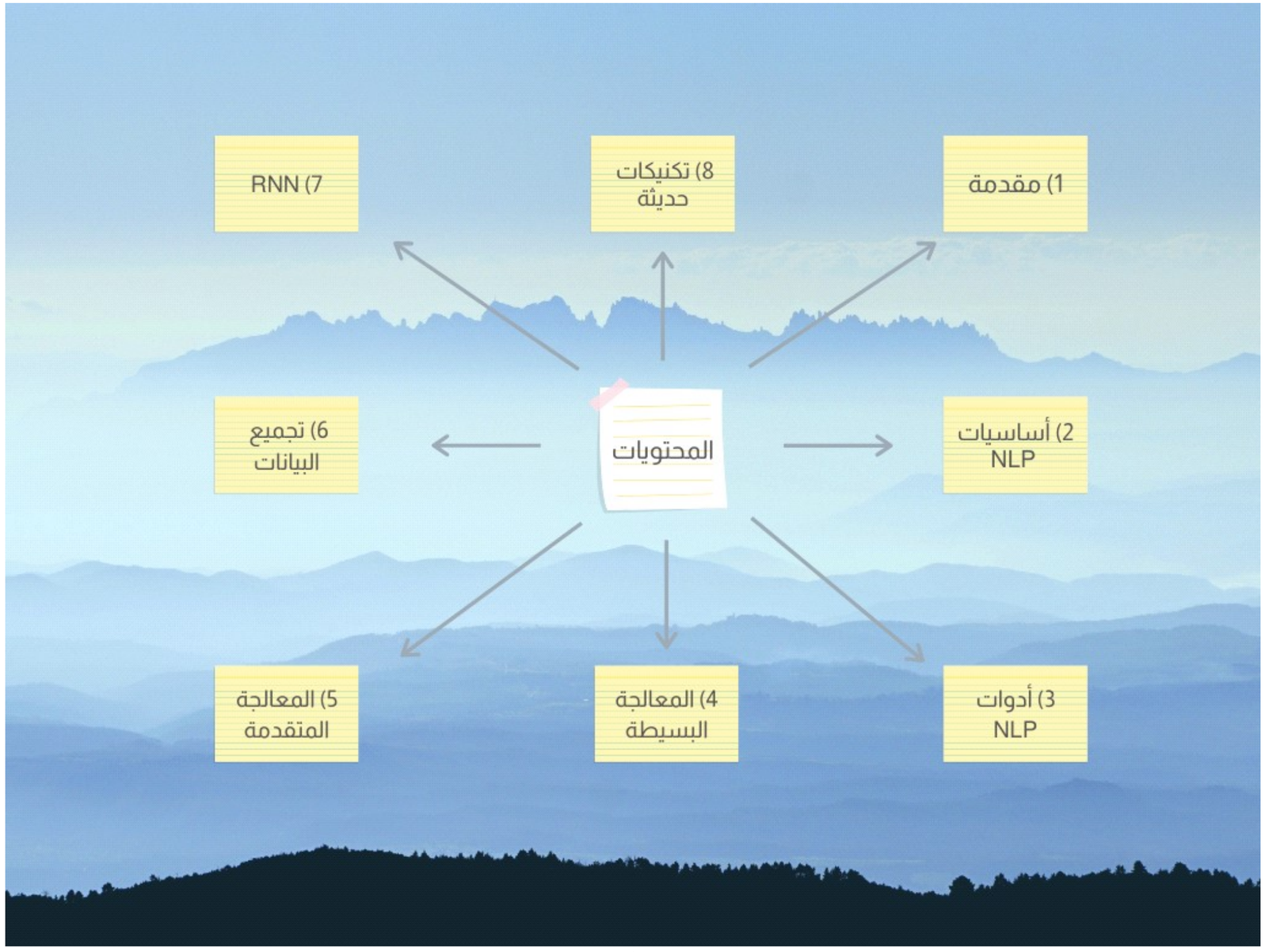
Natural Language Processing المعالجة اللغوية الطبيعية



