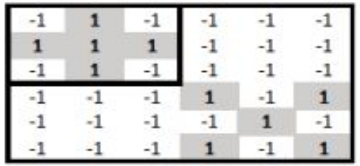
Pretrained Models και NLP

Ηρακλής Βαρλάμης

Περιεχόμενα

- Convolutional Neural Networks
- Σύνθετα μοντέλα
- Image captioning

Convolutional NNs

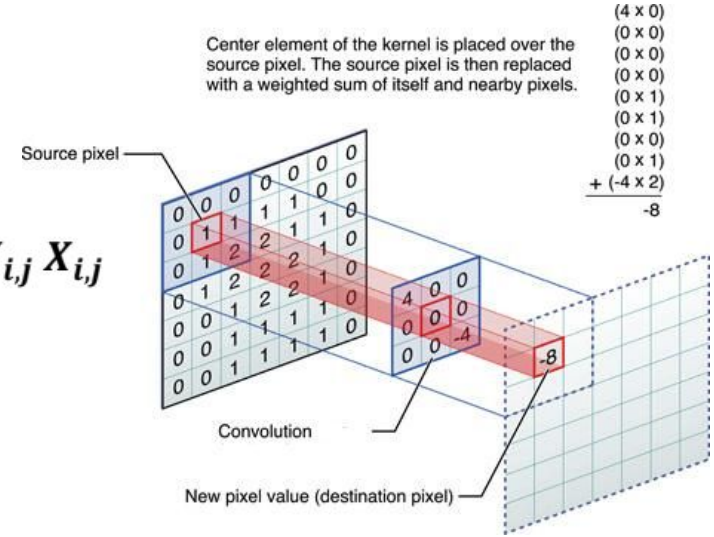


Input

Kernel

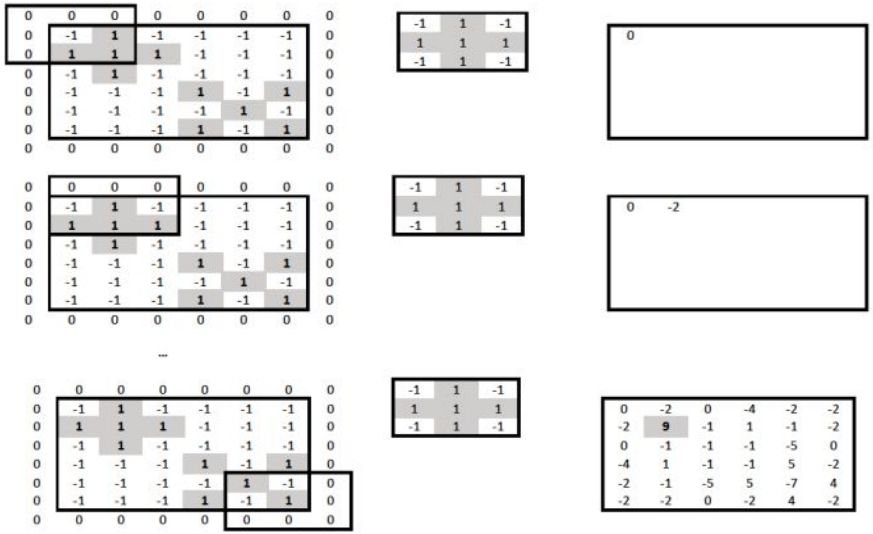
Feature

- Convolution σε εικόνες
- Εφαρμόζουμε ένα kernel μεγέθους 3x3 μετακινώντας τον σε όλη την εικόνα (διάστασης kxk) και καταλήγουμε σε μια νέα εικόνα (features) διαστάσεων (k-2)x(k-2)
- Η τιμή του γνωρίσματος ορίζεται ως $\sum_{i=1}^{3}\sum_{j=1}^{3} W_{i,j}\, X_{i,j}$



Convolutional NNs

- Wide convolution
- Παραγωγή εικόνας με τις αρχικές διαστάσεις με padding στην αρχική είσοδο
- Εφαρμόζοντας περισσότερα kernels μπορούμε να παράξουμε περισσότερα γνωρίσματα
- Τα kernels μαθαίνονται κατά την εκπαίδευση του NN

