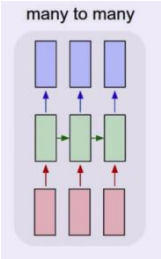
תרגיל 3

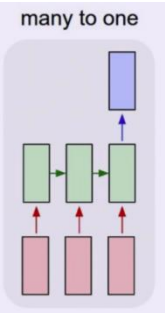
חלק תיארטי

שאלה 1

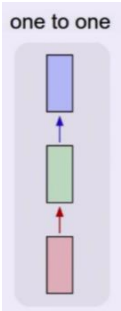


א. כאן קלט מהווה סדרת אותות ופלט זה סדרת מילים. בחירת מילה מתאימה בכל רגע זמן אמורה להתבצע ע"י התחשבות במילים הקודמות שנבחרו,כלומר מה שהוקלט לפני משפיע על הפלט הנוכחי. לכן, הרשת צריכה לעבד סדרת סיגנלים ולפלוט סדרת מילים עם התחשבות בעבר ← עדיף להשתמש ב- many to many. בנוסף, יש לציין שאורך הטקסט תלוי ישירות באורך ההקלטה ← נבחר **many to many RNN** מהסוג הבא:

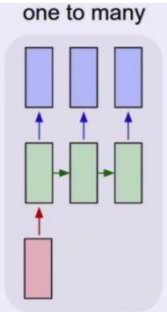
ב. אורך שאלה יכול להשתנות. ניתן לענות על השאלה תוך שימוש במשפט אחד או כמה משפטים. למרות זאת, כדי לענות על שאלה יש להבין קשר בין כל מילה בה. לכן, עיבוד ספרתי לא מתאים כאן, אלא יש לענות ע"י חישוב attention ← יש להשתמש ב **many to many transformer**.



ג. קלט: טקסט ← ממד הקלט משתנה (טקסט יכול להכיל מספר משפטים באורכים שונים). פלט זו מילה שמתארת את רגש הטקסט ← הרשת עוברת מרב ממד לחד ממד ← יש להשתמש ב-**many to one RNN** .



ד. קלט תמיד מהווה תמונה אחת ופלט הרשת זה תיוג מחלקה בודדת ← הבעיה שקולה ל class prediction עם דחיסת תמונה לטנזור ← יש להשתמש ב-**one to one CNN**.



ה. ממד הקלט הינו קבוע, מכיוון שהרשת מקבלת מילה אחת בלבד. למרות זאת, יכול להיות שלמילה אין תרגום מדויק בשפה אחרת, אלא היא שקולה לאוסף מילים. עקב זאת, ממד הפלט אינו קבוע ← יש לבחור **one to many RNN**.
אם יוצאים מנקודת הנחה שלכל מילה יש מילה שקולה אחת בכל שפה אחרת, אז ניתן להשתמש ב-**one to one FNN**.

שאלה 2

א. אורך טקסט משתנה, גודל תמונה משתנה גם אינו קבוע

← יש לבנות רשת מסוג "many to many".

שלב א': עיבוד קלט.

נשתמש ב-ELMo. ראינו בתרגול ש-ELMo מתחשב בהקשר מילה ובסדר מילים . יהיו בו שלוש שכבות: raw, shallow, deep. בסוף תהליך העיבוד, הטקסט יהווה צירוף לינארי שכבות הנ"ל← גודל הקלט אחרי העיבוד יהיה שווה למספר סופי.

שלב ב': מעבר למרחב לטנטי.

ידוע כי קלט מתאר תמונות מסוג מסוים ← ניתן להוציא מהמידע הטקסטואלי נוירונים הכי חשובים ורלווניים. עקב זאת הרשת תצטרך לדעת להוציא כמות סופית של נוירונים שמתארים תמונות ← על מנת לעשות זאת נשתמש ב-mapping.

שלב ג': הכנת מידע לפני שחזור תמונה

סדרת שכבות קונבוציה ו-upsampling אמורה לקבל מידע בצורת טנזור {C, H, W} ← נשתמש בשכבות fully connected.

שלב ד': בניית פלט.

נעביר את המידע דרך סדרת שכבות קונבולוציה ו-upsampling כדי לקבל בסוף תמונה. המעבר הזה שקול ל-"VGG16" בסדר הפוך.

ב. נניח שארבעת הוקטורים יודעים. נקרא להם $X = \{x_1, x_2, x_3, x_4\}$. נניח שממדם d.
שלב א': עיבוד קלט.
נשתמש ב-ELMo. יהיו בו שלוש שכבות: raw, shallow, deep. בסוף תהליך העיבוד, הטקסט יהווה צירוף לינארי שכבות הנ"ל← הקלט ייראה כטנזור {num of words, d}.

שלב ב': חישוב משקל הattention של כל מילה בטקסט עבור כל וקטור לטנטי.
נגדיר שלוש מטריצות W_Q, W_K, W_V, כך ש-W_Q מקבלת וקטורים לטנטיים ו-W_K, W_V מקבלות קידודי טקסט.

$$attention_{\mathrm{i,t}} = softmax(\frac{Q(x_i)K(w_t)}{\sqrt{d}})$$

שלב ג': חיבוש תוכן לכל רבע תמונה:

$$Context_{x_i} = \sum_{t=1}^{Num\ of\ words} attention_{i,t} \cdot W_V(w_t)$$

שאלה 3

א. נשתמש בנוסחה הבאה:

$$\dim_{out} = floor\{\frac{\dim_{in} + 2 \cdot padding - kernel}{stride}\} + 1$$

נקבל:

$$\begin{bmatrix}128\\128\end{bmatrix}\overrightarrow{p=0,k=3,s=2}$$
$$\begin{bmatrix}63\\63\end{bmatrix}\overrightarrow{p=2,k=5,s=1}$$
$$\begin{bmatrix}63\\63\end{bmatrix}\overrightarrow{p=1,k=3,s=1}$$
$$\begin{bmatrix}63\\63\end{bmatrix}\overrightarrow{p=0,k=5,s=2}\begin{bmatrix}30\\30\end{bmatrix}$$

לכן פלט נראה כטנזור 30 על 30 על 1.