المحتويات

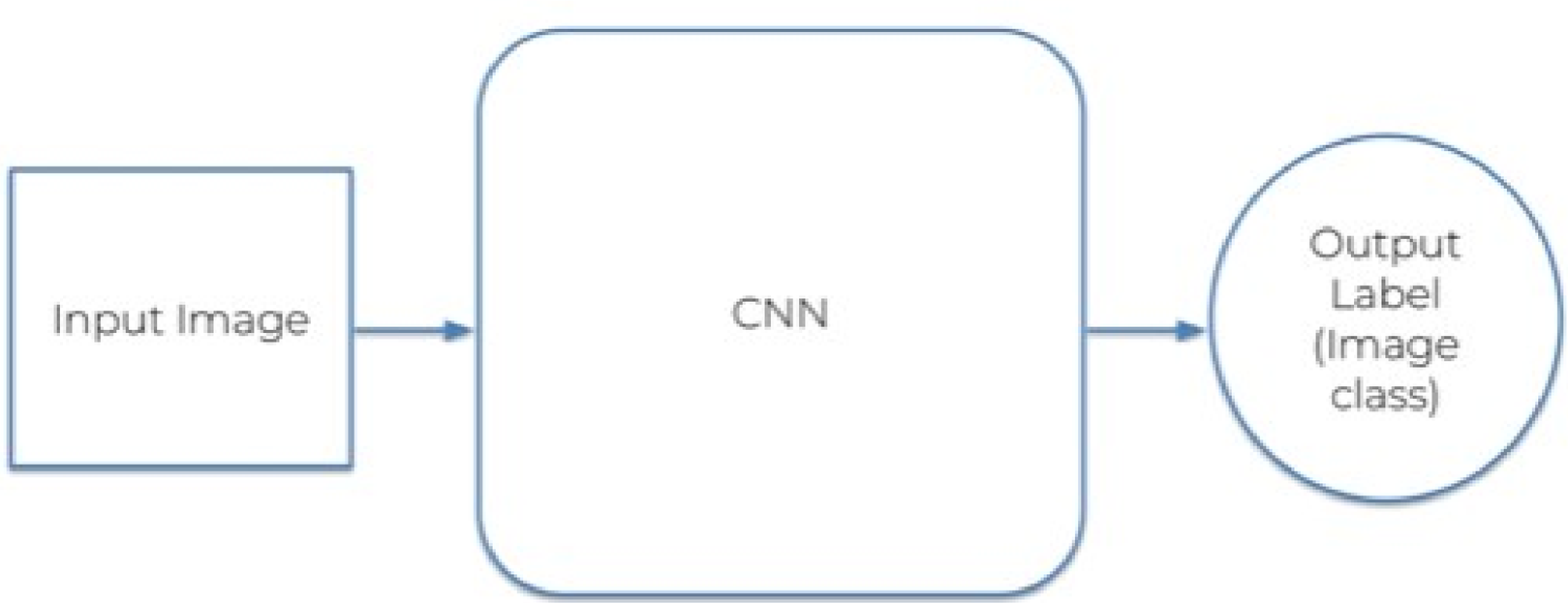
1) مقدمة	المحتويات	ما هو NLP	تاريخ NLP	العقبات و التحديات	التطبيقات				
2) أساسيات NLP	المكتبات	الملفات النصية	ملفات pdf	البحث في النصوص					
3) أدوات NLP	Tokenization	Sent. Segm.	POS	Stem & Lemm	NER	Stopwords	Matchers	Syntactic Struc.	T.Visualization
4)المعالجة البسيطة	Word Meaning	Word embed	T. Vectors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