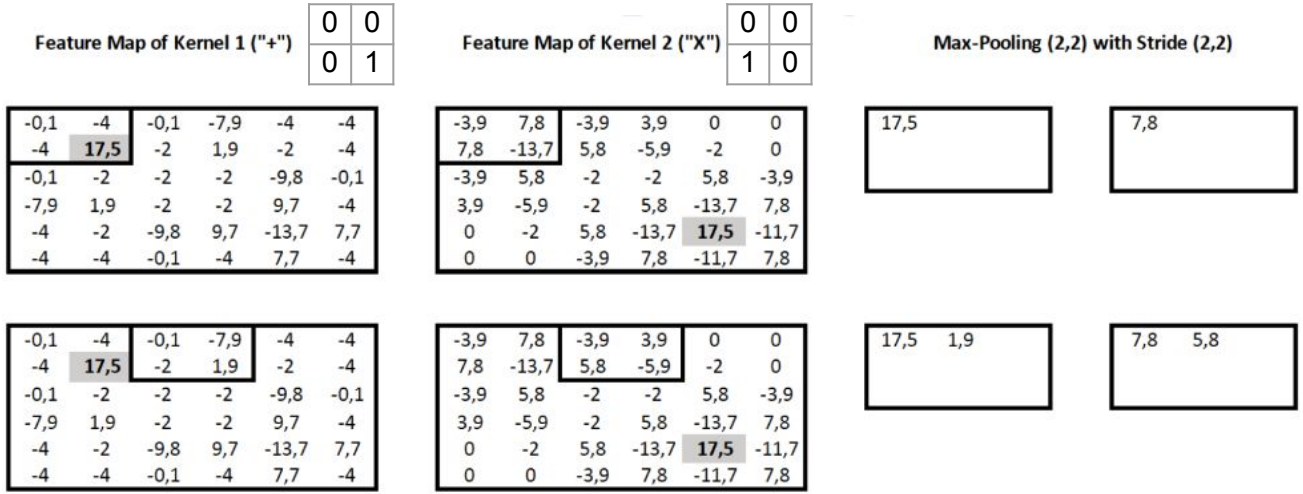


Convolutional layer

- Συνεπώς το συγκεκριμένο επίπεδο μπορεί να αποτυπώσει μια εικόνα στην είσοδο (με c κανάλια, π.χ. c=3 για RGB) με χρήση k kernels σε μια νέα εικόνα με τόσα κανάλια όσα τα γνωρίσματα που παράγονται (c*k).
- Μπορούμε να στοιβάξουμε (stack) περισσότερα convolutional layers
- Τα kernels είναι τα βάρη στο NN και αρχικοποιούνται σε τυχαίες τιμές.

Max-pooling layer

- Παρεμβάλλεται των stacked convolutional layers με στόχο να μειώσει τις διαστάσεις των (ενδιάμεσων) εικόνων (εικόνες με τα γνωρίσματα) που παράγονται.
- Βασικό χαρακτηριστικό και πάλι το kernel που μετακινείται πάνω στην εικόνα σε stride (οριζόντια και κατακόρυφη μετακίνηση σε κάθε βήμα).
- Αλλάζει/διευρύνει το πεδίο αντίληψης (receptive field).



Convolutions σε κείμενο

- Τα convolution layers μπορούν να μάθουν τα n-grams που έχουν σημασία στο όποιο task.
- Υποθέτουμε στο παράδειγμα ότι οι διαστάσεις των embeddings μας είναι γνωστές (π.χ. 4) και οι τιμές δυαδικές

Words	Embeddings			
	Subject	Positive	Stress	Quantity
I	**1**	0	0	0
like	0	**1**	0	0
this	0	0	0	0
movie	0	0	0	0
very	0	0	**1**	0
much	0	0	0	**1**
!	0	0	**1**	0

2 0
0 0
0 0
0 0
0 2
0 0

Global max pooling

2 2

kernels

Filter for “I like”, “we admire”…			
1	0	0	0
0	**1**	0	0

Filter for “very much”, “so much”…			
0	0	**1**	0
0	0	0	**1**

Εμφάνιση των filters σε όλη την πρόταση

Στην πράξη

- Τα φίλτρα μπορεί να είναι bigrams, trigrams κλπ
- Μαθαίνουμε περισσότερα από ένα φίλτρα

