**Τεχνητή Νοημοσύνη**

# Εισαγωγή στην υπολογιστική όραση

**Τμήμα Μηχανικών Η/Υ και Πληροφορικής**
**Πανεπιστήμιο Πατρών**

**Δ. Κοσμόπουλος**

# Άνθρωπος vs υπολογιστής

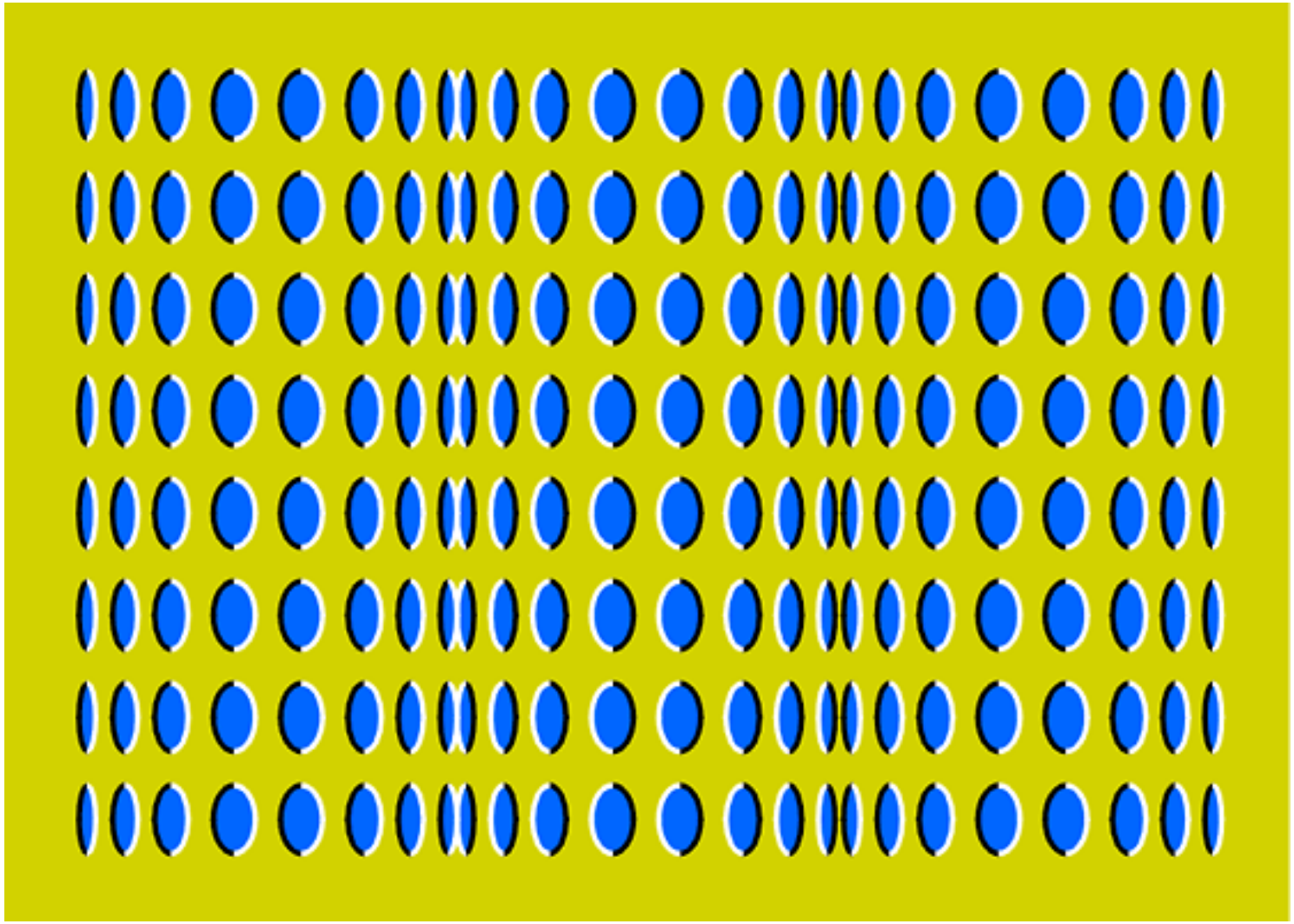

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 0 | 3 | 2 | 5 | 4 | 7 | 6 | 9 | 8 |
| 3 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 2 | 1 | 0 | 3 | 2 | 5 | 4 | 7 | 6 |
| 5 | 2 | 3 | 0 | 1 | 2 | 3 | 4 | 5 |
| 4 | 3 | 2 | 1 | 0 | 3 | 2 | 5 | 4 |
| 7 | 4 | 5 | 2 | 3 | 0 | 1 | 2 | 3 |
| 6 | 5 | 4 | 3 | 2 | 1 | 0 | 3 | 2 |
| 9 | 6 | 7 | 4 | 5 | 2 | 3 | 0 | 1 |
| 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | 0 |

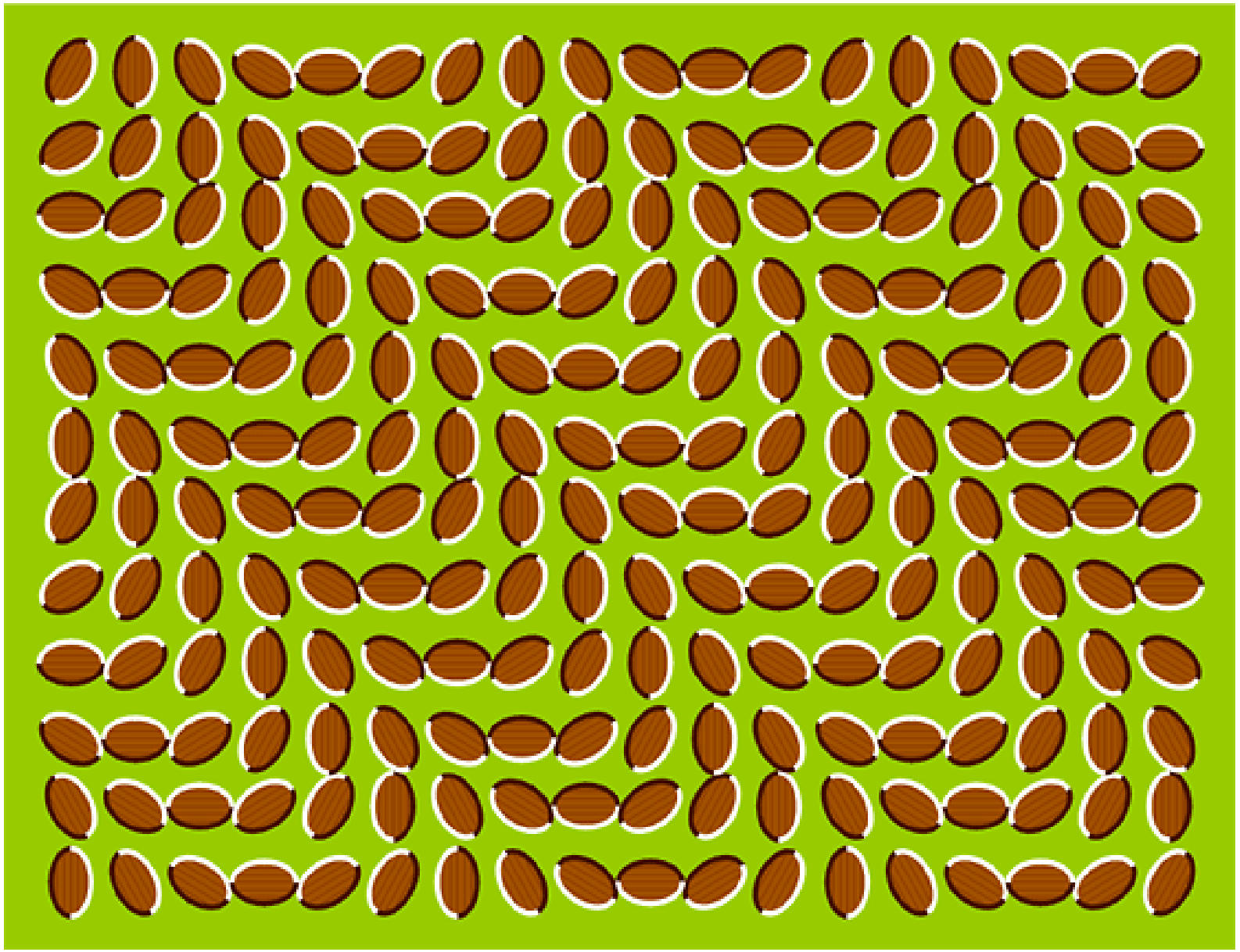
Ποιος είναι καλύτερος;