القسم الثامن : تكنيكات حديثة

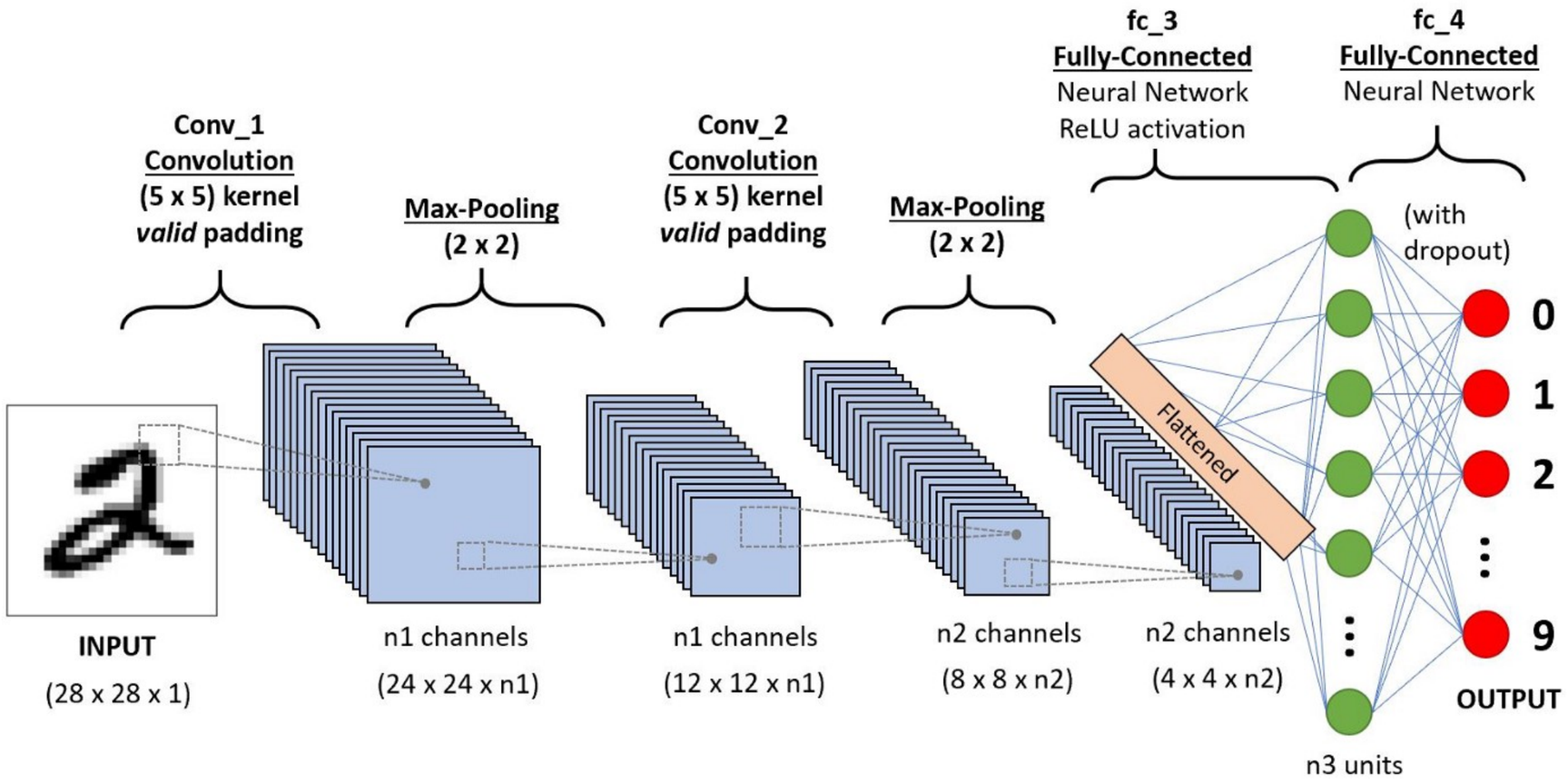
الجزء الثاني : CNN

==========

اولا علينا ان نتعرف علي اساسيات CNN , وهي تكون كالتالي :

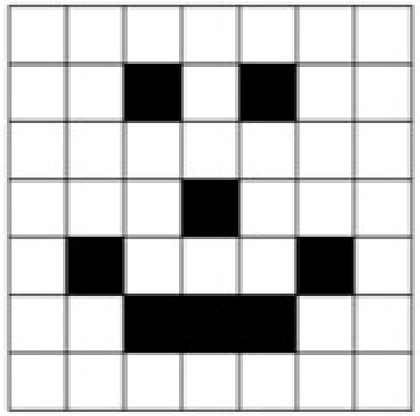


الصور تدخل ك input ثم يكون المخرج هو فئة الصورة
و تكون عبر الخطوات :



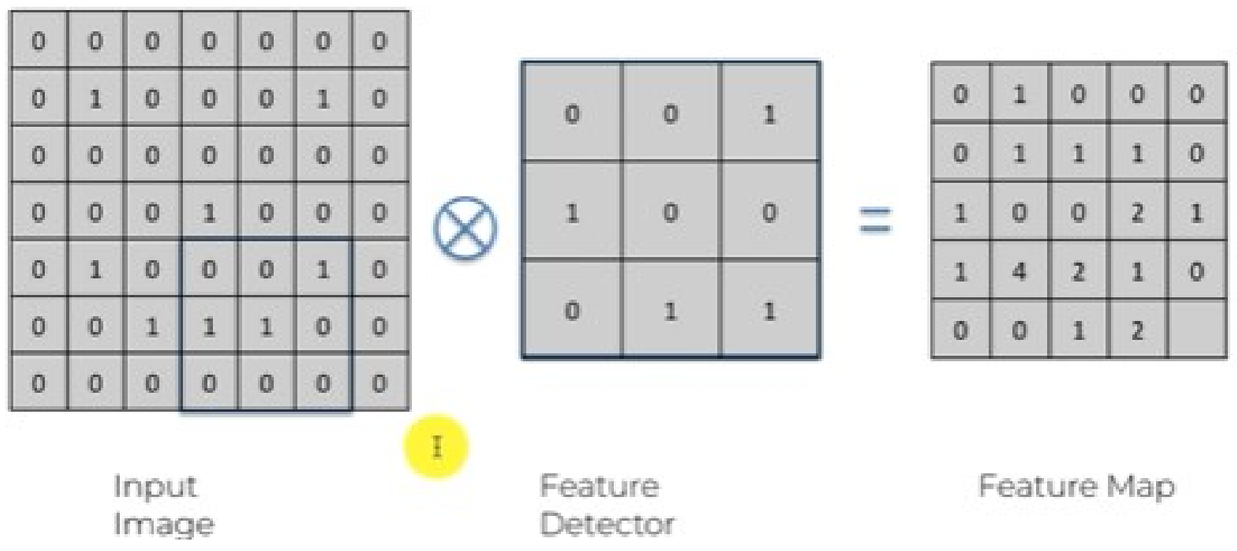
علي ان تكون المهمة الاساسية للخطوة الاولي convolutional هي التعرف علي الفيتشرز و تحديدها