ב. receptive field – כמות נוירונים שרואה כל נוירון בשכבת הקונבולוציה האחרונה.

נשתמש בנוסחה:

$$RF_l = RF_{l-1} + (k_l - 1)\prod_{i=1}^{l-1} s_i$$

נקבל:

$$RF_0 = 1$$
$$RF_1 = 1 + (3-1)\cdot 1 = 3$$
$$RF_2 = 3 + (5-1)\cdot(1\cdot 2) = 11$$
$$RF_3 = 11 + (3-1)\cdot(1\cdot 1\cdot 2) = 15$$
$$RF_3 = 15 + (5-1)\cdot(1\cdot 1\cdot 2\cdot 1) = 23$$

כלומר: receptive field = 23 x 23 .

שאלה 4

יהי אוסף וקטורים $X = \{\overrightarrow{x_1}, \overrightarrow{x_2}, \dots, \overrightarrow{x_n}\}$ שהוא מטריצת פיצר'רים.

יהיו שלוש מטריצות טרנספורמר: W_Q, W_K, W_V.

נניח שהחלפנו סדר וקטורים ע"י הכפלת X במטריצה אחרת בשם P, כאשר $P \in \{0,1\}^{n \times n}$

טענה: $V^{out}(XP) = V^{out}(X)$

הוכחה:

$$Q = XW_Q, K = XW_K, V = XW_V$$

בעבור XP:

$$Q_{new} = (XP)W_Q, K_{new} = (XP)W_K, V_{new} = (XP)W_V$$

נציב בנוסלה לחישוב attention:

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d}}\right)V$$

$$Attention(XP) = Attention(Q_{new}, K_{new}, V_{new}) = softmax\left(\frac{(XP)W_Q[(XP)W_K]^T}{\sqrt{d}}\right)XW_VP =$$
$$= softmax\left(\frac{XW_QP[XW_KP]^T}{\sqrt{d}}\right)XW_VP = softmax\left(\frac{XW_QPP^TW_K^TX^T}{\sqrt{d}}\right)XW_VP$$

נשים לב שהמטריצה $P \in \{0,1\}^{n \times n}$ רק משנה סדר באוסף X ← $PP^T = I$

לכן נקבל:

$$Attention(XP) = softmax\left(\frac{XW_QPP^TW_K^TX^T}{\sqrt{d}}\right)XW_VP = softmax\left(\frac{XW_QIW_K^TX^T}{\sqrt{d}}\right)XW_VP$$
$$= softmax\left(\frac{QK^T}{\sqrt{d}}\right)VP = Attention(Q,K,V)P = Attention(X)P$$

$$V^{out}(X) = \sum_i Attention(x_i)v_i = Attention(X)V$$

$$V^{out}(XP) = Attention(XP)V = Attention(X)PV =$$

$$\sum_{same\ vectors, but\ \ another\ order} Attention(x_{i'})v_{i'} = \sum_i Attention(x_i)v_i = V^{out}(X)$$

כעת נסתכל על הבעיה כאשר יש חשיבות לסדר. נעשה זאת ע"י הוספת וקטורי מיקום:

$$E = e_1, e_2, \dots, e_n$$

טענה: $V^{out}(XP + E) \neq V^{out}(XP + EP)$

הוכחה:

$$Q_{new} = (XP + E)W_Q, K_{new} = (XP + E)W_K, V_{new} = (XP + E)W_V$$

$$
\begin{aligned}
Attention(XP+E) &= softmax\left(\frac{\left(XW_QP+EW_k\right)(XW_VP+EW_V)^T}{\sqrt{d}}\right)(XW_VP+EW_V) \\
&= softmax\left(\frac{\left(XW_QP+EW_k\right)(P^TW_V^TP^T+W_V^TE^T)}{\sqrt{d}}\right)(XW_VP+EW_V)
\end{aligned}
$$

$$
\begin{aligned}
Attention\big((X+E)P\big) &= Attention(XP+EP) \\
&= softmax\left(\frac{\big((XP+EP)W_K\big)\big((XP+EP)W_K\big)^T}{\sqrt{d}}\right)(XP+EP)W_V \\
&= softmax\left(\frac{(XW_KP+EW_KP)(XW_KP+EW_KP)^T}{\sqrt{d}}\right)(XW_VP+EW_VP) \\
&= softmax\left(\frac{(XW_KP+EW_KP)(P^TW_K^TX^T+P^TW_K^TE^T)}{\sqrt{d}}\right)(XW_VP+EW_VP)
\end{aligned}
$$

$$
Attention(XP+E) \neq Attention((X+E)P)
$$

$$
\rightarrow \sum_i Attention(x_ip_i+e_i)v_i \neq \sum_i Attention\big((x_i+e_i)p_i\big)v_i
$$

$$
\rightarrow V^{out}(XP+E) \neq V^{out}(XP+EP)
$$