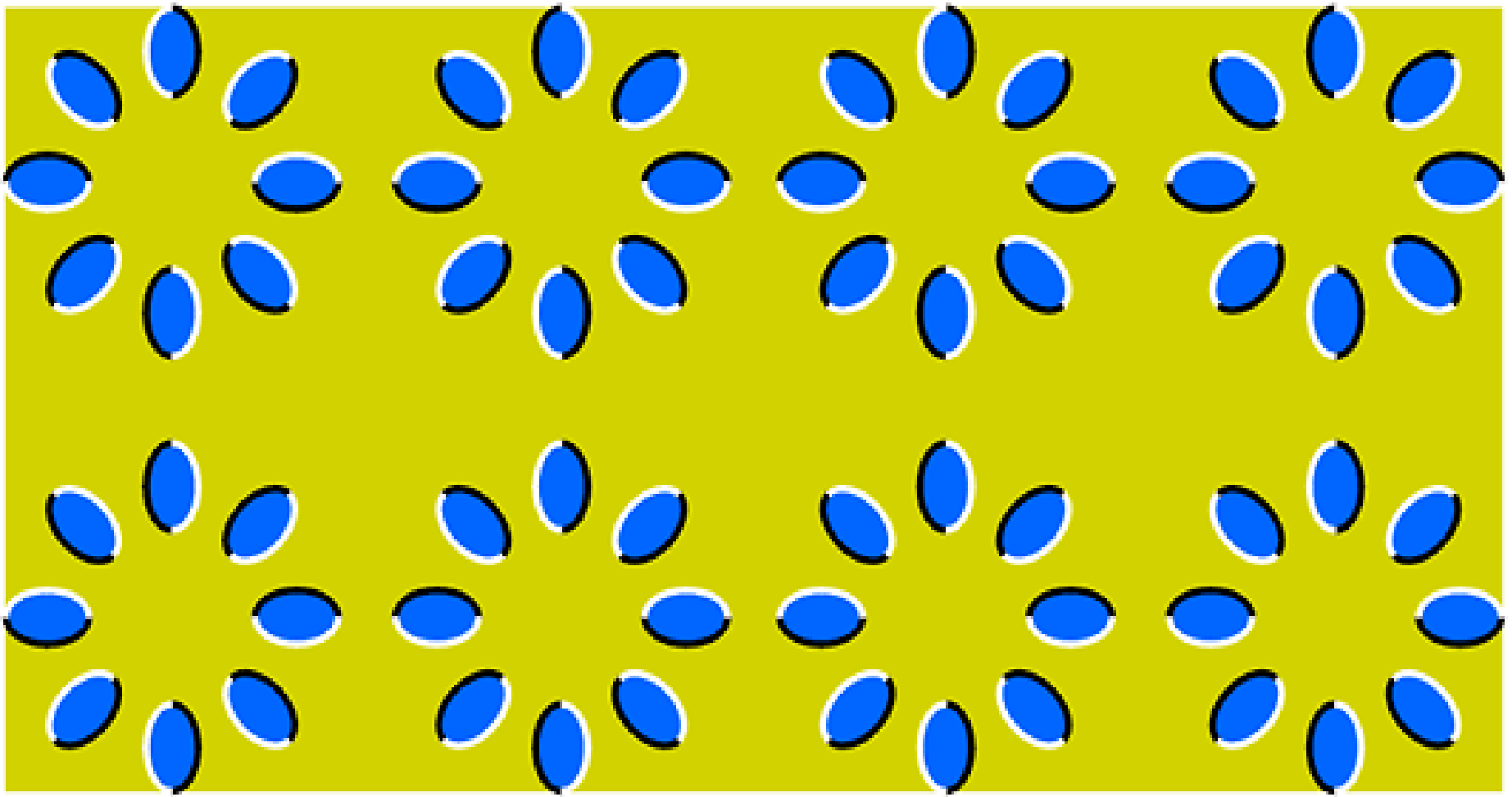
# Άνθρωπος vs υπολογιστής

Ιδανικά η μηχανή θα πρέπει να «βλέπει» και να αντιλαμβάνεται ό,τι και ο άνθρωπος.

Ή μήπως όχι;