و هنا شكل توضيحي لها :

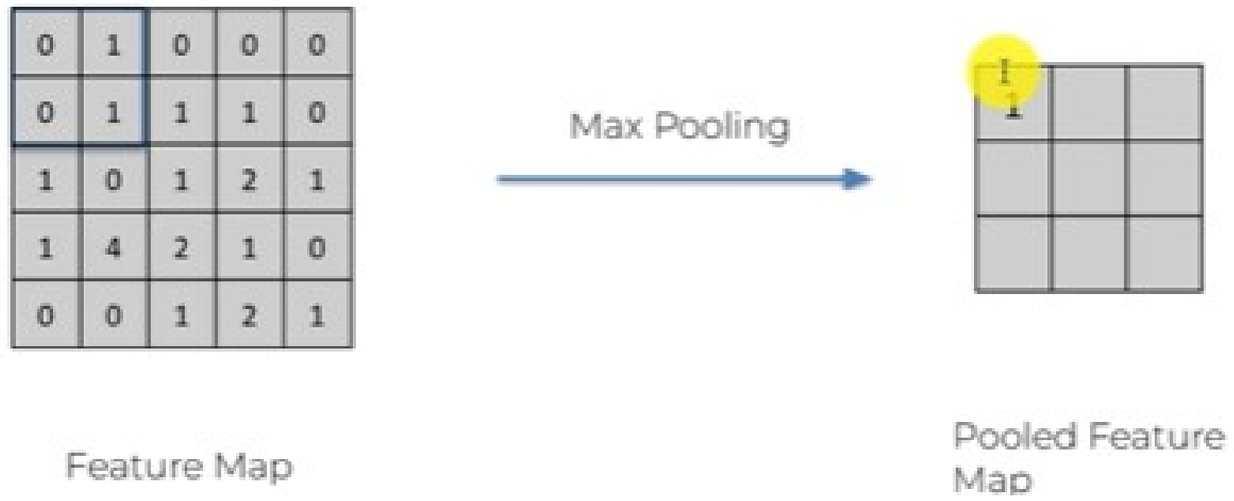


0	0	0	0	0	0	0
0	1	0	0	0	1	0
0	0	0	0	0	0	0
0	0	0	1	0	0	0
0	1	0	0	0	1	0
0	0	1	1	1	0	0
0	0	0	0	0	0	0

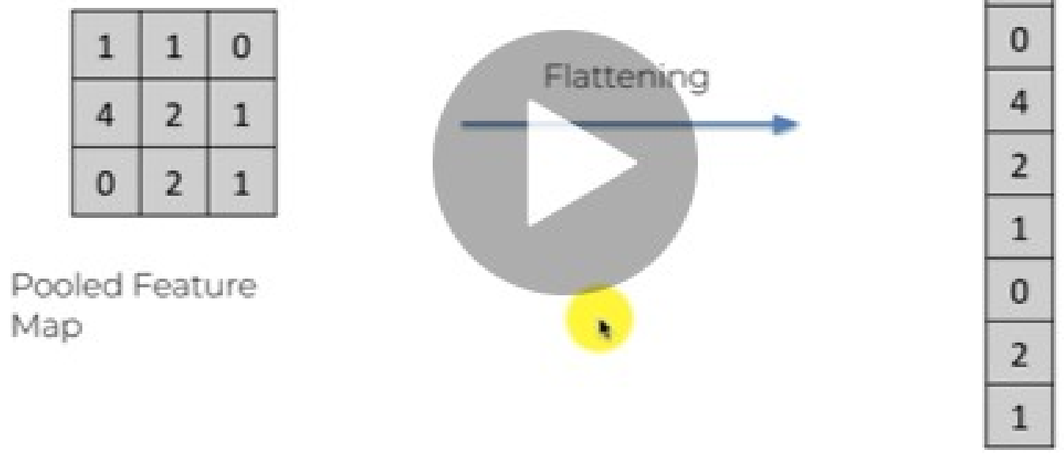
و تتم هكذا :