Words	Embeddings			
	d_1	d_2	d_3	d_4
I	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$x_{1,4}$
like	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$x_{2,4}$
this	$x_{3,1}$	$x_{3,2}$	$x_{3,3}$	$x_{3,4}$
movie	$x_{4,1}$	$x_{4,2}$	$x_{4,3}$	$x_{4,4}$
very	$x_{5,1}$	$x_{5,2}$	$x_{5,3}$	$x_{5,4}$
much	$x_{6,1}$	$x_{6,2}$	$x_{6,3}$	$x_{6,4}$
!	$x_{7,1}$	$x_{7,2}$	$x_{7,3}$	$x_{7,4}$

$$h_2 = \mathrm{ReLU}(Wx + b) \in \mathbb{R}^{3\times 1}$$

$$x^T = \langle x_{2,1}, x_{2,2}, \dots, x_{3,3}, x_{3,4} \rangle \in \mathbb{R}^{1\times 8}$$

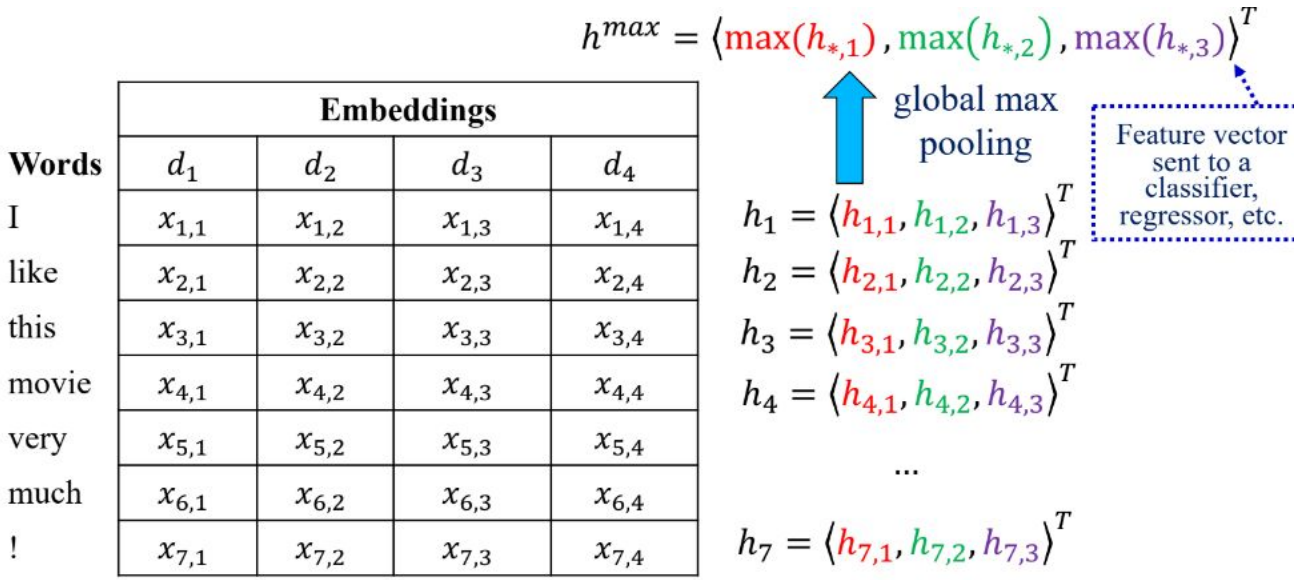
A bigram filter			
$w_{1,1}$	$w_{1,2}$	$w_{1,3}$	$w_{1,4}$
$w_{2,1}$	$w_{2,2}$	$w_{2,3}$	$w_{2,4}$

$$w = \langle w_{1,1}, w_{1,2}, w_{1,3}, \dots, w_{2,3}, w_{2,4} \rangle$$

$$b$$

$$W = \begin{bmatrix} w_{1,1,1} & w_{1,1,2} & w_{1,1,3} & \cdots & w_{1,2,3} & w_{1,2,4} \\ w_{2,1,1} & w_{2,1,2} & w_{2,1,3} & \cdots & w_{2,2,3} & w_{2,2,4} \\ w_{3,1,1} & w_{3,1,2} & w_{3,1,3} & \cdots & w_{3,2,3} & w_{3,2,4} \end{bmatrix} \in \mathbb{R}^{3\times 8} \qquad b = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} \in \mathbb{R}^{3\times 1}$$

