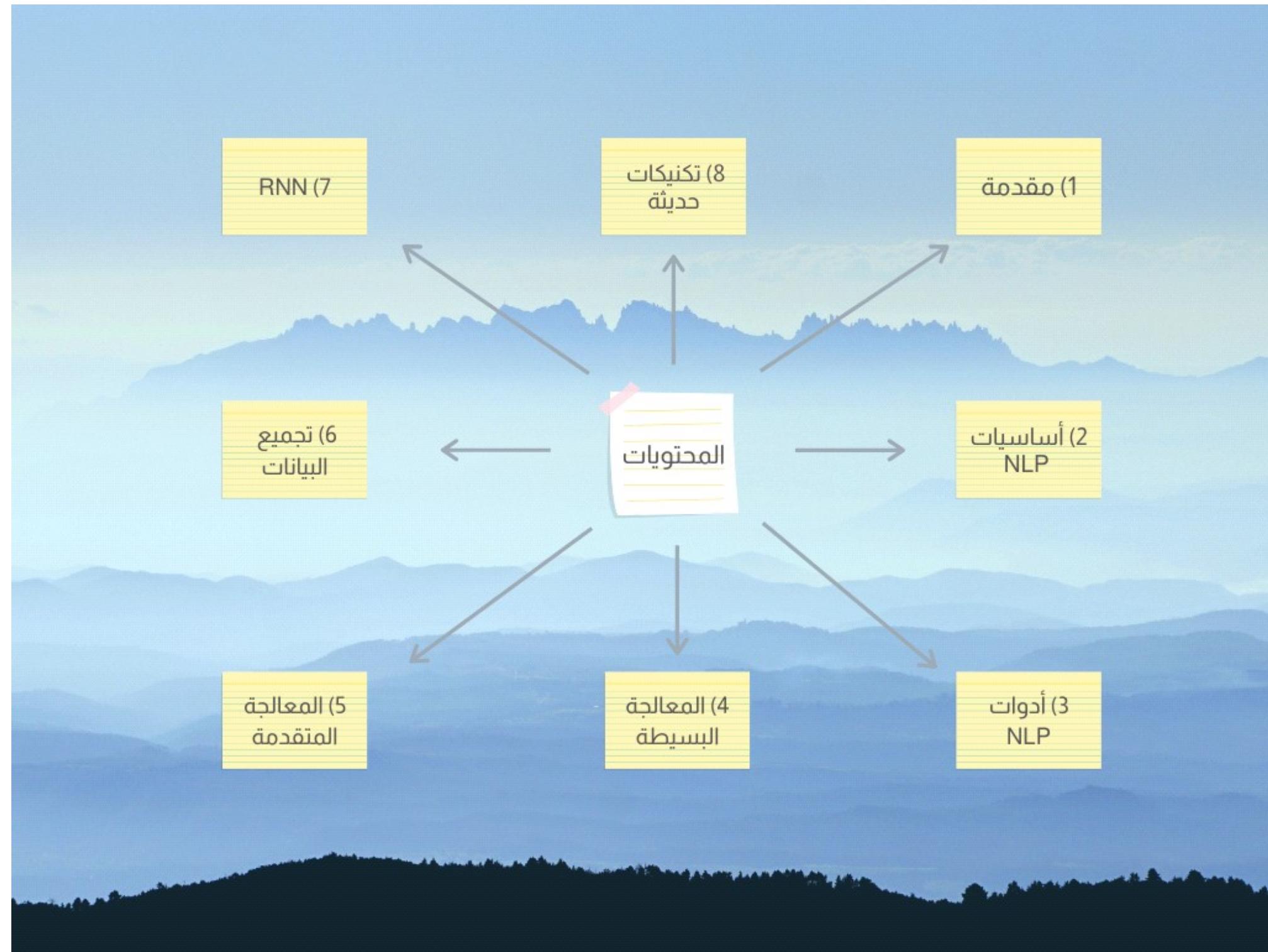


NATURAL LANGUAGE PROCESSING

المعالجة اللغوية الطبيعية



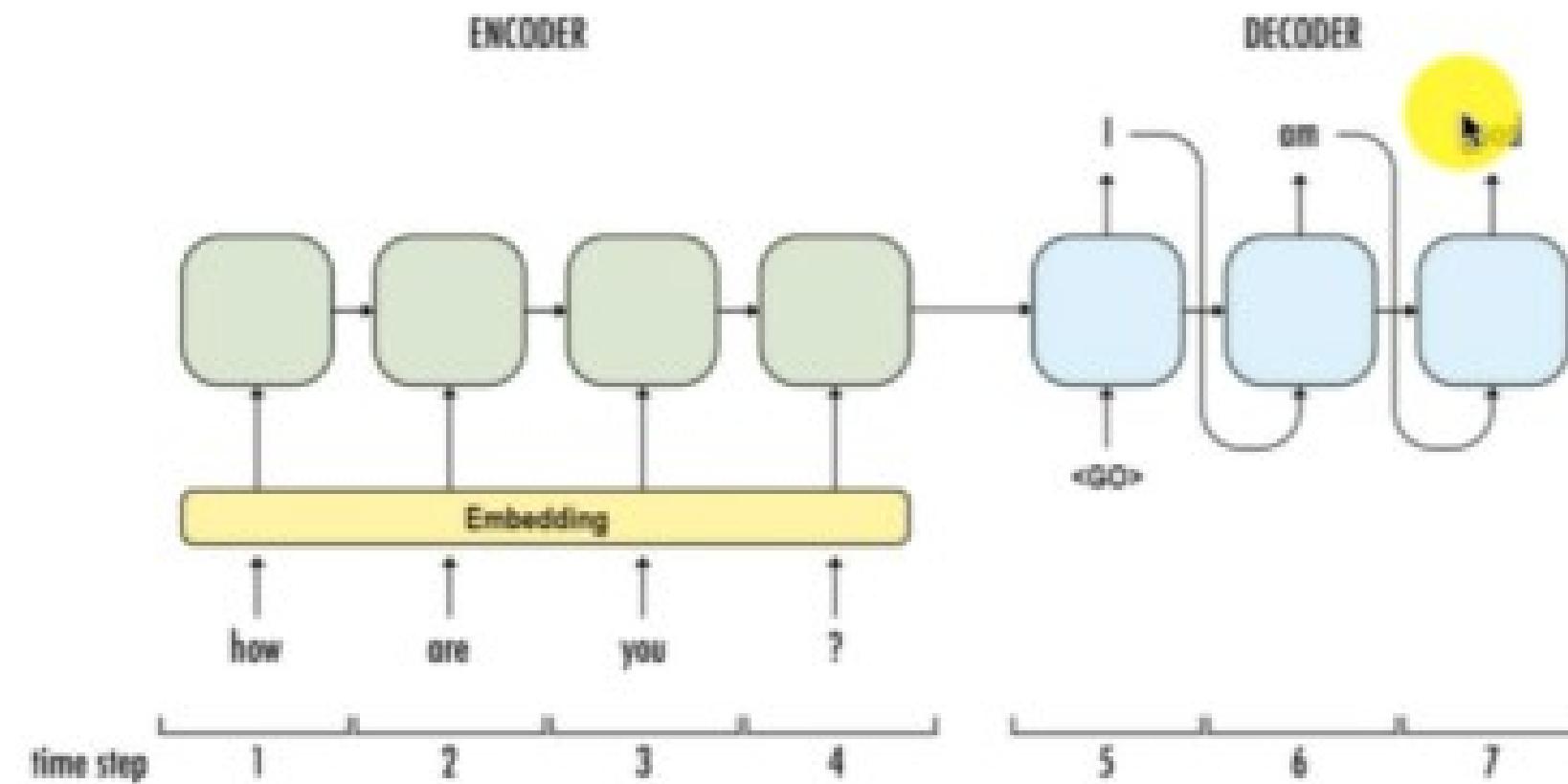
المحتويات

					التطبيقات	العقبات و التحديات	NLP تاريخ ملفات pdf	ما هو NLP الملفات النصية	المحتويات المكتبات	1) مقدمة
T.Visualization	Syntactic Struc.	Matchers	Stopwords	NER	Stem & Lemm	POS	Sent. Segm.	Tokenization	NLP	2) أساسيات NLP
	Dist. Similarity	Text Similarity	TF-IDF	BOW	Word2Vec	T. Vectors	Word embed	Word Meaning		3) أدوات NLP
T. Generation	L. Modeling	NGrams	Lexicons	GloVe	NMF	LDA	T. Clustering	T. Classification		4) المعالجة البسيطة
	Summarization & Snippets		Ans. Questions		Auto Correct	Vader	Naïve Bayes	Sent. Analysis		5) المعالجة المتقدمة