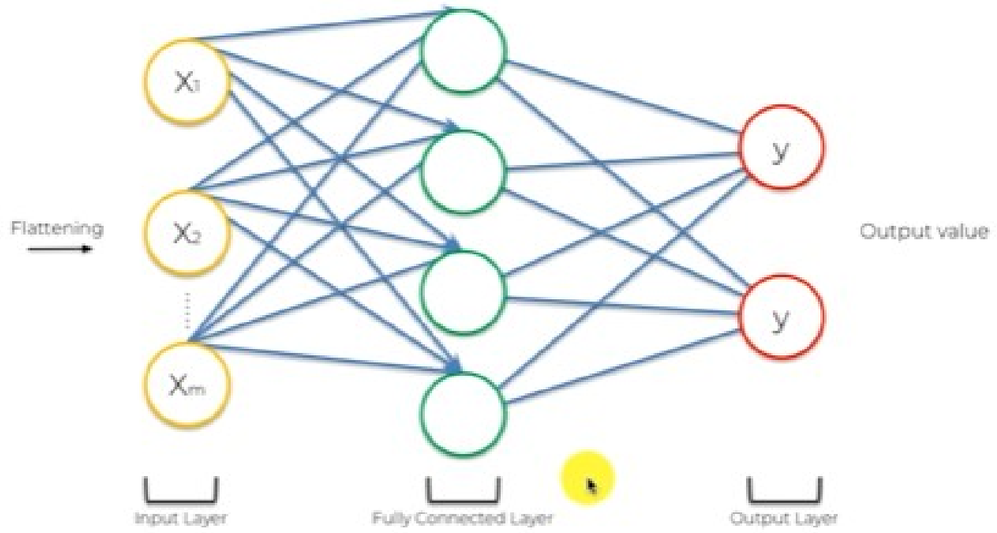
و خطوة الـ MaxPooling تهدف لتقليل حجم الداتا مع المحافظة علي الارقام الهامة