Εφαρμογή των φίλτρων



Words	Embeddings			
	d_1	d_2	d_3	d_4
I	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$x_{1,4}$
like	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$x_{2,4}$
this	$x_{3,1}$	$x_{3,2}$	$x_{3,3}$	$x_{3,4}$
movie	$x_{4,1}$	$x_{4,2}$	$x_{4,3}$	$x_{4,4}$
very	$x_{5,1}$	$x_{5,2}$	$x_{5,3}$	$x_{5,4}$
much	$x_{6,1}$	$x_{6,2}$	$x_{6,3}$	$x_{6,4}$
!	$x_{7,1}$	$x_{7,2}$	$x_{7,3}$	$x_{7,4}$

$$W = \begin{bmatrix} w_{1,1,1} & w_{1,1,2} & w_{1,1,3} & \dots & w_{1,2,3} & w_{1,2,4} \\ w_{2,1,1} & w_{2,1,2} & w_{2,1,3} & \dots & w_{2,2,3} & w_{2,2,4} \\ w_{3,1,1} & w_{3,1,2} & w_{3,1,3} & \dots & w_{3,2,3} & w_{3,2,4} \end{bmatrix} \in \mathbb{R}^{3\times 8} \qquad b = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} \in \mathbb{R}^{3\times 1}$$

Περιεχόμενα

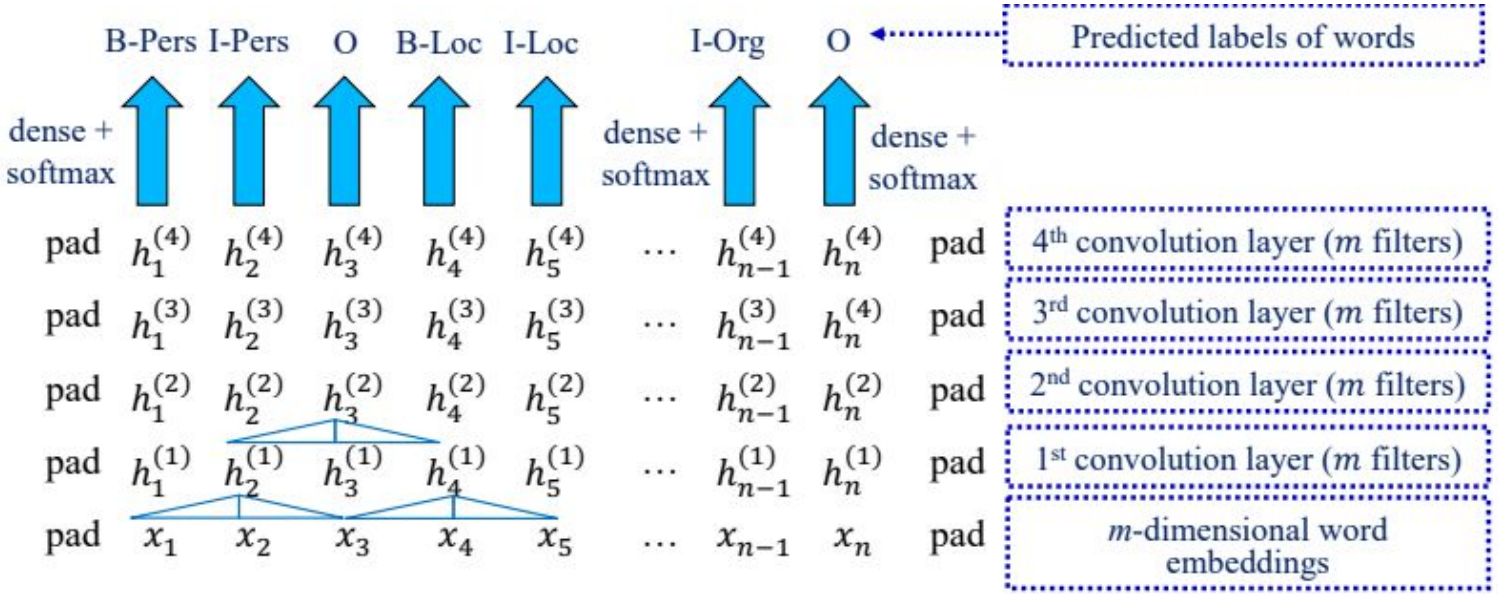
- Convolutional Neural Networks
- Σύνθετα μοντέλα
- Image captioning