# Το φόρεμα

https://en.wikipedia.org/wiki/The_dress

# Υπολογιστική όραση

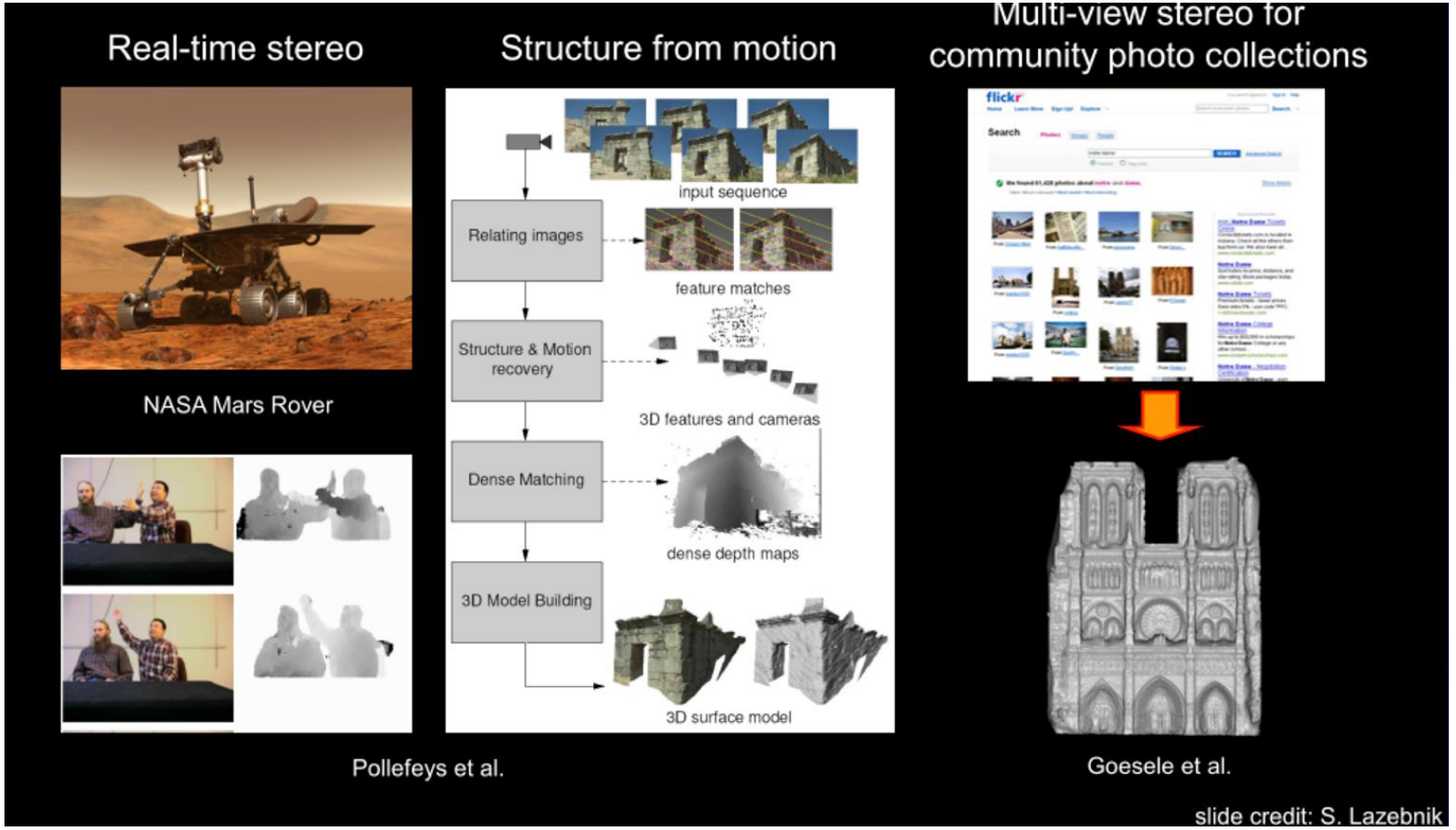

Κατανόηση της ιστορίας

Εξαγωγή σημασιολογικής πληροφορίας

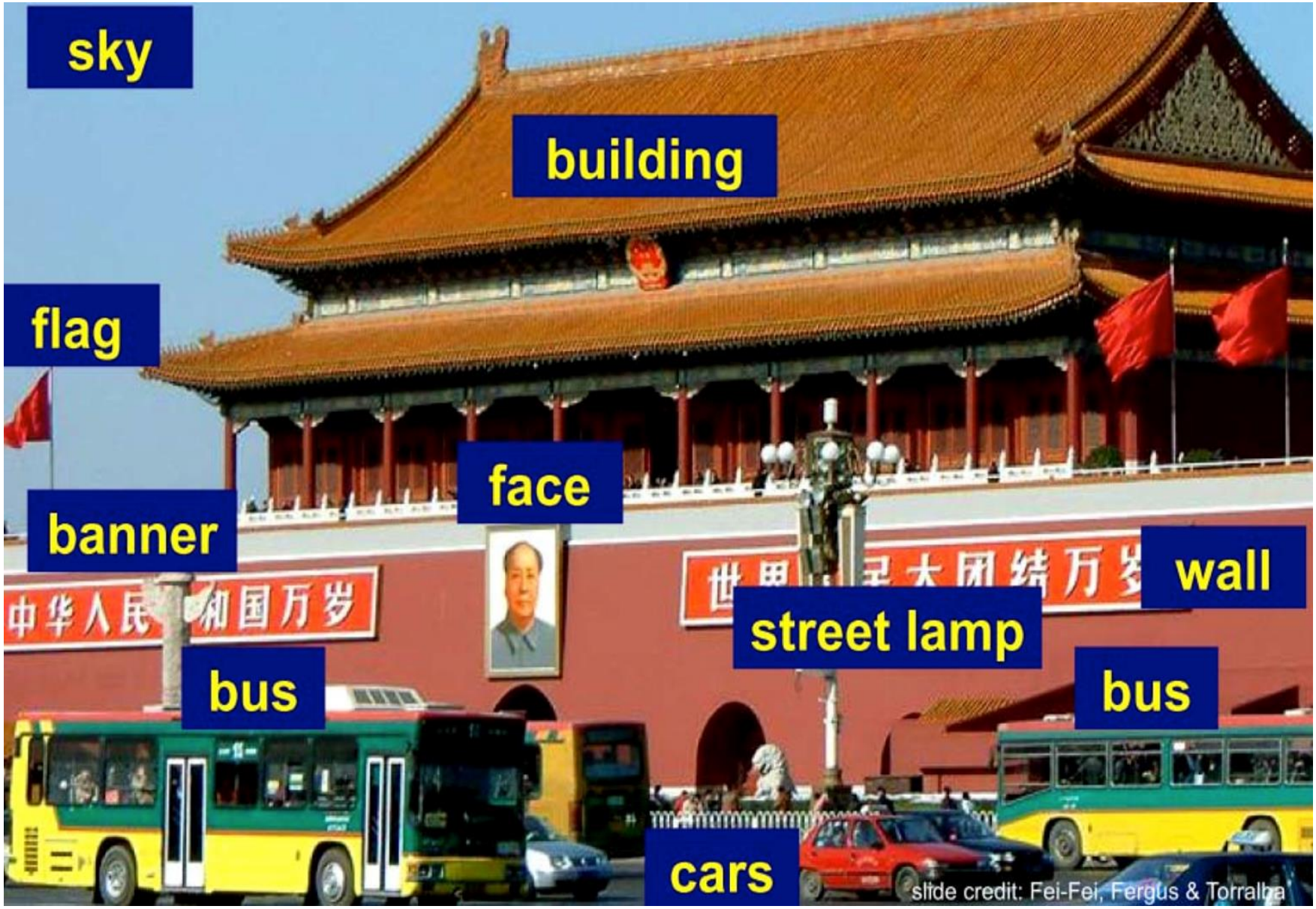
Εξαγωγή μετρήσεων από τον 3Δ κόσμο

Παραγωγή αληθοφανών εικόνων - βίντεο

# Εξαγωγή 3Δ πληροφορίας

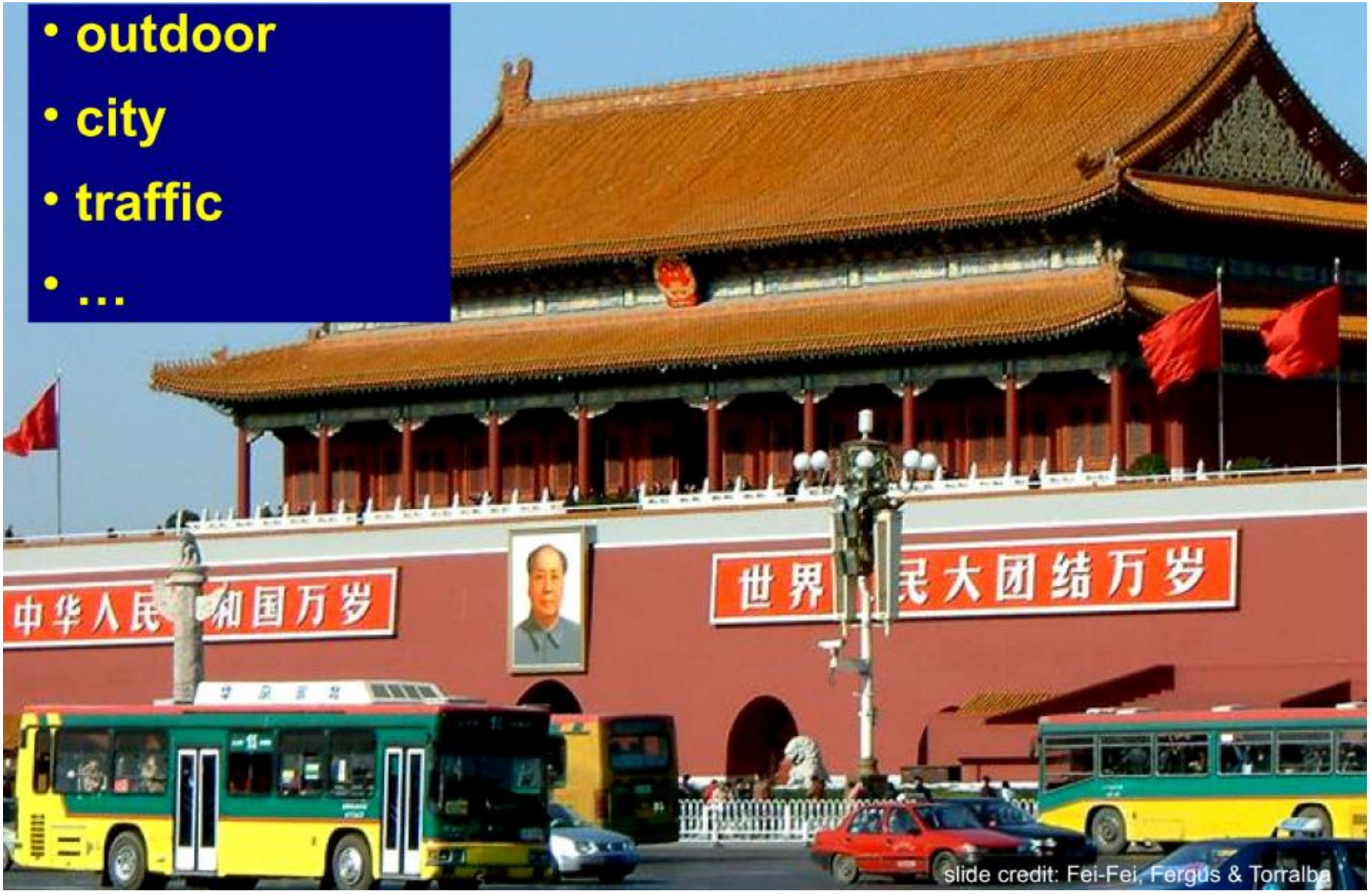


# Σημασιολογική πληροφορία



# Σημασιολογική πληροφορία

# Σημασιολογική πληροφορία



# Γιατί;


Personal photo albums
Movies, news, sports
Google Image Search
Picasa
flickr
webshots
picsearch
YouTube
Broadcast Yourself
Surveillance and security
Medical and scientific images
slide credit: S. Lazebnik