Search Engine	Relative Extraction		Information Retrieval		Information Extraction		Data Scraping	Tweet Collecting		6) تجميع البيانات
					Rec NN\TNN	GRU	LSTM	RNN		7) RNN (
Chat Bot	Gensim	FastText	Bert	Hug. Face	Attention Model	T. Forcing	CNN	Word Cloud		8) تكنيات حديثة

القسم الثامن : تكنولوجيات حديثة

الجزء الرابع : Attention Model

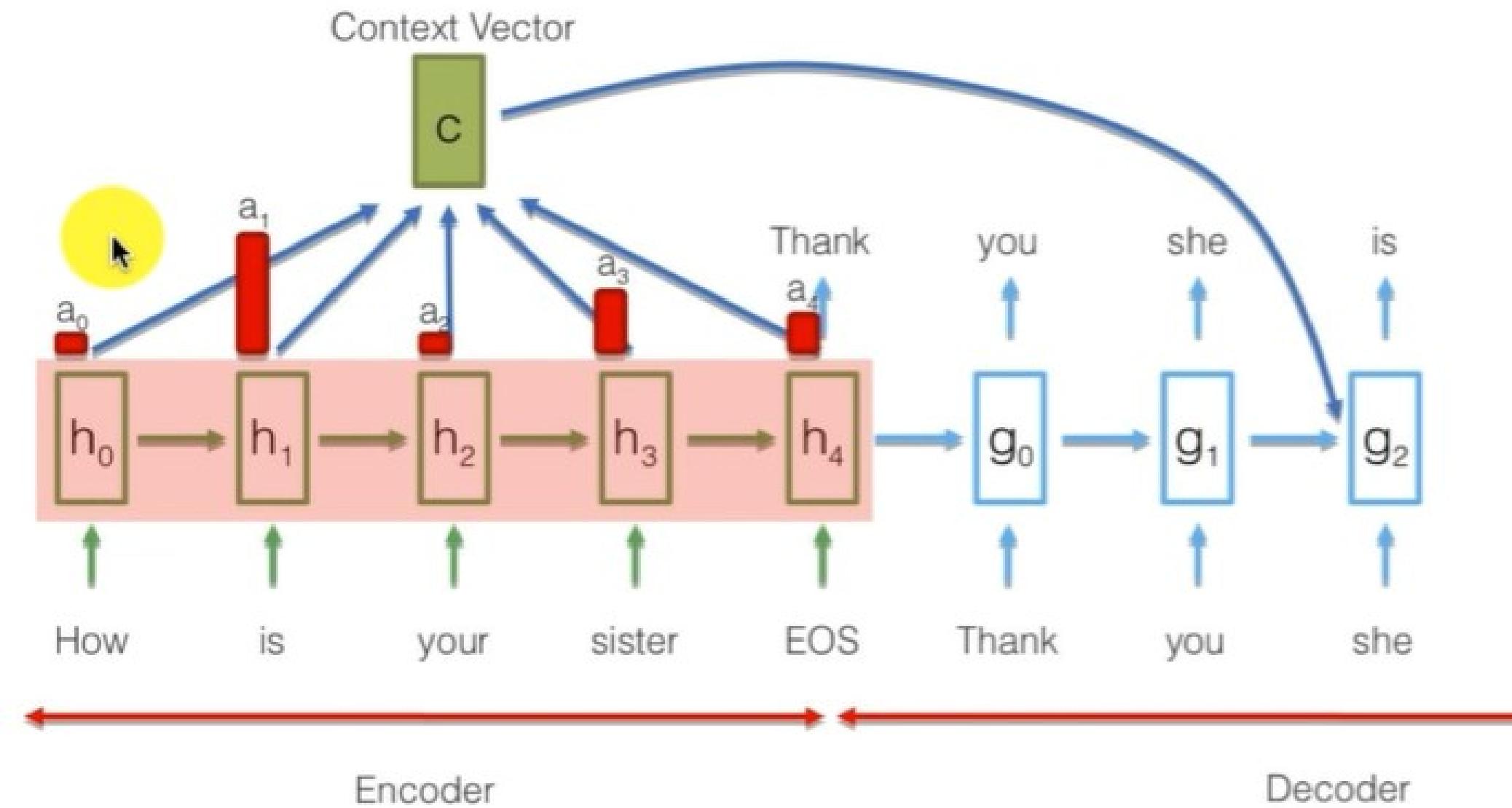
اذا بدأنا بالأمثلة المعتادة لل seq2seq , فهي كانت كالتالي :



فقد كان هناك المشفر و المفسر , حيث المشفر encoder يقوم باستقبال الكلمات الداخلة عليه بشكل متتالي (كلمة وراء الكلمة) و تقوم RNN بمعالجة الكلمات ، ويقوم المفسر decoder باستنتاج الكلمات المطلوب استخراجها كلمة وراء الكلمة)

و من أهم عيوب هذا النظام ان الكلمات الأولى يقل تأثيرها في استخراج الـ output لأن تأثيرها على الشبكة يكون اقل بكثير من الكلمات الاخيرة