Named entity recognition

- Εντοπισμός επωνύμων οντοτήτων στα κείμενα:
  - Άνθρωποι (person), τοποθεσίες (location), οργανισμοί (organization), γεωπολιτικές οντότητες (GPE), υποδομές (facility), μάρκες οχημάτων (vehicle), γονίδια (gene) κλπ.
  - Μπορεί να αποτελούνται από περισσότερες λέξεις
- Επίλυση ασαφειών: Paris (η πόλη ή η Paris Hilton)
- Πιο σπάνια: εντοπισμός χρονικών εκφράσεων, γεγονότων
- Μπορεί να λυθεί ως πρόβλημα κατηγοριοποίησης ή ως πρόβλημα sequence labeling

Words	IOB Label	IO Label
American	B-ORG	I-ORG
Airlines	I-ORG	I-ORG
,	O	O
a	O	O
unit	O	O
of	O	O
AMR	B-ORG	I-ORG
Corp.	I-ORG	I-ORG
,	O	O
immediately	O	O
matched	O	O
the	O	O
move	O	O
,	O	O
spokesman	O	O
Tim	B-PER	I-PER
Wagner	I-PER	I-PER
said	O	O
.	O	O

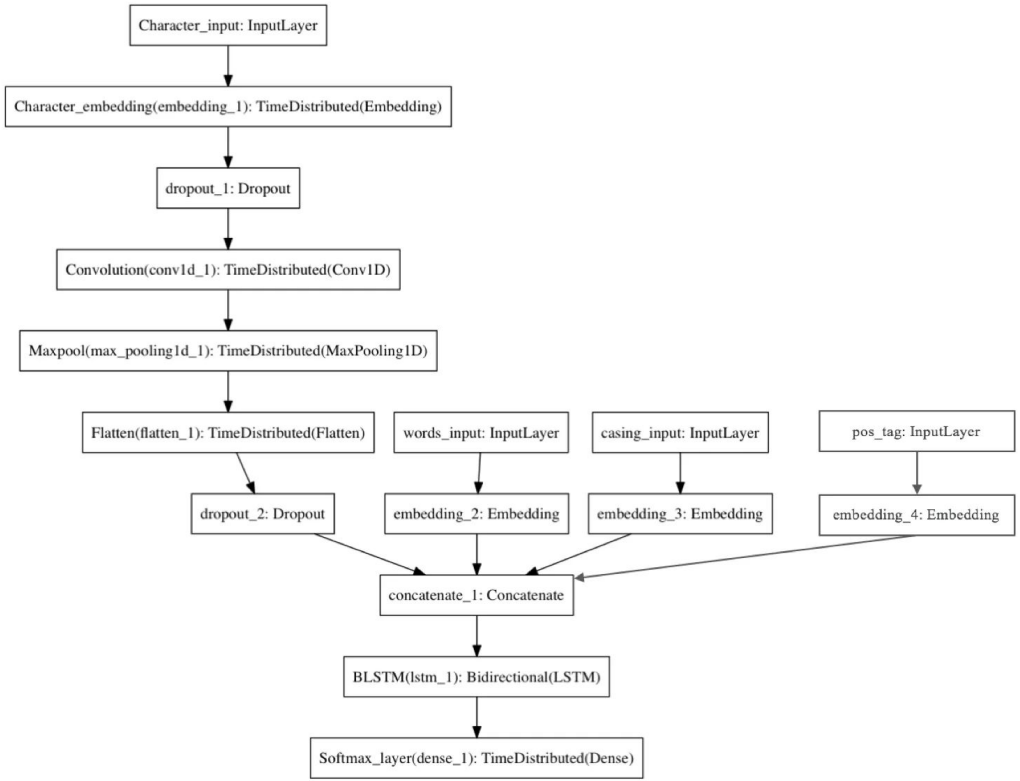
CNNs for token classification



$$h_i^{(1)} = \mathrm{ReLU}\left(W^{(1)}[x_{i-1}; x_i; x_{i+1}] + b^{(1)}\right) + x_i \in \mathbb{R}^{m\times 1}$$