# Γιατί;

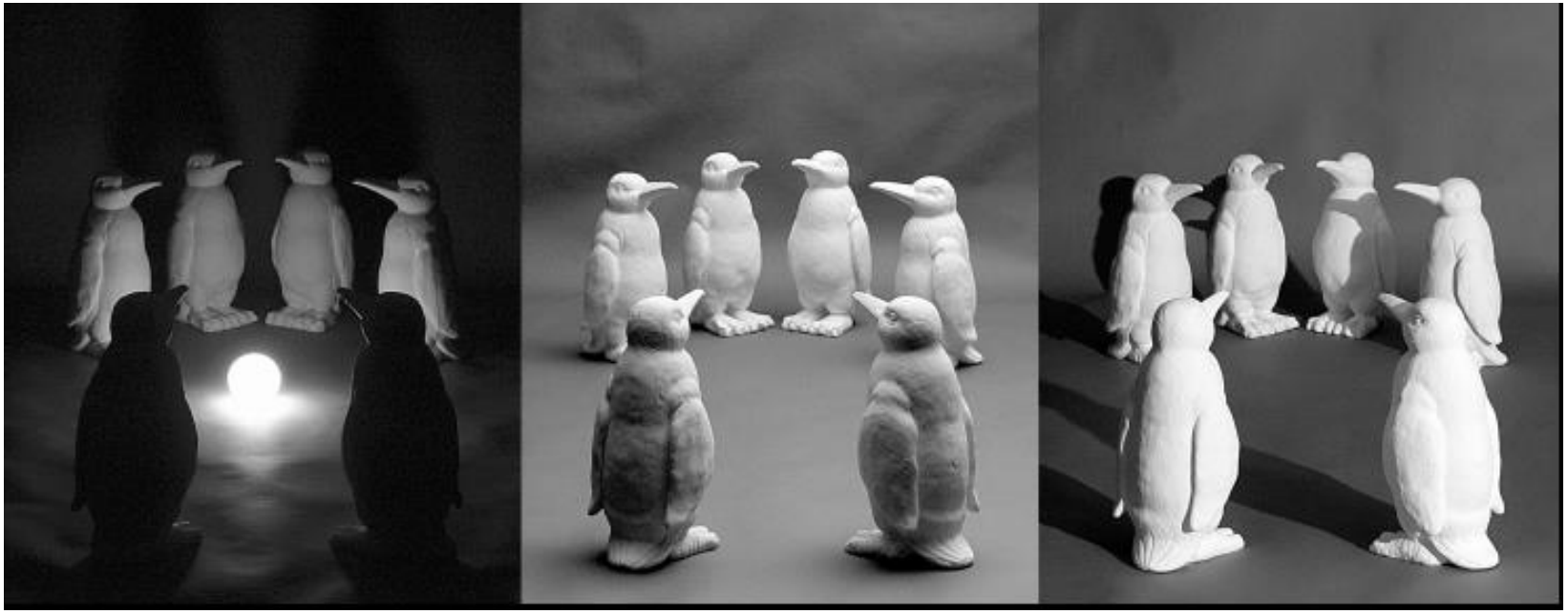
Η υπολογιστική όραση είναι παντού

Είναι ενδιαφέρουσα

Παρουσιάζει προκλήσεις

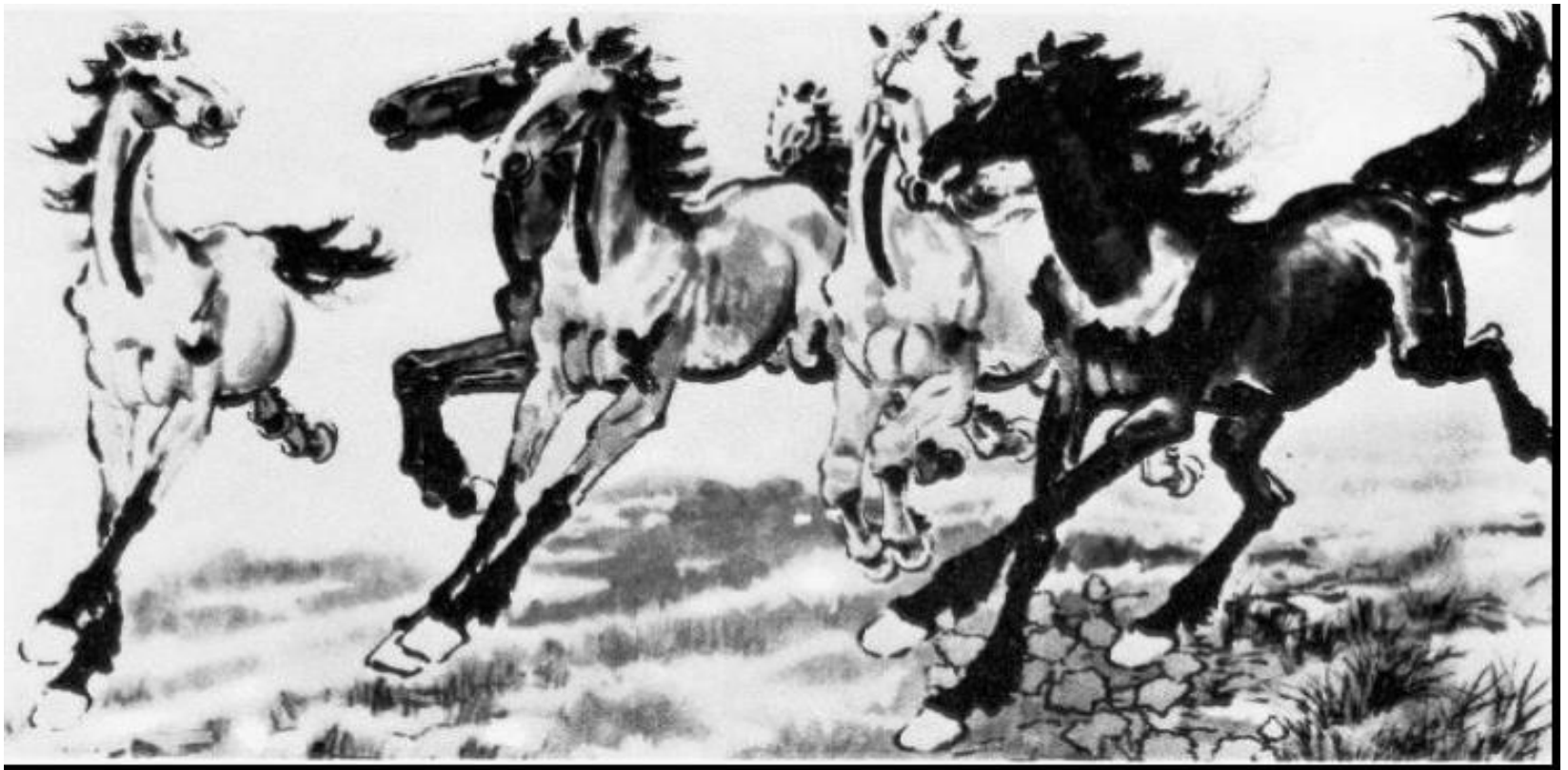
# Γωνία θέασης

# Φωτισμός



# Κλίμακα

# Παραμόρφωση



Xu, Beihong 1943

slide credit: Fei-Fei, Fergus & Torralba

# Επικαλύψεις

# Οπτικό υπόβαθρο


Slide credit: S. Lazebnik

# Γρήγορη κίνηση

# Μεταβλητότητα

# Μη διακριτικά χαρακτηριστικά

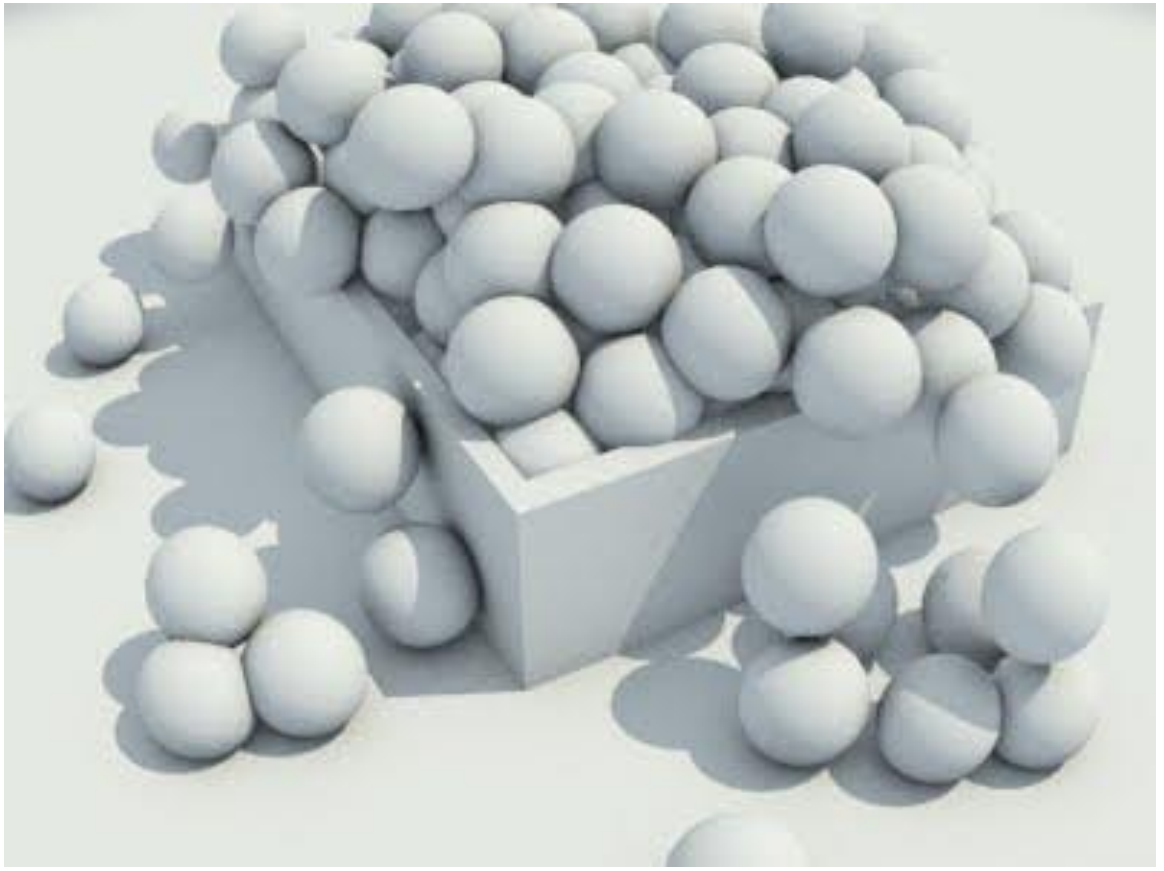
# Δυσκολίες ή ευκαιρίες

- Η ερμηνεία εικόνων – βίντεο είναι δύσκολη
- Υπάρχει πληθώρα χαρακτηριστικών
- Διερεύνηση / εκμετάλλευση αυτών!

# Εκτίμηση βάθους: προοπτική

Slide credit: S. Lazebnik

# Κατηγοριοποίηση βάθους: επικαλύψεις



# Εξαγωγή μορφής: υφή


Slide credit: S. Lazebnik

# Εξαγωγή θέσης: σκιά

# Ομαδοποίηση: χρώμα, υφή, εγγύτητα

Slide credit: S. Lazebnik

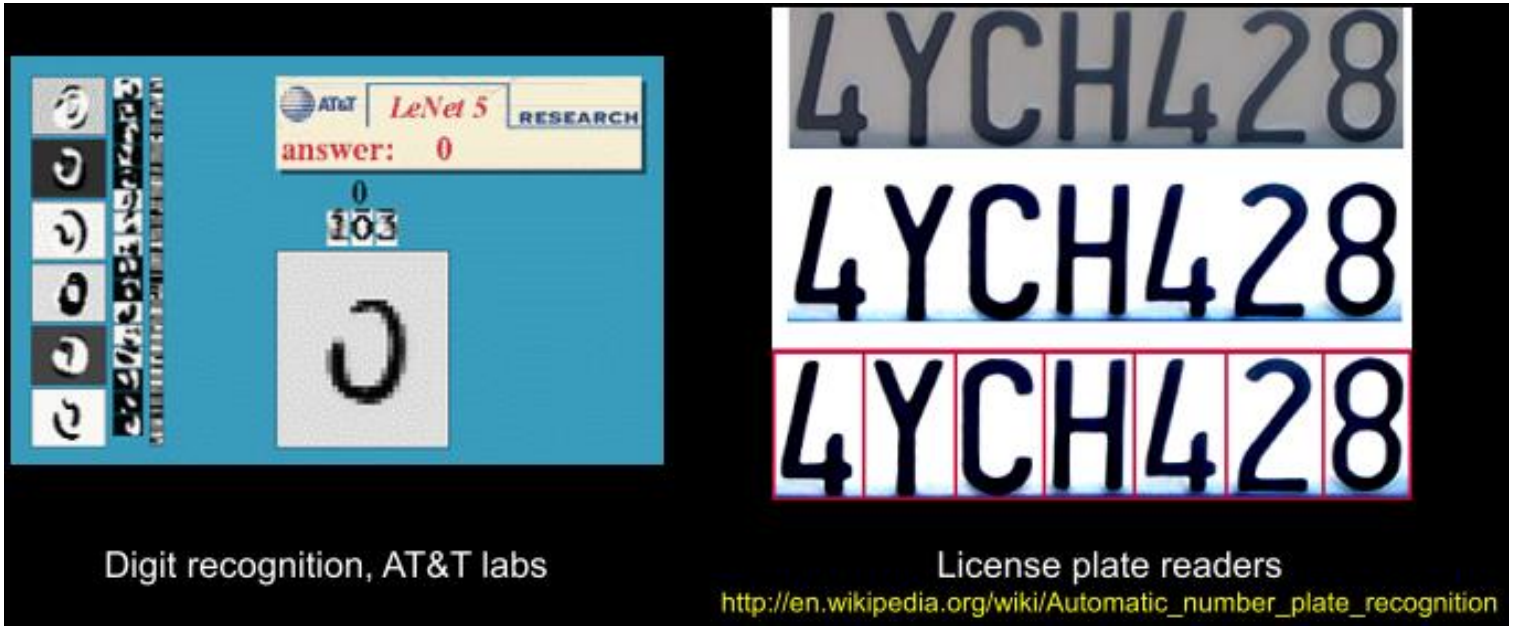
# Ταξινόμηση κίνησης: κοινή συμπεριφορά

# Μονοδιάστατη εικόνα / βίντεο

- Μερικές φορές παραπλανητικό
- 2D-3D αντιστοιχία μερικές φορές δύσκολη να εξαχθεί
- Χρειαζόμαστε πρότερη γνώση