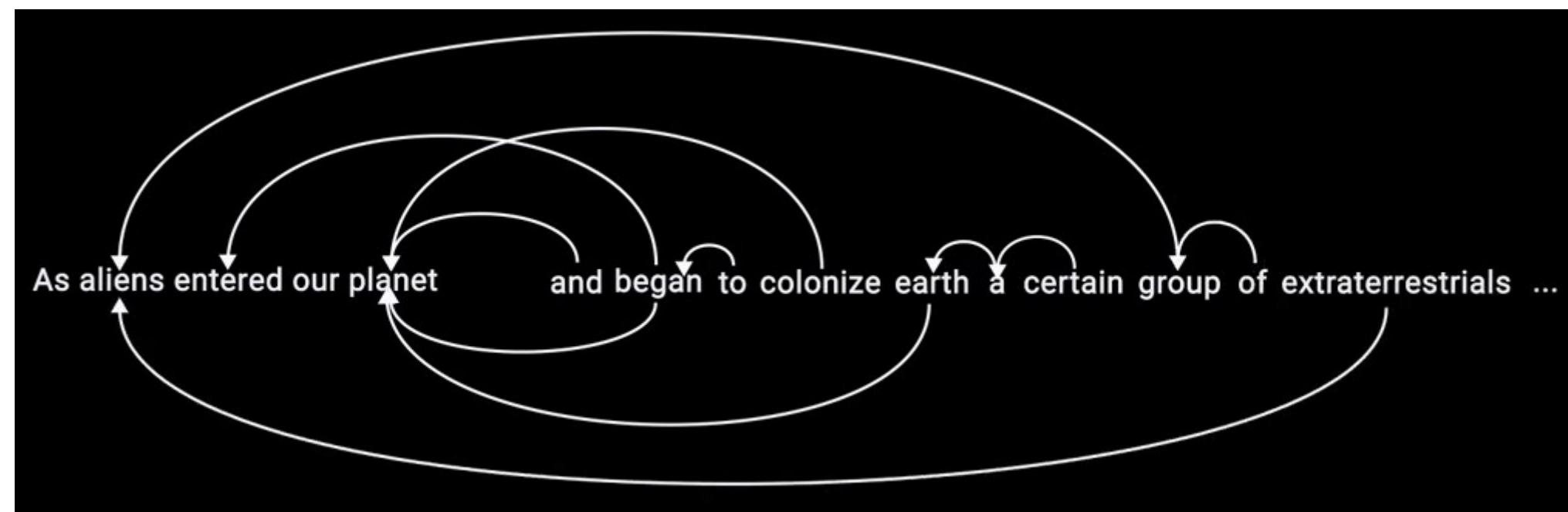
و لحل هذه المشكلة تم عمل تعديل في الشبكة وهو ما يسمى نموذج الانتباه attention model



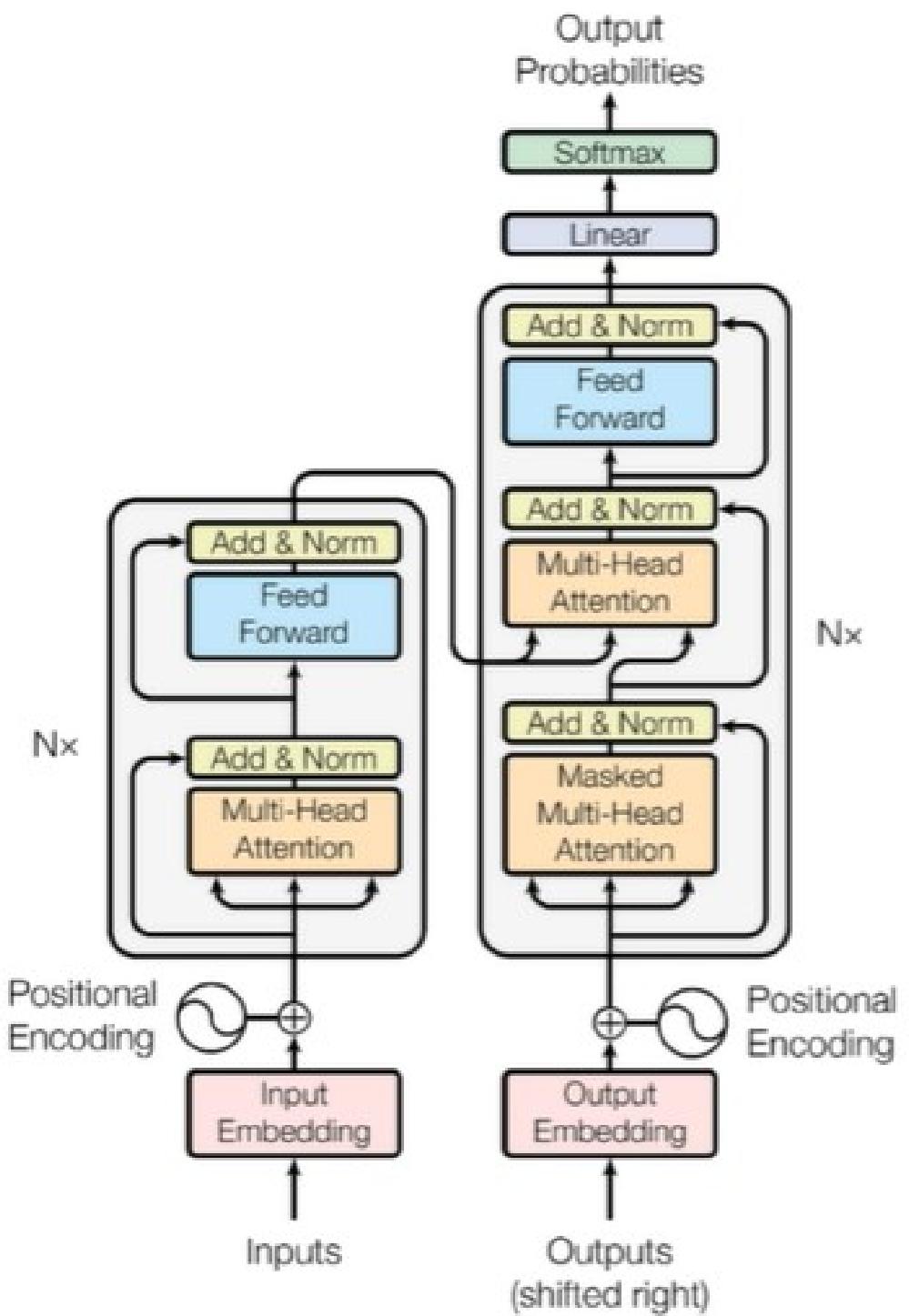
وهو أن يكون هناك وزن weight لكل كلمة من كلمات الـ input بحيث تشير أهميتها وتأثيرها على الكلمة المنشودة في الـ output وبالتالي يتم تحديد أي كلمة بالتحديد من المدخل هي الأكثر تأثيراً على هذه الكلمة في المخرج . .

و بالتالي ففي المدخل thank you she is fine ، كان المخرج how is your sister الكلمة الثانية هي الاكثر تأثيرا علي كلمة is الرابعة في المخرج و بالتالي فسيكون هناك قيم الاوزان h0 , h1 , h2 .. لكل كلمة من الكلمات وهي الخاصة بالكلمة الرابعة في المخرج

