



איור 2: בלוק LSTM

כפי שראינו, דעיכת הגרדיאנט וכשל בדחיסת מידע מונעים מרשת RNN את היכולת להתרכז בחלקים החשובים ביותר בסדרה. כתוצאה מכך, הרשת מתקשה לאתר את התלות בין מרכיבי השונים שזו כאמור המטרה העיקרית בעיבוד שפה טבעית. כיצד ניתן להתגבר על בעיות אלה ברשתות המקבלות כקלט מידע סדרתי?

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

זהו בסיס הרעיון שהרשת הבאה מנסה ליישם. [LSTM](#) (וגרסתה האלטרנטיבית GRU) היא ארכיטקטורה איטרטיבית שבדומה ל-RNN מכילה זיכרון לטווח קצר  $h_t$ , (המצב הפנימי ב-RNN) ובנוסף, רכיב זיכרון ארוך  $c_t$ . בניגוד ל-RNN שמתמצתת את המידע שהתקבל עד לנקודת הזמן הנוכחית כזיכרון קצר טווח, ההפרדה לזיכרון לטווח קצר וארוך ב-LSTM מאפשרת לרשת לשמר את מאפייני הקלט המהותיים גם בטווח הארוך ופותרת את בעיית צוואר הבקבוק של RNN. מבנה הרשת מאפשר לה לשמור את שני סוגי הזכרונות והמילה החדשה בכל איטרציה, ולנתח את המילה החדשה באמצעותם.

ניהול זיכרון מתבצע באמצעות שלושה שערים: שער השכחה, שער הקלט ושער הפלט (איור 2). **שער השכחה** אחראי על הסרת מידע מהזיכרון לטווח הארוך לאחר שזה התגלה כלא רלוונטי יותר. **שער הקלט** אחראי על הזרמת המידע לתוך תא הזיכרון. שער זה משתמש במידע המתקבל מהמצב הפנימי הקודם ומקלט חדש, ומחליט מה מהמידע החדש רלוונטי ויש להוסיפו לתא הזיכרון. **שער הפלט** אחראי על חישוב מוצא הרשת -  $h_t$ , המהווה למעשה את הזיכרון קצר הטווח.

היתרון של LSTM מתבטא בעיקר במשפטים ארוכים, לדוגמה, במשפט "She left" תרגום המילה "left" על ידי RNN תהייה "הלכה" ולא "הלך" (או "שמאל") מכיון שההקשר הוא קרוב. אולם, אם ניקח את המשפט "The goalie was determined to protect the goal" את המילה האחרונה ניתן לתרגם כ"משימה/מטרה" או "שער (כדורגל)", במקרה זה, מילה ההקשר "goalie" (שוער) מרוחקת מהמילה שאנו רוצים לתרגם, ולכן RNN עלולה לטעות בתרגום, במקרה זה יבוא לידי ביטוי הזיכרון ארוך הטווח, שמתחשב במילה "goalie" בתרגום המילה "goal".

מדוע LSTM סובלת פחות מדעיכת הגרדיאנט? הסיבה נובעת מכך שרכיב הזיכרון אינו מתעדכן בפלט של פונקציית האקטיבציה כמו ב-RNN, אלא בשימוש במוצא שער השכחה ושער הקלט. שער השכחה מסיר חלקים מהקלט על ידי יצירת וקטור של ערכים רציפים במקטע של  $[0,1]$ , כאשר 0 משמעו שכחה מוחלטת, ו-1 מהווה שימור מוחלט של המידע. הקלט והמצב הפנימי קובעים כמה מידע יש לשכוח, אולם הם אינם קובעים מה יהיה תוכן תא הזיכרון לאחר העדכון. מן הצד השני, הוספת מידע חדש לזיכרון אינה מתבצעת כדריסה של התוכן על ידי הקלט, אלא כחיבור שלו. שער הקלט מייצר וקטור שמתווסף למידע הקיים בתא

הזיכרון בפעולת חיבור (element-wise addition) ובכך מאפשר את העברת האינפורמציה ברשת ומונע את דעיכת הגרדיאנט, מכיוון שפעולה זו פחות רגישה לשינויים בקלט כפי שמוצא פונקציית האקטיבציה רגישה אליהם. כתוצאה מכך, הנגזרות שאינן מושפעות באופן ישיר מפונקציות האקטיבציה ולא מתאפסות, מאפשרות את עדכון המשקולות (בשונה מ-RNN שהנגזרות מתאפסות די מהר). בנוסף, מכיוון ששער השכחה מייצר וקטור הקובע אילו חלקים ברכיב הזיכרון יש לשכוח ואילו לשמר, כאשר ערך הוקטור קרוב ל-1, המידע יכול לזרום ברשת מבלי להינזק באופן משמעותי או במילים אחרות, מתאפשרת שמירת מידע ארוך טווח (מידול מתמטי של נושא זה ניתן למצוא [בקישוב](#)). בזכות השימוש ברכיב זיכרון, המצב הפנימי מצטמצם להיות זיכרון לטווח קצר בלבד והשפעת צוואר הבקבוק שנוצרת ב-RNN פוחתת גם היא.

למרות היתרונות המובהקים של הרשת על פני RNN במשימות בהן ישנו צורך לנתח תלויות ארוכות טווח, קיימות לה מגבלות בניתוח שפה טבעית הנובעות ממורכבות הארכיטקטורה ואופן ההפעלה האיטרטיבי שלה. הבעיה הראשונה משותפת לכלל הרשתות האיטרטיביות, והיא חוסר היכולת להפעיל את הרשת במקביל, דבר הגורר הפעלה איטית של הרשת ואי ניצול מספק של משאבי החישוב, בנוסף, לא מאפשרת למדל תלויות בין מקטעים שונים של הקלט.

מכיוון שכל תא LSTM מכיל מספר גדול יותר של פרמטרים בהשוואה ל-RNN, אימון והפעלת הרשת דורשים משאבי חישוב (זיכרון וכוח עיבוד) רבים יותר. בנוסף, על מנת שהרשת תמדל את התלויות ארוכות הטווח באופן אפקטיבי ישנו צורך בכמות משמעותית של דאטה בעלת מספר טוקנים גבוה (=אורך הסדרה). בנוסף, משך האימון הוא גם כן חסרון משמעותי של הרשת. חיפוש תלויות ארוכות טווח בסדרה הינה משימה מורכבת יותר ביחס לחיפוש דפוסים מקומיים (local patterns) בדאטה, ולכן נדרש זמן אימון ארוך יותר.

#### נקודות מפתח לסיכום הפוסט:

- LSTM הינה יחידת עיבוד איטרטיבית המכילה רכיב זיכרון שאוגר 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 הרשת פותרת את דעיכת הגרדיאנט על ידי עדכון רכיבי הזיכרון באמצעות הקלט באופן עקיף, ולא באופן ישיר כמו ב-RNN.
- עקב ריבוי הפרמטרים של הרשת, ומידול המידע ארוך הטווח, משך האימון וכמות המשאבים הנדרשים לאמן את הרשת מהווים עקב אכילס שלה.