$$h_i^{(j)} = \mathrm{ReLU}\left(W^{(j)}\left[h_{i-1}^{(j-1)}; h_i^{(j-1)}; h_{i+1}^{(j-1)}\right] + b^{(j)}\right) + h_i^{(j-1)} \in \mathbb{R}^{m\times 1}$$

Σύνθετα δίκτα με CNN και RNN

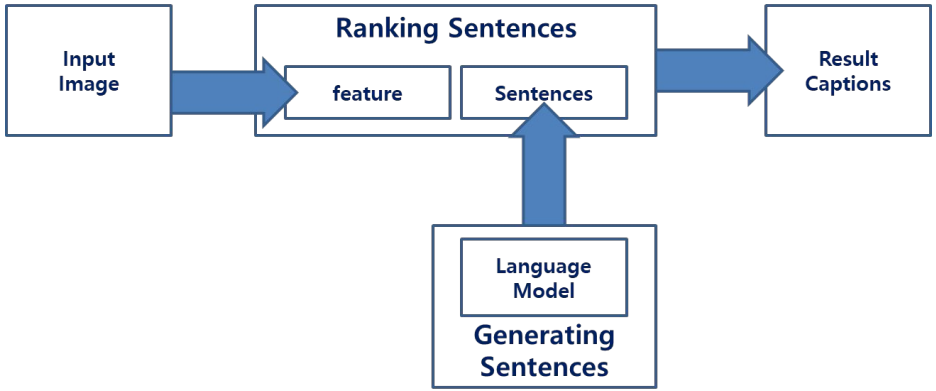


Περιεχόμενα

- Convolutional Neural Networks
- Σύνθετα μοντέλα
- Image captioning

Ορισμός

- Σε ένα δεδομένο χώρο προτάσεων $S=V^N$
- Βρες προτάσεις για μια συγκεκριμένη εικόνα
- Brute force
  - Βρες ένα υποσύνολο πιθανών προτάσεων
  - Υπολόγισε το σκορ κάθε πιθανής πρότασης
  - Κράτησε αυτές που περνούν το κατώφλι



Στατιστικό γλωσσικό μοντέλο

- Σχετική πρόταση = πιθανή πρόταση
- Αναθέτει μια πιθανότητα σε κάθε πρόταση με βάση την ακολουθία λέξεων

$p(w_N, w_{N-1},..., w_1) = \Pi^N_{i=1} p(w_i|w_{i-1}, ..., w_1) = p(w_1)p(w_2|w_1)p(w_3 | w_2, w_1)$ …
$p(w_N | W_{N-1},...w_1)$

π.χ. A woman holding a camera in a crowd.

p(a)p(woman|a)p(holding|woman,a)...p(crowd|a,...a)

- Ουσιαστικά μόνο οι τελευταίες λέξεις καθορίζουν την πιθανότητα της επόμενης

p(crowd|a,in,camera,a,holding,woman,a) ~p(crowd|a, in, camera)

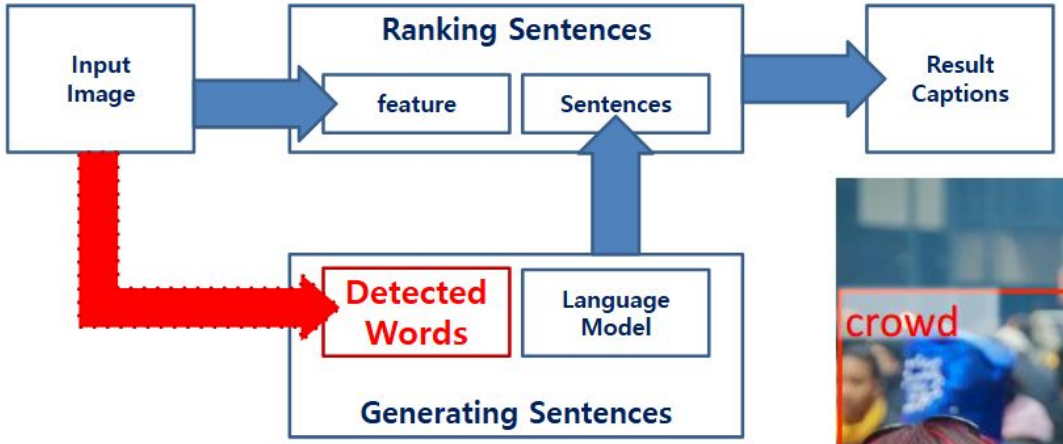
Τρόποι αναζήτησης (δημιουργίας) της περιγραφής