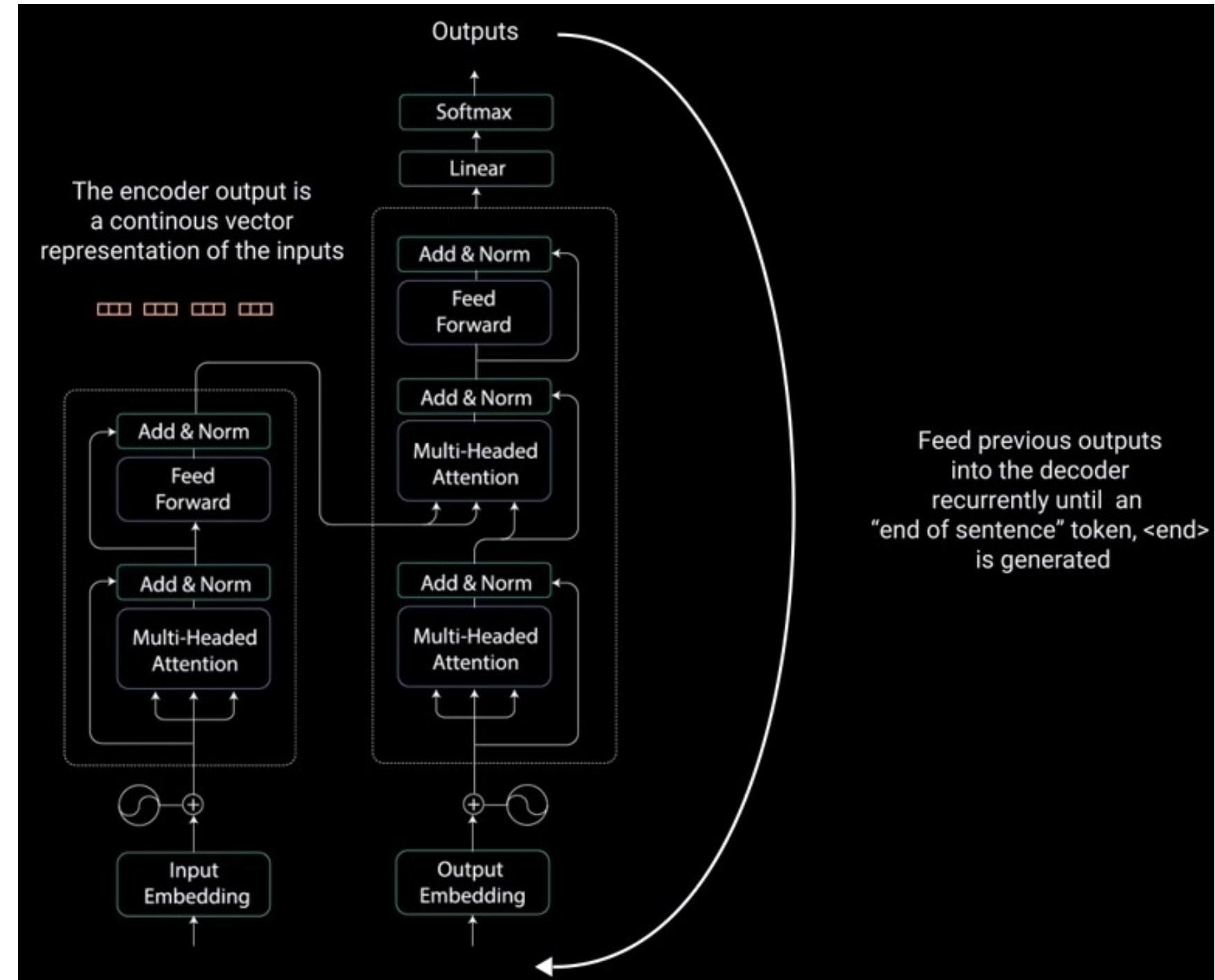
و يتميز نموذج الانتباه أن ذاكرته في التعامل مع الكلمات البعيدة افضل بكثير من RNN , GRU, LSTM ، وهو ما يكون مميزا في تطبيقات مثل Text Generation



و بالتالي علينا التعرف على نموذج الانتباه ، والذي يكون له هذا التصميم :



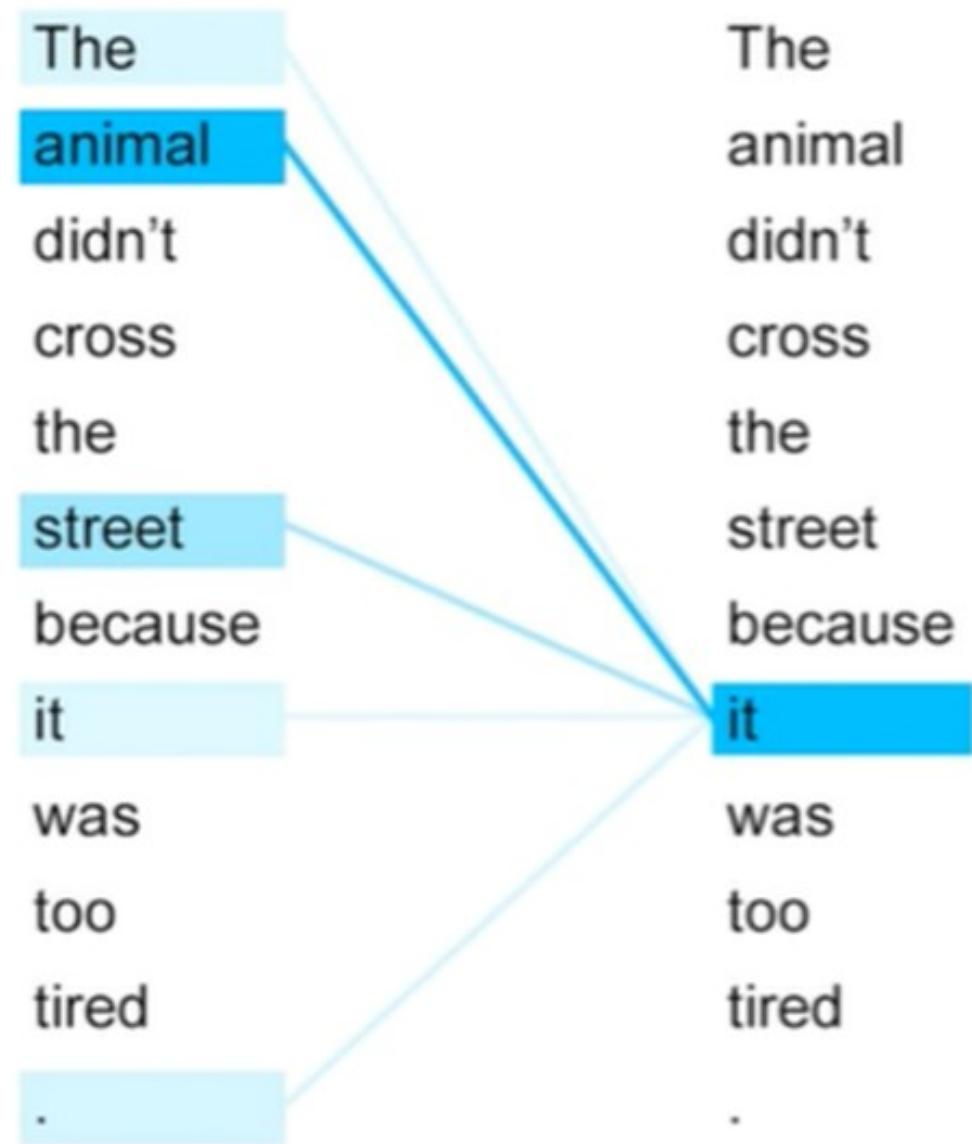
حيث يكون له المدخل من تحت علي اليسار ، ثم يتم انتاج المخرج فوق علي اليمين ، ثم يدخل مرة اخري من تحت علي اليمين و بالتالي سيمكن النموذج من استيعاب العلاقة بين الكلمات و بعضها فالمدخلات تكون عبارة عن Sequence و output كـ Sequence من الكلمات ، ثم يقوم النموذج بإخراج الـ `<end>` الذي يعاد إدخاله مرة أخرى