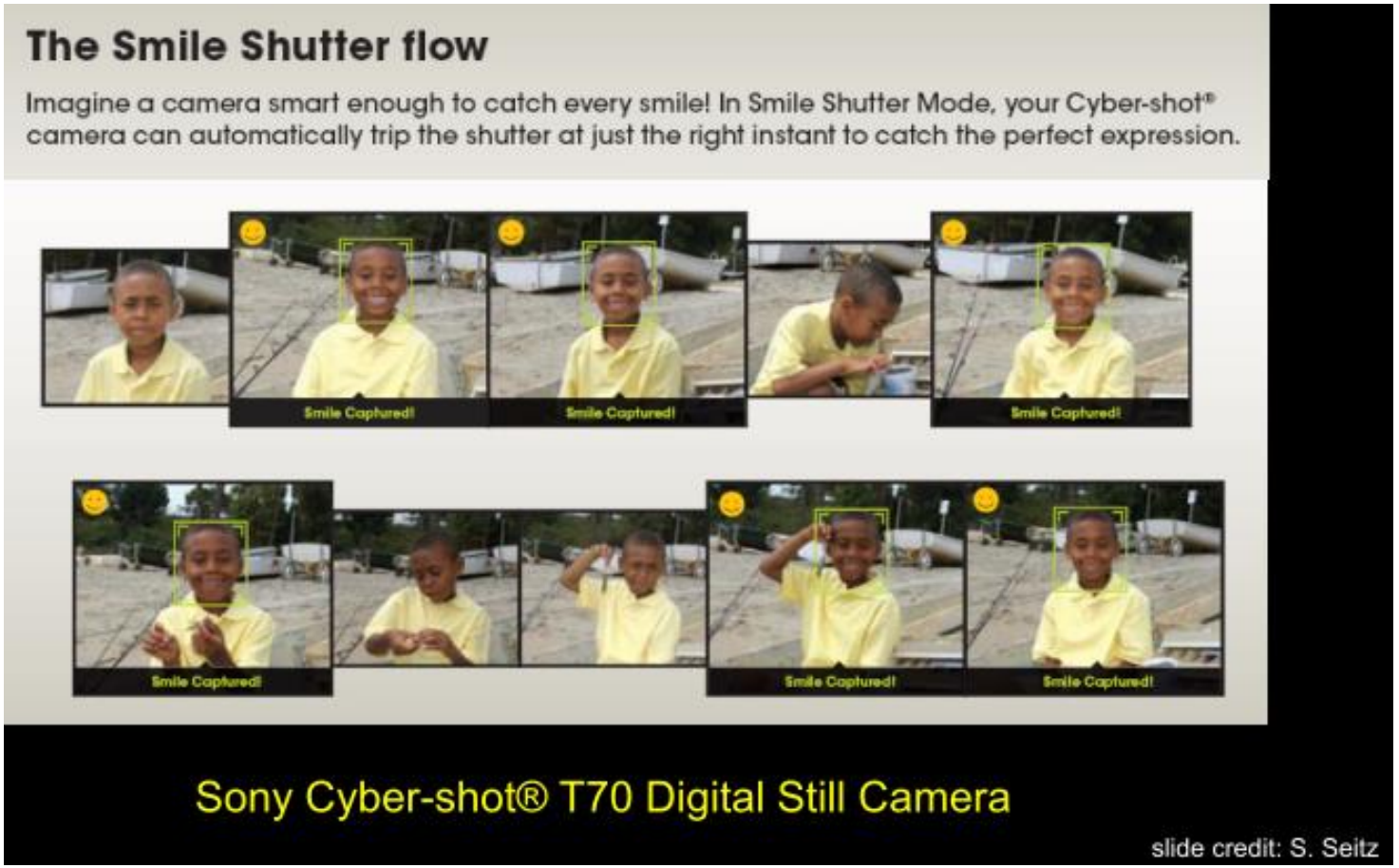
# Σχετικά Προϊόντα

# Αναγνώριση χαρακτήρων



# Παρακολούθηση προσώπου

# Ανίχνευση χαμόγελου



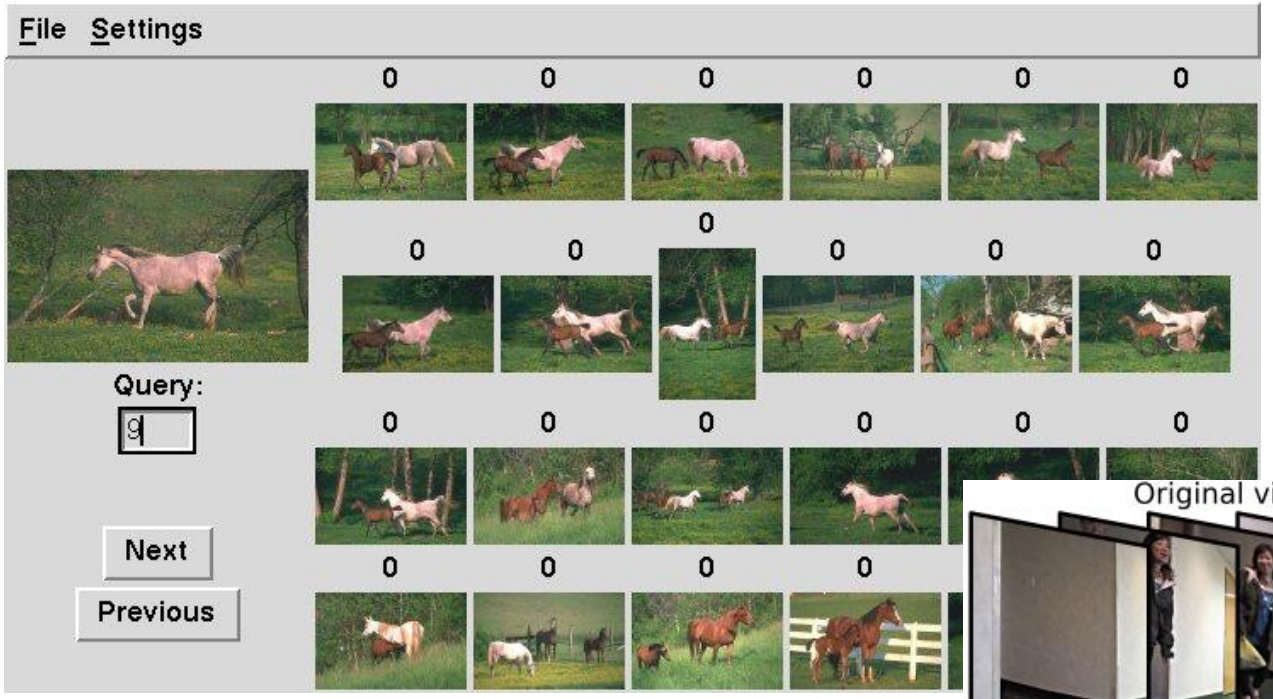
# Αθλητισμός


OAK 7 4TH
TEN 14 11:04
Sportvision first down line
Nice explanation on www.howstuffworks.com