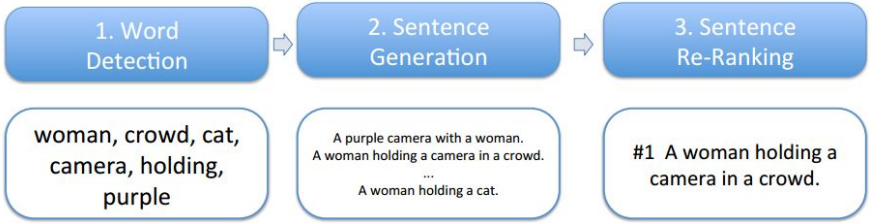
- Greedy search

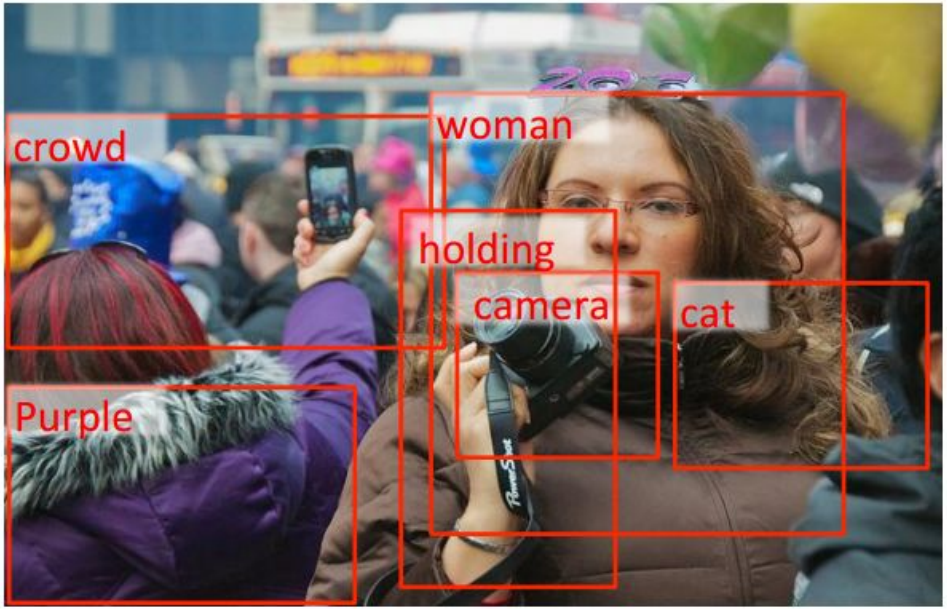
$$p(w_N, w_{N-1}, \cdots, w_1) = \prod_{i=1}^{N} p(w_i | w_{i-1}, \cdots, w_1)$$
$$= p(w_1)p(w_2|w_1)p(w_3|w_2, w_1) \cdots p(w_N|w_{N-1}, \cdots w_1)$$

  - Και παράγει την επόμενη λέξη που έχει τη μεγαλύτερη πιθανότητα
- Beam search
  - Σαν το greedy μόνο που κρατά κάθε φορά τα k-best μονοπάτια

From Captions to Visual Concepts and Back (CVPR 2015)