, فإذا كان لدينا جملة

The animal didn't cross the street because it was too tired

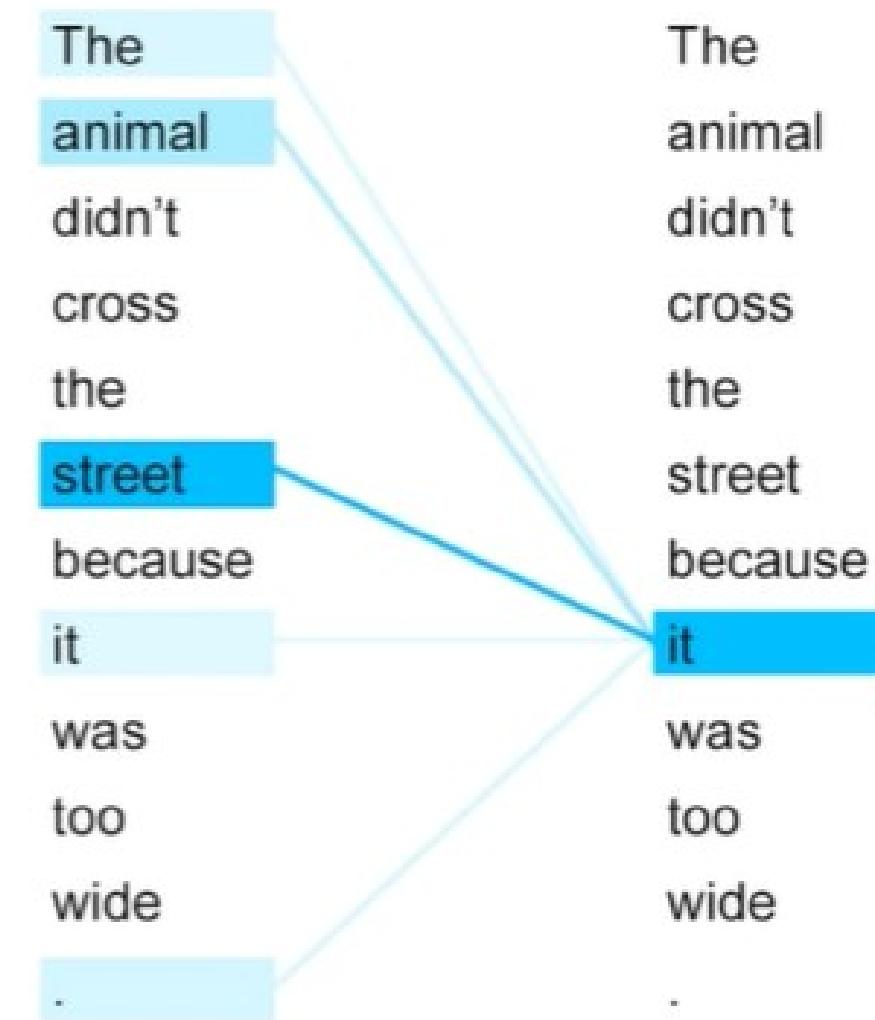
فسيتمكن النموذج من معرفة أن الكلمة **it** هنا تعود إلى الـ **animal**



: بينما لو كانت الجملة

The animal didn't cross the street because it was too wide

هذا سيتمكن النموذج من معرفة ان كلمة it تعود الي الشارع بسبب كلمة wide

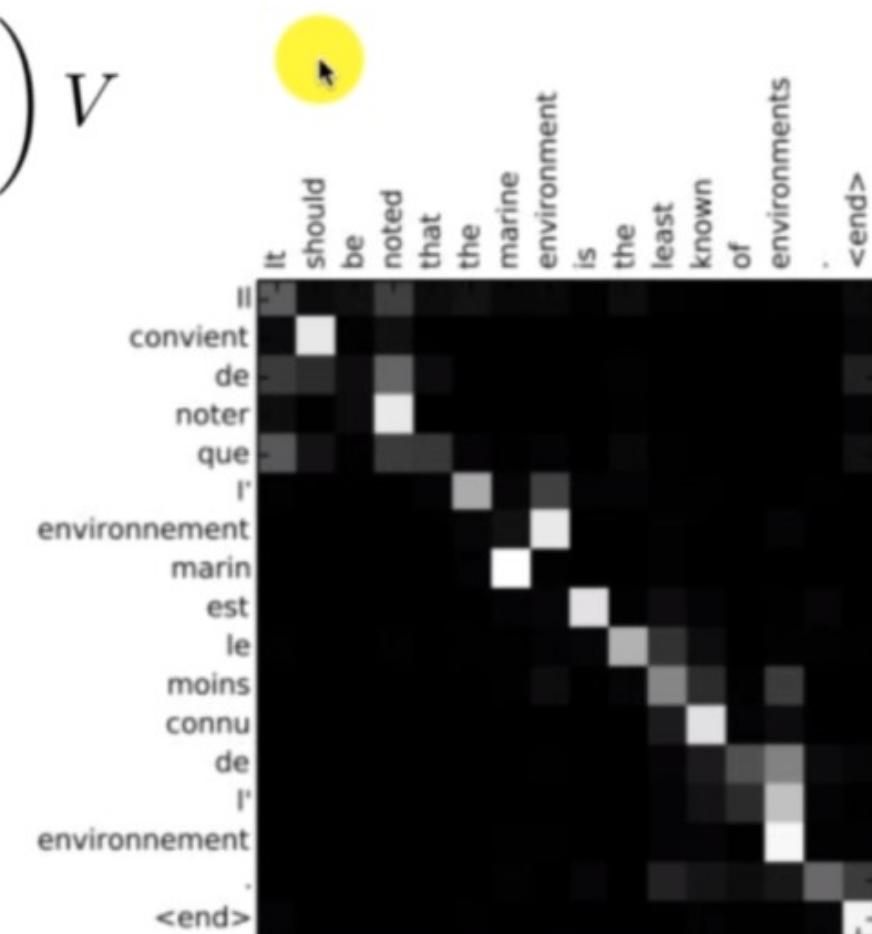


ففي النموذج الأول سيكون لدينا 7 أوزان اكبر لكلمة animal ذات اللون الأزرق و قيم اقل لباقي الكلمات ذات اللون السماوي ، والعكس في الجملة الثانية

و يمكن عمل خريطة بعلاقة الكلمات الاصلية و المترجمة بعضها بعض ، ومن اعتمد على من اثناء الترجمة

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

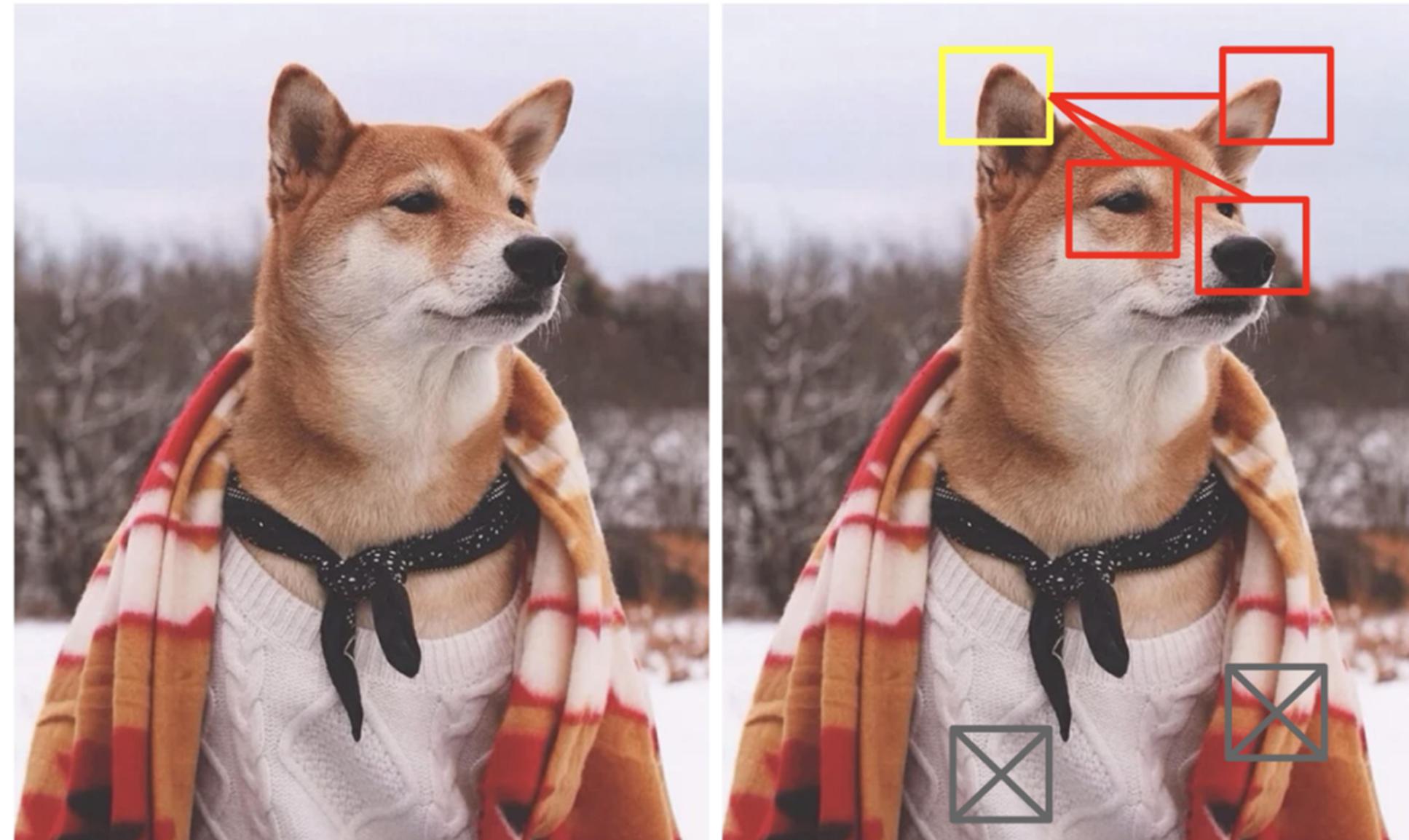
QK^T example:



