NAIM Mouna
mouna.naim@ens-paris-saclay.fr

Lab session 1
ALTEGRAD 2024

13/10/2024

1 Question 1

The attention mechanism can be improved as follows:

- Instead of encoding a vector into a single vector, the model uses a 2D matrix where each row captures different semantic components of the sentence. Formally if $\mathbf{H}$ represents the LSTM hidden states for the sentence, the attention mechanism generates a matrix of weights $\mathbf{A}$, applied to $\mathbf{H}$ to produce the embedding matrix $\mathbf{M}$:

$$\mathbf{M} = \mathbf{AH} \tag{1}$$

  To allow the model to focus on multiple aspects of the sentence, the attention mechanism computes several attention vectors in parallel (multiple attention hops). The matrix attention is computed as:

$$\mathbf{A} = \text{softmax}\left(\mathbf{W_s}^2 \tanh\left(\mathbf{W_s}^1 \mathbf{H}^T\right)\right) \tag{2}$$

  Where $\mathbf{W_{s1}}$ and $\mathbf{W_{s1}}$ are learnable weight matrices.

- To ensure that each attention hop focuses on different parts of the sentence, a penalization term is used to minimize redundancy between the rows of $\mathbf{A}$. The penalization term is based on the Frobenius norm:

$$\mathbf{P} = \left\|\mathbf{A}\mathbf{A}^T - I\right\|_F^2 \tag{3}$$

  This penalization allows the matrix $\mathbf{A}$ to be close to orthogonal, meaning each row focuses on different parts of the sentence. This helps avoid redundancy and makes the attention mechanism more efficient in covering various sentence aspects.

2 Question 2

- Recurrent models process input sequences sequentially. For each step $\mathbf{t}$, the hidden state h_t depends on the previous hidden state h_{t-1}, which forces the model to process each token one by one. This prevents parallel computation, since future tokens must wait for the computation of earlier ones. However, self-attention avoids this dependency. It processes all input tokens at once since the attention mechanism computes relationships between all tokens simultaneously, which facilitates parallel processing.

- In contrast to recurrent operations, self-attention relates all tokens to each other in a single step. Each token can attend to every other token in the sequence with just one attention operation, regardless of the distance between them. This makes self-attention effective at capturing dependencies between distant tokens, which is important for tasks that require understanding global context.

3 Question 3

by	Dane	Youssef	I	was	kind	of	looking	forward	to	this	one	.	
I	enjoy	Eddie	Murphy	and	I	love	it	when	a	star	OOV	a	vehicle
for	themselves	or	when	someone	who	writes	decides	to	mark	their	own	directorial	debut
.													
But	when	the	star	's	head	gets	too	big	for	the	rest	of	his
body	,	there	's	always	a	danger	of	a	OOV	Hollywood	vanity	production	.
Will	the	filmmaker	keep	it	OOV	?							
or	will	he	just	waste	amounts	of	money	(	the	studio	's	,	ours
)	and	time	(	the	studio	's	,	ours	&	his	own	)	patting
himself	on												
Sadly	,	it	's	the	latter	here	.						
Another	thing	I	really	like	is	when	someone	breathes	new	and	fresh	life	into
an	exhausted	and	OOV	genre	.								

Figure 1: Coefficients per word

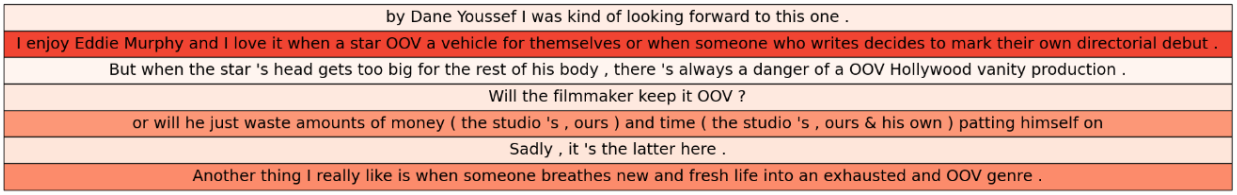

Figure 2: Coefficients per sentence

These two figures show how the attention architecture uses coefficients to focus on the most significant words and sentences in a document. The more red the word or sentence is, the higher coefficient it has.
The use of attention, in this case, allows the model to focus on key words or sentenses like ”enjoy” and ”waste” to better understand the position of the review, whether it is positive or negative. This helps us see how the model understands the text by focusing on the elements that most influence its decision process.

4 Question 4

The main limitation of HAN is that sentences are encoded in isolation. Each sentence is encoded independently of the other sentences in the document. This means that the model does not take into account the overall document context when encoding each sentence, which can limit the quality of the global representation and make it difficult to understand the logical flow within a paragraph.
As a result, the sentence encoding does not prevent redundancy.