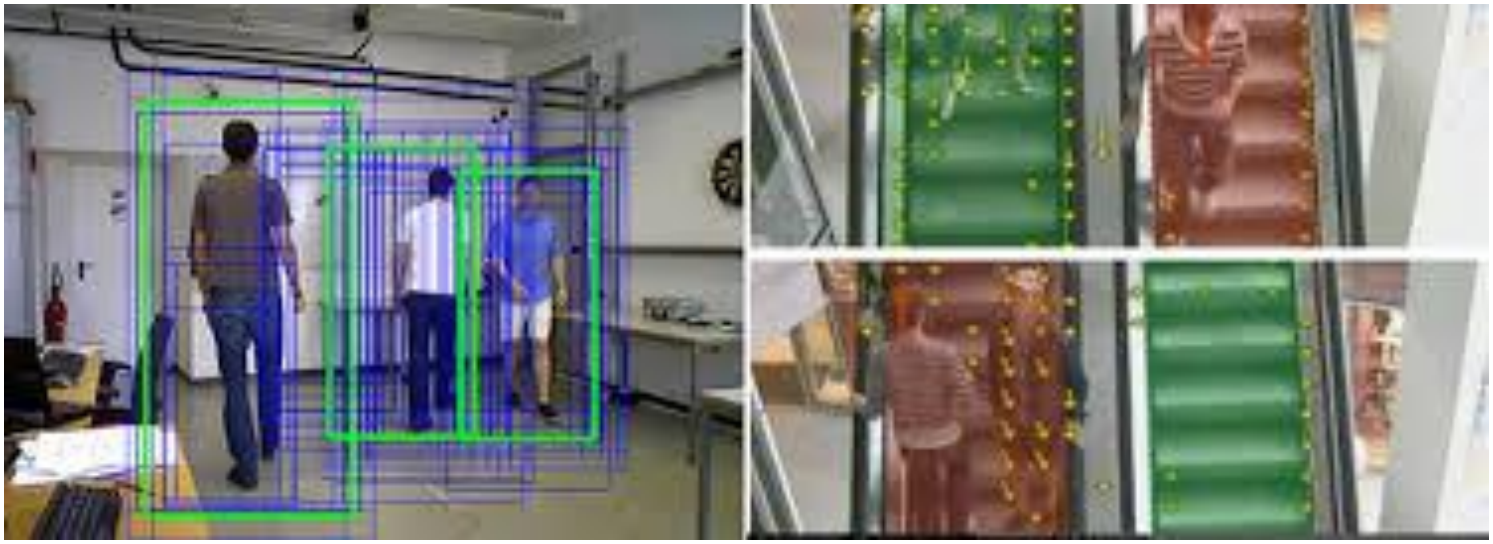
# Αυτόνομη οδήγηση

# Αναγνώριση χειρονομίας

# Ανάκτηση περιεχομένου

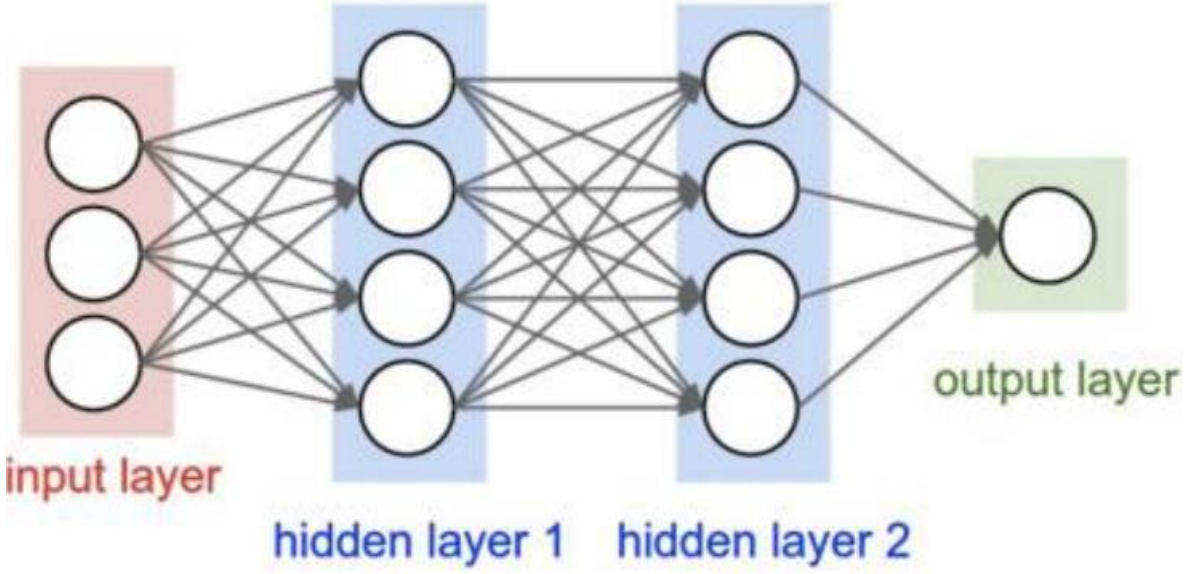



# Αναγνώριση συμπεριφοράς



# Βελτίωση - αποκατάσταση

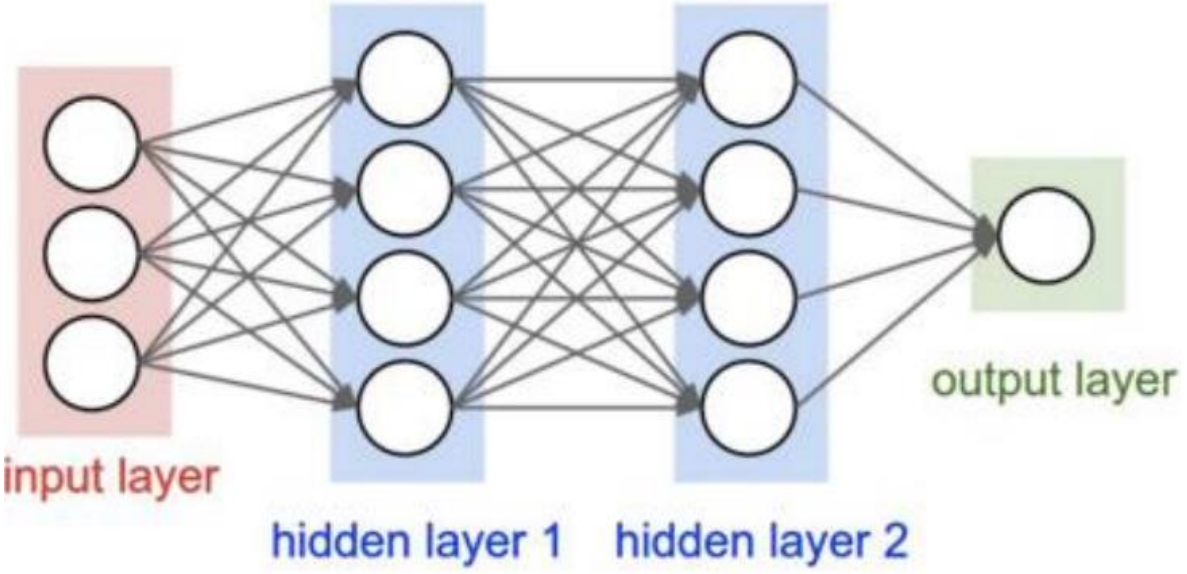
# Νευρωνικά δίκτυα και εικόνες




Ένα από τα βασικά προβλήματα είναι η αναγνώριση αντικειμένων σε εικόνες.

Πώς μπορεί να υλοποιηθεί;

# Νευρωνικά δίκτυα και εικόνες




Η είσοδος είναι η εικόνα (εικονοστοιχεία).

Αυτό δημιουργεί προβλήματα ως προς την διάσταση της εισόδου και ως προς την δυνατότητα αναπαράστασης.

# Πλήρης vs τοπική σύνδεση

Οι φυσικές εικόνες/βίντεο έχουν την ιδιότητα να είναι «στατικές», που σημαίνει ότι τα στατιστικά ενός μέρους της εικόνας είναι ίδια με οποιοδήποτε άλλο μέρος.

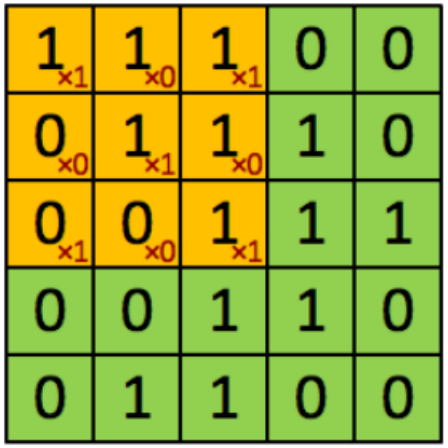
- Αυτό υποδηλώνει ότι τα χαρακτηριστικά που μαθαίνουμε σε ένα μέρος της εικόνας μπορούν επίσης να εφαρμοστούν και σε άλλα μέρη της εικόνας, και μπορούμε να χρησιμοποιήσουμε τα ίδια χαρακτηριστικά σε όλες τις τοποθεσίες.
- Αυτό οδηγεί στις εξής παραδοχές:
  - Τοπικά πεδία υποδοχής (local reception fields)
  - Δέσμευση βαρών (weight tying)

# Συνέλιξη και pooling

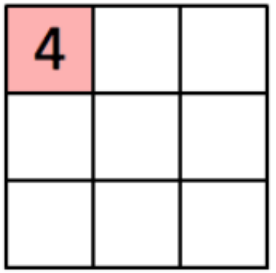
- Αρχικά εφαρμόζουμε συνέλιξη
- Αν η εικόνα είναι r x c και το μητρώο συνέλιξης είναι a x b τότε η εικόνα που προκύπτει είναι (r-a+1)x(c-b+1)

| | | |
|---|---|---|
| 1 | 0 | 1 |
| 0 | 1 | 0 |
| 1 | 0 | 1 |

**Μητρώο συνέλιξης**



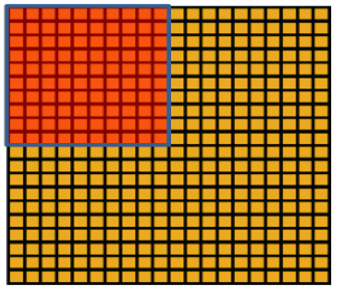

**εικόνα**



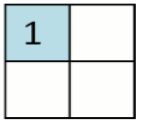

**χαρακτηριστικό από συνέλιξη**

# Συνέλιξη και συγκέντρωση

- Για την περιγραφή μιας μεγάλης εικόνας/βίντεο, μια φυσική προσέγγιση είναι η συγκέντρωση των τοπικών στατιστικών αυτών των χαρακτηριστικών σε διάφορες τοποθεσίες.
- Για παράδειγμα, θα μπορούσε κανείς να υπολογίσει τη μέση (ή μέγιστη) τιμή ενός συγκεκριμένου χαρακτηριστικού σε μια περιοχή της εικόνας.
- Αυτά τα συνοπτικά στατιστικά έχουν πολύ χαμηλότερη διάσταση (σε σύγκριση με τη χρήση όλων των εξαγόμενων χαρακτηριστικών) και μπορούν επίσης να βελτιώσουν τα αποτελέσματα (λιγότερη υπερπροσαρμογή).
- Αυτή η λειτουργία ονομάζεται μερικές φορές «μέση συγκέντρωση» (mean pooling) ή «μέγιστη συγκέντρωση» (max pooling), ανάλογα με τη λειτουργία συγκέντρωσης που εφαρμόζεται.
- Οι μονάδες συγκέντρωσης είναι αναλλοίωτες ως προς τη μετατόπιση.