مع التأكيد أن نموذج الانتباه له جذور في علم النفس ، وفي التعامل البشري مع البيانات التي يستخدمها فالإنسان حينما ينظر إلى صورة ما ، مثل هذه الصورة

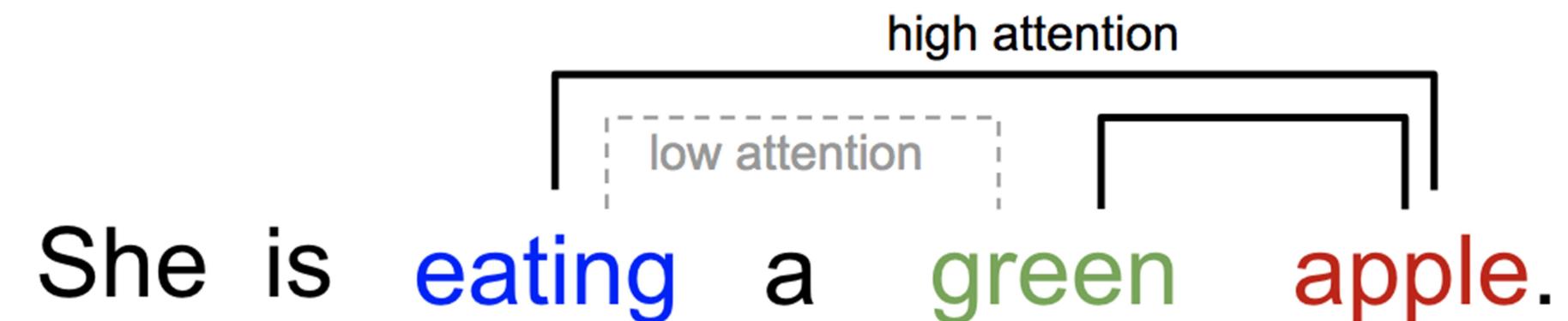


فهذه الصورة لأنها ملئة بالتفاصيل ، فالعين تتجه الي كل جزء علي حدة ، تقوم بعمل إلمام بالتفاصيل فيه ، وحينما تقوم بالتركيز علي جانب ما ، فهي تتجاهل مؤقتا باقي التفاصيل ، حتى تقوم بالإلمام بالصورة الكاملة

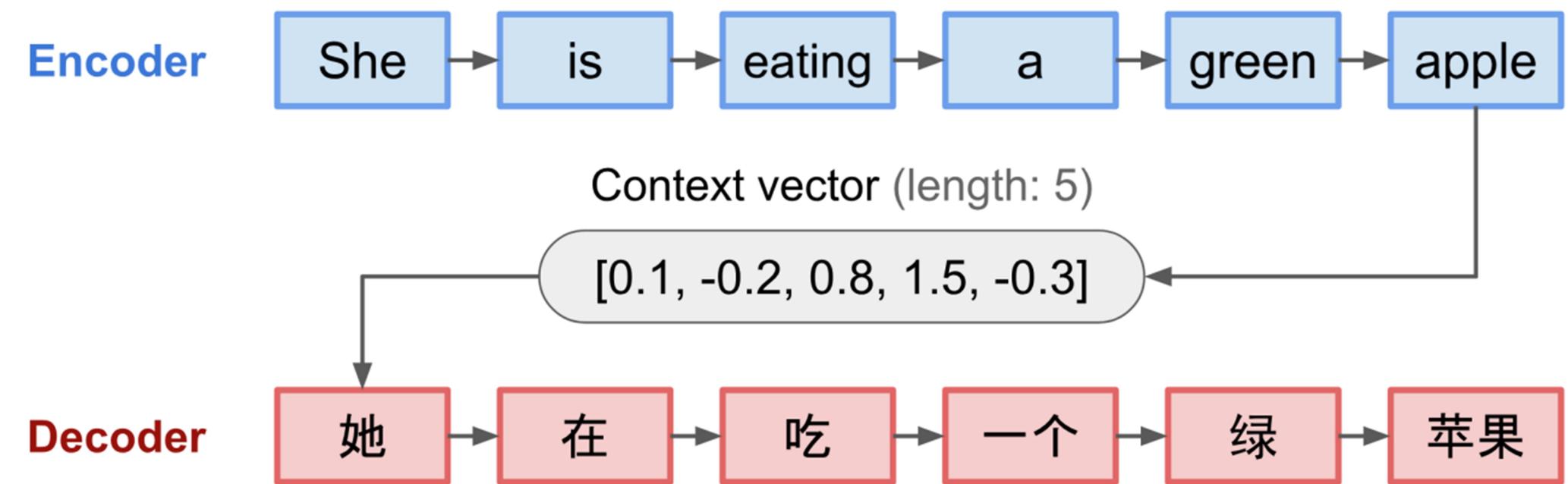


وهذا ما يقصد به بالانتباه ، التركيز على اشياء محددة لنقوم باستنتاج اشياء أخرى

كذلك في النصوص ، فحينما نأتي لكلمة eating توقع بعدها ان هناك كلمة تدل على طعام ، و بالتالي لاستنتاج الكلمة apple فسيتم الانتباه لكلمة eating فهناك علاقة قوية بينهما ، كذلك هناك علاقة قوية بين green ، apple وبين العلاقة بين eating ، green ضعيفة بالرغم انهم متاليان



و قد ابتكر نموذج الانتباه ، لحل معضلة كبيرة كانت في نموذج seq2seq ، وهو التعامل مع الجمل الطويلة ، والتي يصعب على اي موديل التعامل معها ، او تذكر الكلمات البعيدة ، حينما يقوم باستنتاج الجملة الجديدة