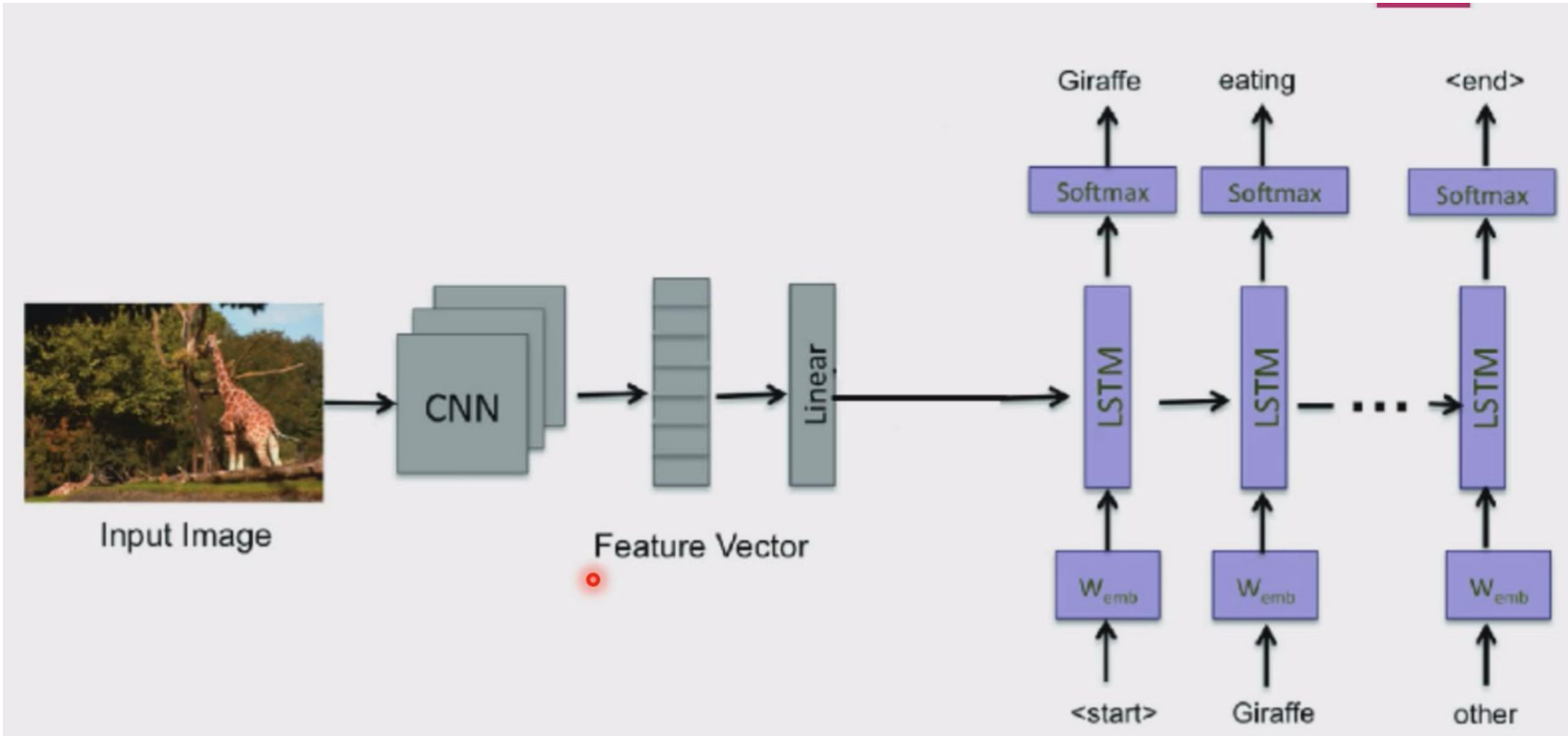
Συνδυασμός NLP και Computer Vision (+audio)

- Περιγραφή ιατρικών (και άλλων) εικόνων
- Λεκτική περιγραφή video (σε ανθρώπους με προβλήματα όρασης)
- Μετατροπή λεκτικών περιγραφών σε εικόνες ή βίντεο

- Text-to-speech, speech-to-text
- Μετατροπή νοηματικής σε λόγο ή κείμενο και αντίστροφα

- Νέοι τρόποι για τη διάδραση ανθρώπου-μηχανής

Παράδειγμα


Giraffe
eating
<end>
Softmax
Softmax
Softmax
LSTM
LSTM
LSTM
Linear
CNN
Input Image
Feature Vector
W_emb
W_emb
W_emb
<start>
Giraffe
other