Χαρακτηριστικά από συνέλιξη



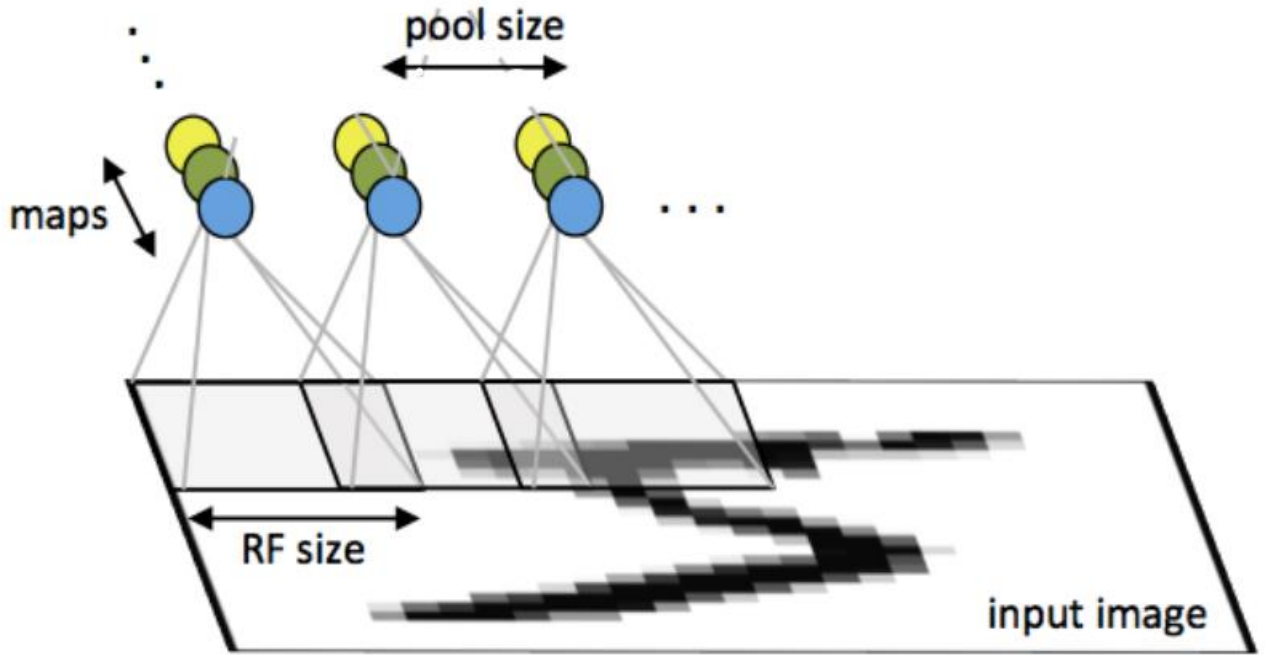

Χαρακτηριστικό από συγκέντρωση

# Συνελικτικό νευρωνικό δίκτυο (CNN)

- Ένα CNN αποτελείται από έναν αριθμό συνελικτικών (convolutional) και υποδειγματοληπτικών (subsampling) επιπέδων, τα οποία ακολουθούνται προαιρετικά από πλήρως συνδεδεμένα επίπεδα.
- Η είσοδος σε ένα συνελικτικό επίπεδο είναι μια εικόνα διαστάσεων m x m x r, όπου m είναι το ύψος και το πλάτος της εικόνας και r είναι ο αριθμός των καναλιών, π.χ. μια εικόνα RGB έχει r=3.
- Το συνελικτικό επίπεδο θα έχει k φίλτρα (ή πυρήνες) διαστάσεων n x n x q, όπου το n είναι μικρότερο από τη διάσταση της εικόνας και το q μπορεί είτε να είναι ίδιο με τον αριθμό των καναλιών r είτε μικρότερο και να διαφέρει για κάθε πυρήνα.
- Το μέγεθος των φίλτρων δημιουργεί τη δομή τοπικής σύνδεσης, όπου κάθε φίλτρο διέρχεται συνελικτικά από την εικόνα, παράγοντας k χάρτες χαρακτηριστικών διαστάσεων m−n+1.
- Κάθε χάρτης στη συνέχεια υποδειγματοληπτείται, συνήθως με μέση (mean) ή μέγιστη (max) συγκέντρωση (pooling) σε γειτονικές περιοχές διαστάσεων p x p, όπου το p κυμαίνεται από 2 για μικρές εικόνες (π.χ. MNIST) και συνήθως δεν ξεπερνά το 5 για μεγαλύτερες εισόδους.
- Είτε πριν είτε μετά το επίπεδο υποδειγματοληψίας, προστίθεται ένα προσαυξητικό bias και εφαρμόζεται μια σιγμοειδής μη γραμμικότητα (sigmoidal nonlinearity) σε κάθε χάρτη χαρακτηριστικών. Το παρακάτω σχήμα απεικονίζει ένα πλήρες επίπεδο σε ένα CNN που αποτελείται από συνελικτικά και υποδειγματοληπτικά υποεπίπεδα.

# Συνελικτικό νευρωνικό δίκτυο (CNN)

Μονάδες ίδιου χρώματος έχουν τα ίδια βάρη



# Συνελικτικό νευρωνικό δίκτυο (CNN)

- Επειδή υπάρχουν διακυμάνσεις φωτεινότητας, ακολουθείται μια διαδικασία λεύκανσης (whitening).

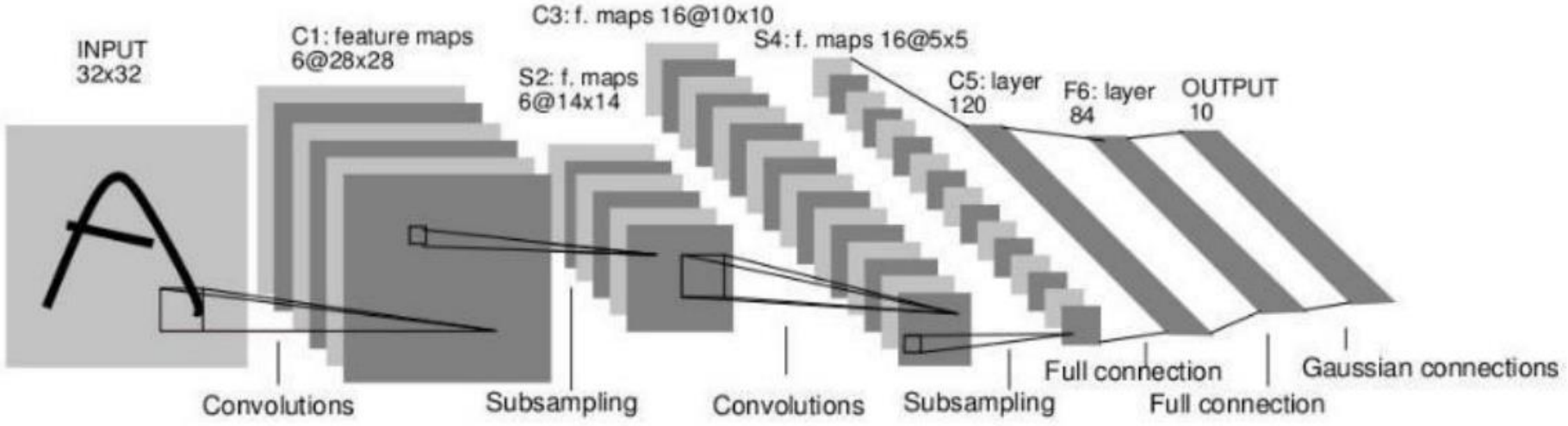
- Η διαδικασία αυτή είναι σύμφωνη με όσα συμβαίνουν στον οπτικ'ο φλοιό του εγκεφάλου

```
avg = mean(x, 1);     % Compute the mean pixel intensity value separately for each patch.
x = x - repmat(avg, size(x, 1), 1);
```

# Βαθιά Συνελικτικά Νευρωνικά Δίκτυα

Νευρωνικό δίκτυο με εξειδικευμένη δομή συνδεσιμότητας

- Στοίβαξη πολλαπλών σταδίων εξαγωγέων χαρακτηριστικών
- Τα υψηλότερα στάδια υπολογίζουν πιο γενικευμένα και αναλλοίωτα χαρακτηριστικά
- Στρώμα ταξινόμησης στο τέλος




Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86(11): 2278–2324, 1998.

# Συνάρτηση softmax

Υποθέστε την ύπαρξη K κλάσεων (π.χ., χειρονομίες).
• Δεδομένου ενός δοκιμαστικού δείγματος **x**, θέλουμε η υπόθεσή μας να εκτιμήσει την πιθανότητα p(y=k|x) για κάθε τιμή k=1,…,K  k=1,…,K.