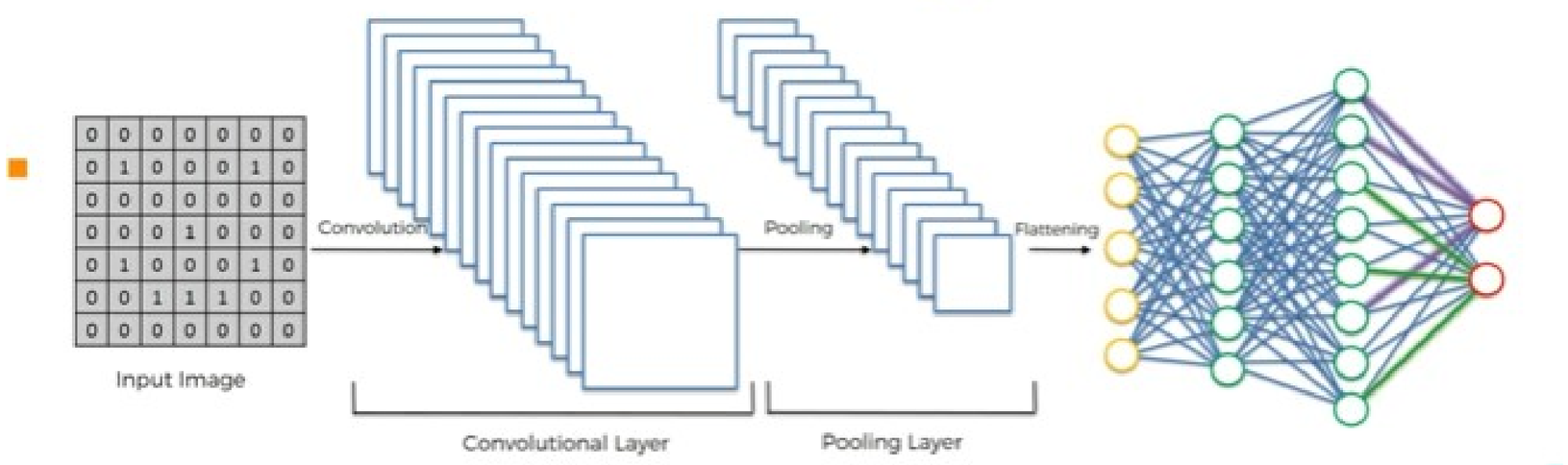
ثم خطوة الـ Flatten



ثم بعد ذلك الطبقات الخفية



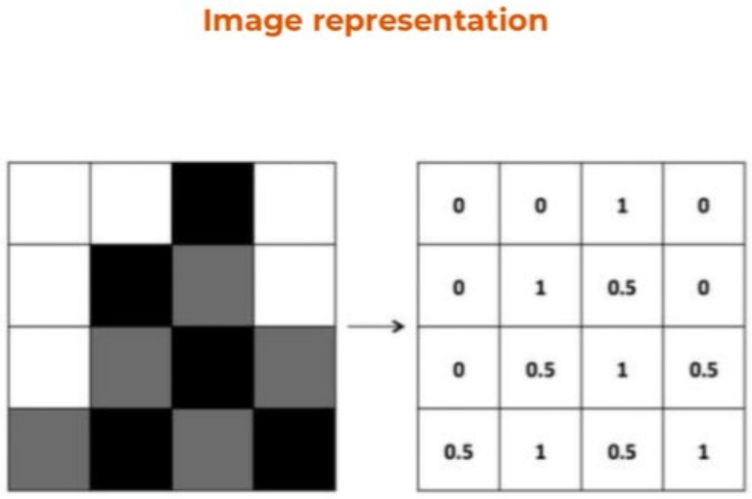
و هنا الخطوات كاملة

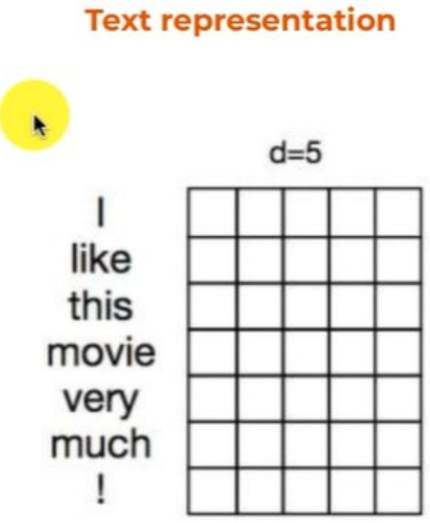


_

و هنا نأتي للسؤال الهام , اذا كانت CNN تعمل مع الصور , فكيف يمكن استخدامها مع النصوص . .

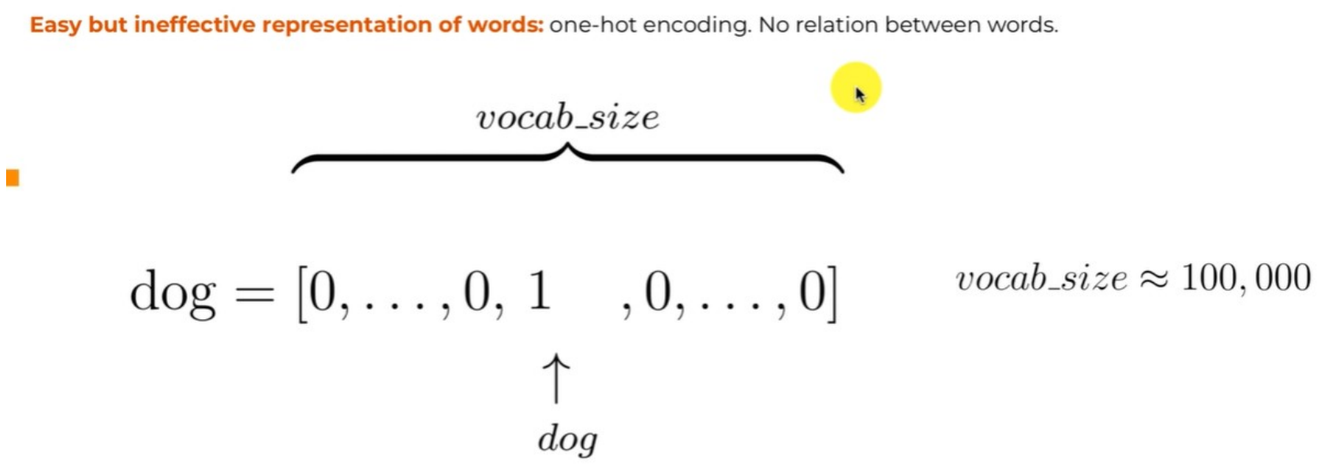
الـ CNN تقوم بالبحث عن الـ features في العناصر , وهو ما سنحاول تطبيقه كذلك مع النصوص , فالخطوة الاولي ان نقوم بتحويل النصوص الي مصفوفة ليتم التعامل معها بالـ CNN , وهو ما سيكون كالتالي :