اما الان , فإنه لدى استنتاج كلمة ما , ف يتم التركيز على عدد محدد من الكلمات , بأوزان معينة حتى يسهل عليه استنتاج الكلمة بدقة أعلى و وقت أقل , فالسطور القادمة فيها استنتاج كلمات معينة , و الكلمة التي يتم استنتاجها باللون الأحمر , بينما كثافة اللون الأزرق السابق في الكلمات تدل على مدى الانتباه عليها لاستنتاج هذه الكلمة

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

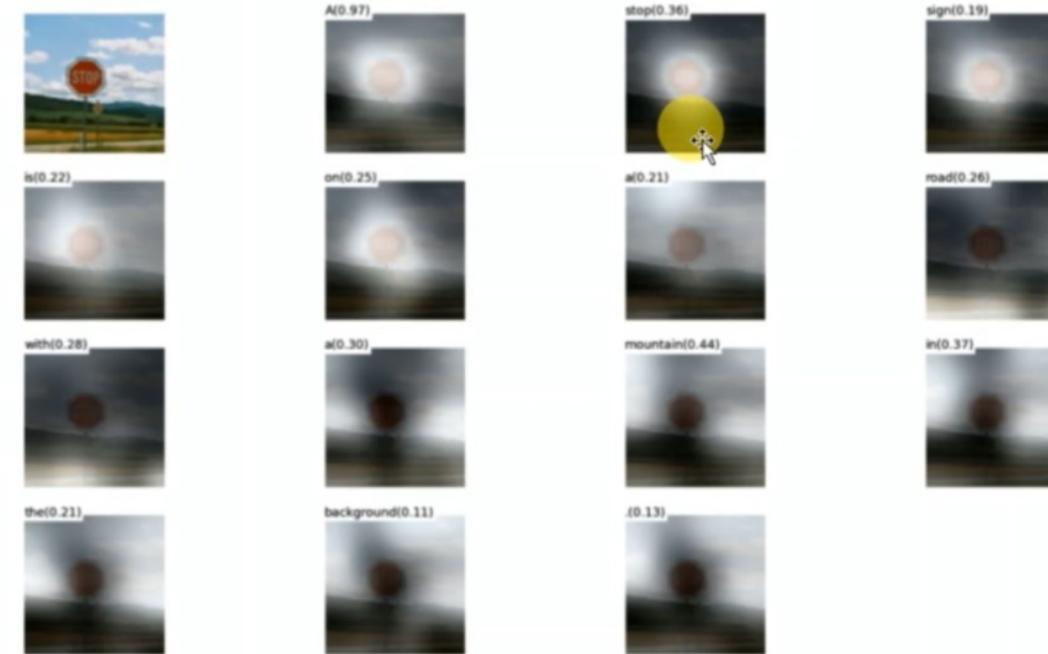
The FBI is chasing a criminal on the run .

نفس الأمر يتم في تطبيق image captioning وهو الخاص باستنتاج وصف للصورة المستخدمة ، فهذا التطبيق يعتمد في استنتاجه لكل كلمة على التركيز على مكان محدد في الصورة ، لمعرفة المحتوي بها



و هنا مثال آخر

Image Captioning with Attention Mechanism



مثال اخر



A man holding a couple plastic containers is walking down an intersection towards me.



