארכוכות הטווח באופן אפקטיבי ישנו צורך בכמות משמעותית של DATA בעלת מספר TOKENים גבוי (=אוריך הסדרה). בנוסף, משך האימון הוא גם כן חסרון משמעותי של הרשות. חיפוש תלויות ארכוכות טווח בסדרה הינה שימוש מורכבת יותר ביחס לחיפוש דפוסים מקומיים (local patterns) בDATA, ולכן נדרש זמן אימון ארוך יותר.

נקודות מפתח לסיקום הפוסט:

- LSTM הינה ייחdet עיבוד איטרטיבית המכילה רכיב זיכרון שאונר מידע ארוך טווח, בנוסף על המצב הפנימי(הנקרא זיכרון קצר טווח). עדכון רכיבי הזיכרון מתבצע באמצעות 3 שערים, שער השכחה, שער הקלט ושער הפלט.
- הארכיטקטורה של הרשות פותרת את דעיכת הנדריאנט על ידי עדכון רכיבי הזיכרון באמצעות הקלט באופן עקייף, ולא באופן ישיר כמו ב-RNN.
- עקב ריבוי הפרמטרים של הרשות,omidול המידע ארוך הטווח, משך האימון וכמות המשאים הנדרשים לאמן את הרשות מהווים עקב אכילס שלה.