فنري هنا كيف ان الجملة البسيطة I like this movie very much , والتي بها 6 كلمات , تكون مصفوفة بست صفوف و بها اعمدة حسب المعلومات التي سيتم استخراجها منها ,لكن كيف يمكن تحويل النصوص الي جدول

هناك الطريقة التقليدية , وهي طريقة one hot encoder مثل هذه , لكنها عقيمة و بها مشكلة ان عدد الاعمدة سيكون كبير جدا بناء علي عدد الكلمات و سيكون اغلب القيم اصفار , كما ان الارقام لا ترسم اي علاقات بين معاني الكلمات او مدي اقترابها او ابتعادها عن بعضها البعض



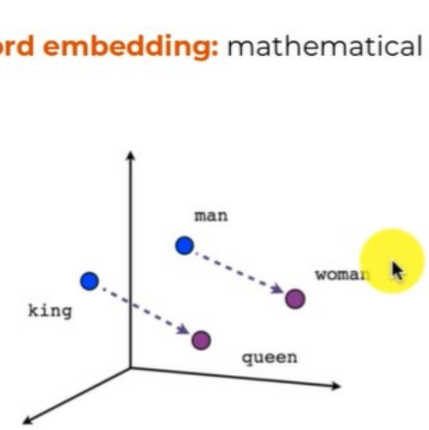
كما ان لدينا فكرة التضمين embedding حيث يكون هناك عدد من القيم (مثلا 64) حيث يكون هناك قيم تتراوح بين الصفر و الواحد في كل قيمة منهم

Word embedding: make each vector smaller -> adds relation between words.

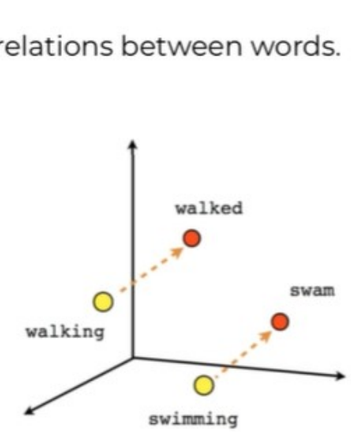
$$\text{dog} = \overbrace{[0.194, 0.047, \ldots, 0.126]}^{emb\_dim} \qquad emb\_dim \approx 64$$

و اهم ميزة لـ embedding انه يقوم بربط الكلمات معا و اظهار مدي التقارب او التباعد بين الكلمات , كذلك في الازمنة و المرادفات

Word embedding: mathematical relations between words.



Male-Female



Verb tense

[king] - [man] + [woman] = [queen]

[Paris] - [France] + [Italia] = [Rome]

_

ماذا عن عملية التدريب نفسها ؟ ؟

تكون الخطوة الاولي في ايجاد الكلمات المتجاورة في النص او المرتبطة ببعضهم البعض , ففي الجملة المشهورة لـ annie frank:

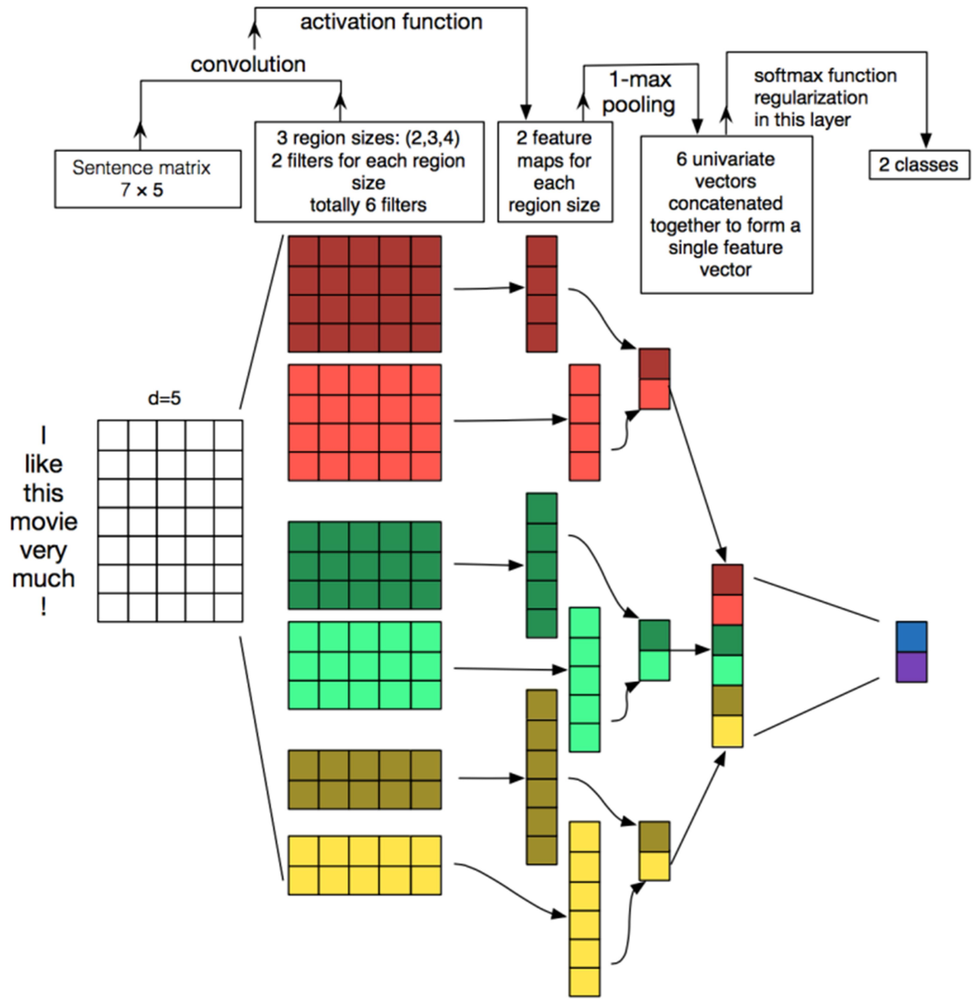
in spite of everything I still believe that people are really good at heart