A man



holding a couple
plastic containers

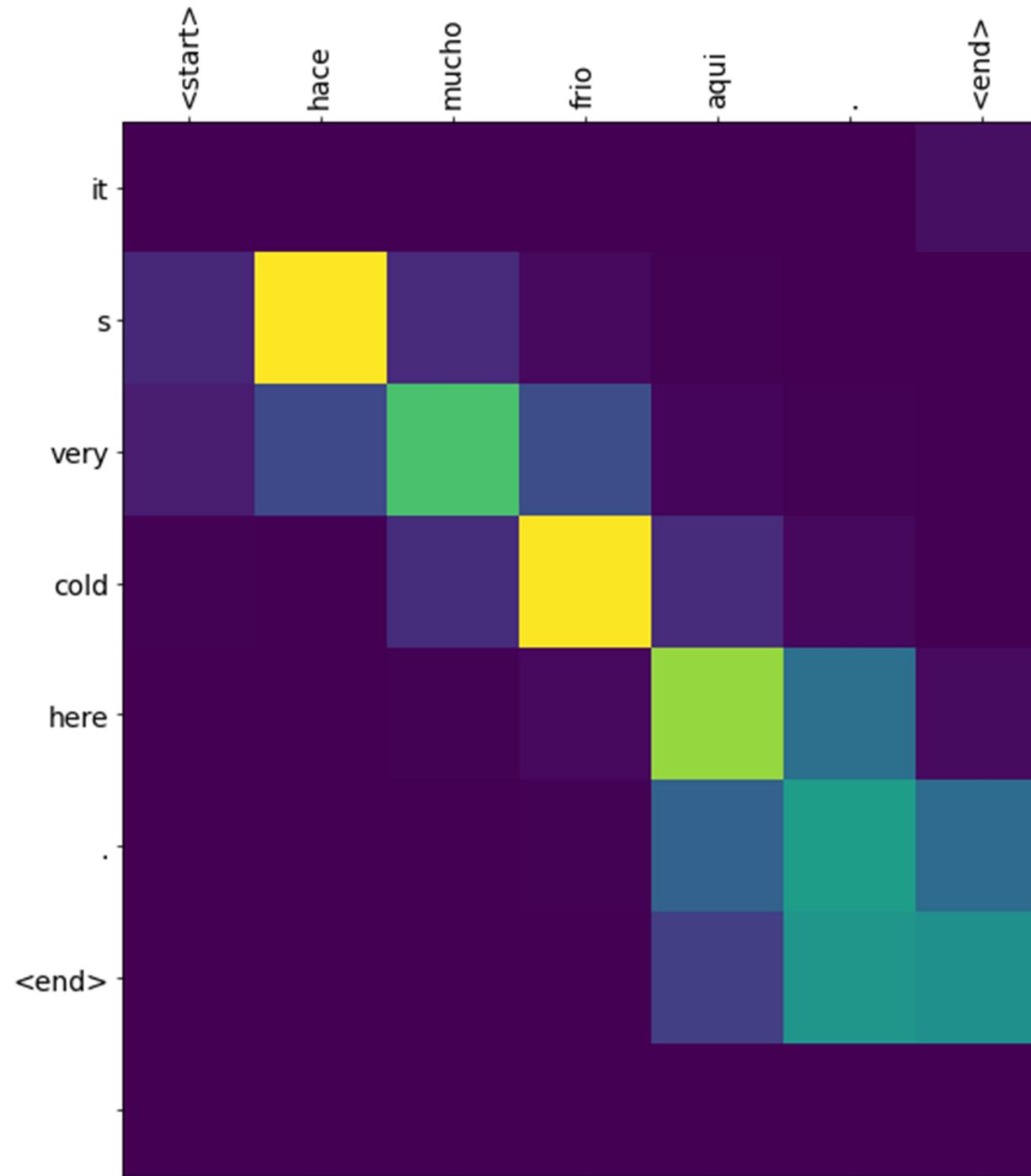


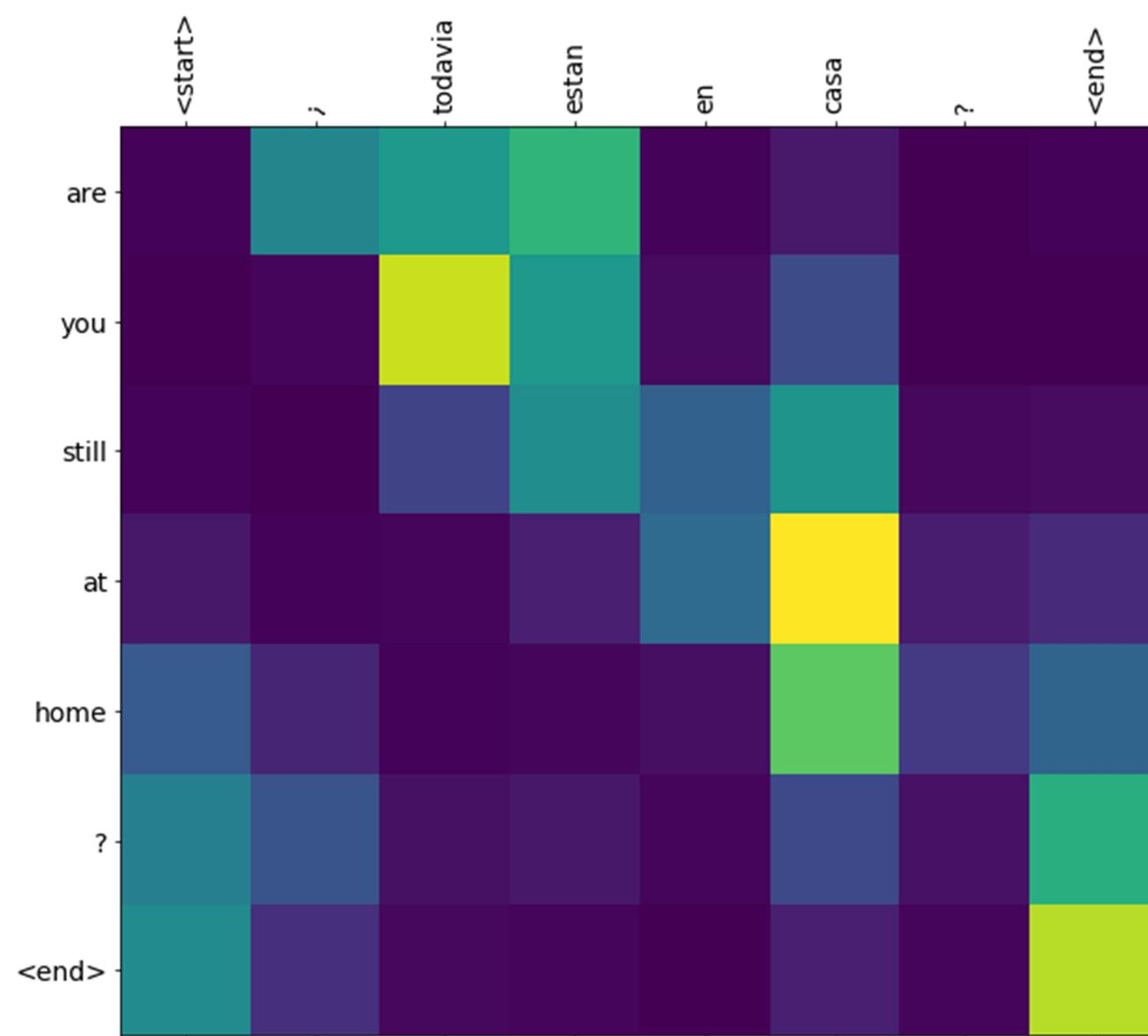
is walking down
an intersection

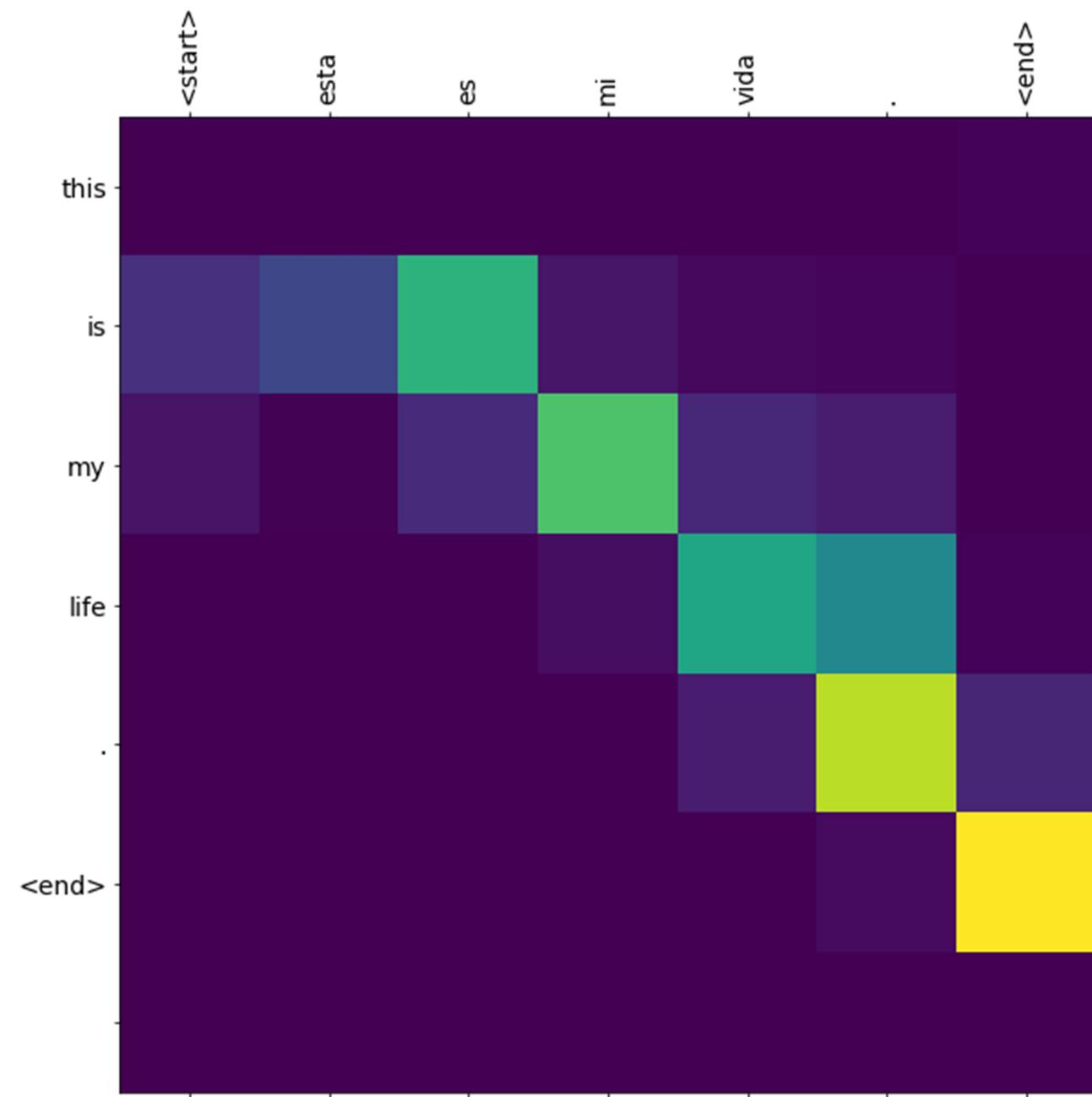


towards
me.

كما يمكن رسم heatmap للعلاقة بين الكلمات المستنيرة و الكلمات الأصلية في نموذج الترجمة , حيث تعتمد كل كلمة مستنيرة على عدد من الكلمات الأصلية , على اساس المعنى الخاص بها







--*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*