نجد ان كلمة good مرتبطة بكلمات اربع هي : are , really , at , heart

فنقوم بإيجاد مصفوفة one hot encoder للجملة كلها و التي يكون فيها ارقام 1 و 0 حسب تواجد الكلمة , ثم يتم ضربها في مصفوفة embedding matrix حيث يكون هناك تضمين لكل كلمة من الكلمات , فحينما يتم ضرب الرقم 1 لكلمة ما (good مثلا) فيها يتم اختيار قيمها فقط , ثم يتم استخدام اداة softmax لتوقع الكلمة المطلوبة

_

ماذا عن كيفية استخدام CNN في التعامل مع NLP



و الإجابة عن السؤال تعتمد علي الإجابة علي سؤال آخر وهو : ما هو الفارق بين المصفوفة الناتجة من صورة و المصفوفة الناتجة عن نصوص ؟

في البداية علينا ان نتعرف علي مصفوفات الداتا العادية , فالصفوف تكون للـ sample size بينما الاعمدة تكون للمعلومات الخاصة بها features

بينما في مصفوفة الـ text تكون الصفوف لكل كلمة علي حدة , اما الاعمدة فهي لقيم التضمين embedding, وهذا ليس له نفس فكرة ال edges الموجودة في الصور , وبالتالي تطبيق فكرة filter 2*2 التي تقوم باللف علي الصورة , لن يكون لها معني منطقي .

لذا سيتم عمل فلتر يكون له عرض مماثل لعرض الجدول الاساسي , اي ان يكون بنفس عدد قيم ال embedding (الفلترات الملونة) , وسيكون اللف رأسيا فقط , من أعلي لأسفل لأن العرض هو نفسه .

و تكون العملية الرياضية كالتالي : الفتلر (البني وهو 4*5) يتم ضرب كل قيمه في اول اربع صفوف من مصفوفة النصوص و يكون له قيمة محددة , ثم يتحرك خطوة للاسفل و يتم الضرب و حساب القيمة , وهكذا , فنجد أن الناتج سيكون قيم فقط , وهي عدد خطوات النزول , لذا هناك مصفوفة 4*1 علي اليمين .

و يتم تكرار الامر مع اكثر من فلتر , وكل فلتر يوجد علي يمينه مصفوفة الناتج بنفس اللون .

و يتم تصميم الفلترز بحيث يكون هناك من يغطي 4 كلمات معا (الفلتر الاول والثاني) او يغطي ثلاث كلمات (الثالث و الرابع) او يغطي كلمتين (الخامس و السادس) , و عدد الفلترز يكون اضعاف هذه الارقام ,فقد يكون هناك 50 فلتر من كل تصميم .

الخطوة التالية هي عمل maxpooling لكل فئة , بحيث يتم اختيار اعلي قيمة في كل فئة او لون , ويتم رصهم معا في فيكتور واحد , وبالطبع لن نحتاج لخطوة ال flatten لان القيم كلها من بعد واحد

بعد ان يتم رص هذه القيم , يتم إدخالها في NN و يكون هناك عدد من المخرجات المطلوبة ليتم تدريبها عليها