• Δηλαδή, θέλουμε να εκτιμήσουμε την πιθανότητα η ετικέτα της κλάσης να παίρνει καθεμία από τις K διαφορετικές πιθανές τιμές. Έτσι, η υπόθεσή μας θα εξάγει ένα διάνυσμα διαστάσεων K (του οποίου τα στοιχεία αθροίζονται σε 1) που μας δίνει τις εκτιμώμενες πιθανότητες για τις K κλάσεις.
Συγκεκριμένα, η υπόθεσή μας $h_\theta(x)$ παίρνει τη μορφή:

$$h_\theta(x) = \begin{bmatrix} P(y=1|x;\theta) \\ P(y=2|x;\theta) \\ \vdots \\ P(y=K|x;\theta) \end{bmatrix} = \frac{1}{\sum_{j=1}^{K} \exp(\theta^{(j)\top} x)} \begin{bmatrix} \exp(\theta^{(1)\top} x) \\ \exp(\theta^{(2)\top} x) \\ \vdots \\ \exp(\theta^{(K)\top} x) \end{bmatrix}$$

$$\theta = \begin{bmatrix} | & | & | & | \\ \theta^{(1)} & \theta^{(2)} & \cdots & \theta^{(K)} \\ | & | & | & | \end{bmatrix}.$$

# Συνάρτηση softmax

Στην παρακάτω εξίσωση, το **1**{.} είναι η συνάρτηση δείκτη (indicator function), έτσι ώστε:
**1**{αληθής δήλωση} = 1
**1**{ψευδής δήλωση} = 0

- Η συνάρτηση κόστους μας θα είναι:

$$J(\theta) = -\left[\sum_{i=1}^{m}\sum_{k=1}^{K} 1\left\{y^{(i)} = k\right\} \log \frac{\exp(\theta^{(k)\top} x^{(i)})}{\sum_{j=1}^{K} \exp(\theta^{(j)\top} x^{(i)})}\right]$$

Μπορεί να λυθεί επαναληπτικά λαμβανοντας υπόψη ότι:

$$\nabla_{\theta^{(k)}} J(\theta) = -\sum_{i=1}^{m}\left[x^{(i)}\left(1\{y^{(i)} = k\} - P(y^{(i)} = k|x^{(i)};\theta)\right)\right]$$

Συνήθως προσθέτουμε ένα **softmax επίπεδο** στο τέλος μιας βαθιάς αρχιτεκτονικής.

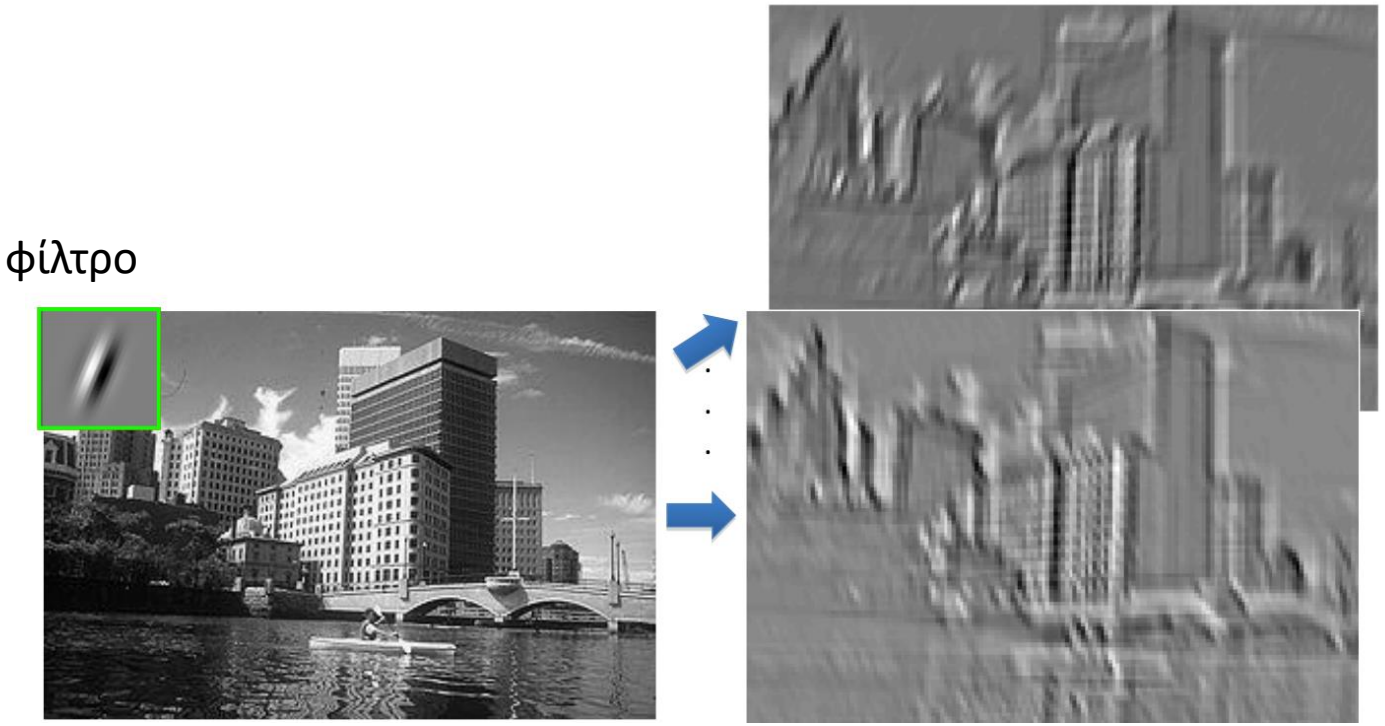
# Συνάρτηση softmax

Η αφαίρεση του $\psi$ από κάθε $\theta(j)$ δεν επηρεάζει καθόλου τις προβλέψεις της υπόθεσής μας!

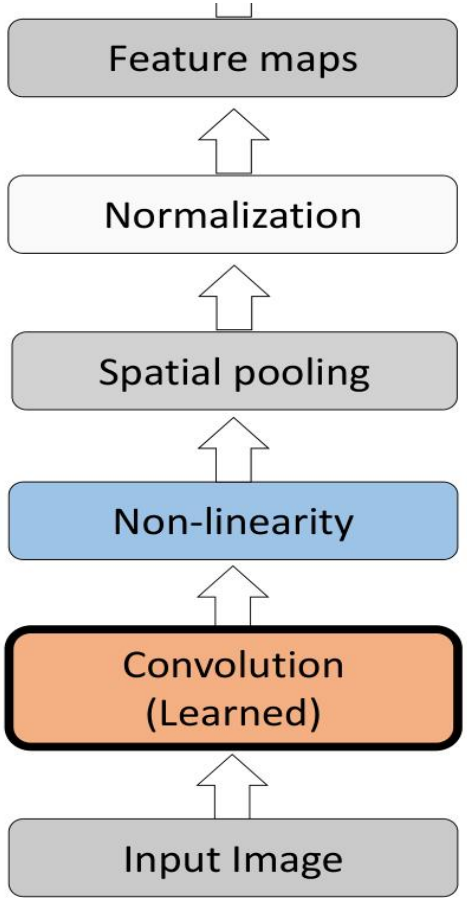
$$
\begin{aligned}
P(y^{(i)} = k | x^{(i)}; \theta) &= \frac{\exp((\theta^{(k)} - \psi)^\top x^{(i)})}{\sum_{j=1}^{K} \exp((\theta^{(j)} - \psi)^\top x^{(i)})} \\
&= \frac{\exp(\theta^{(k)\top} x^{(i)}) \exp(-\psi^\top x^{(i)})}{\sum_{j=1}^{K} \exp(\theta^{(j)\top} x^{(i)}) \exp(-\psi^\top x^{(i)})} \\
&= \frac{\exp(\theta^{(k)\top} x^{(i)})}{\sum_{j=1}^{K} \exp(\theta^{(j)\top} x^{(i)})}.
\end{aligned}
$$

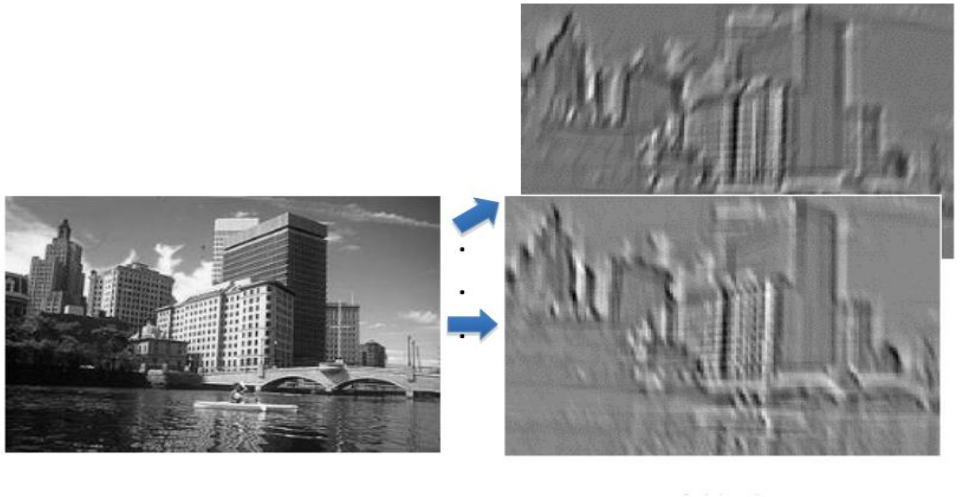
Αντί να βελτιστοποιούμε πάνω σε K·n παραμέτρους ($\theta^{(1)},\theta^{(2)},\ldots,\theta^{(K)}$), όπου $\theta(k)\in R^n$, μπορούμε να ορίσουμε $\theta^{(K)}=0$ να βελτιστοποιούμε μόνο ως προς τις (K−1)·n υπόλοιπες παραμέτρους.

# Βαθιά Συνελικτικά Νευρωνικά Δίκτυα

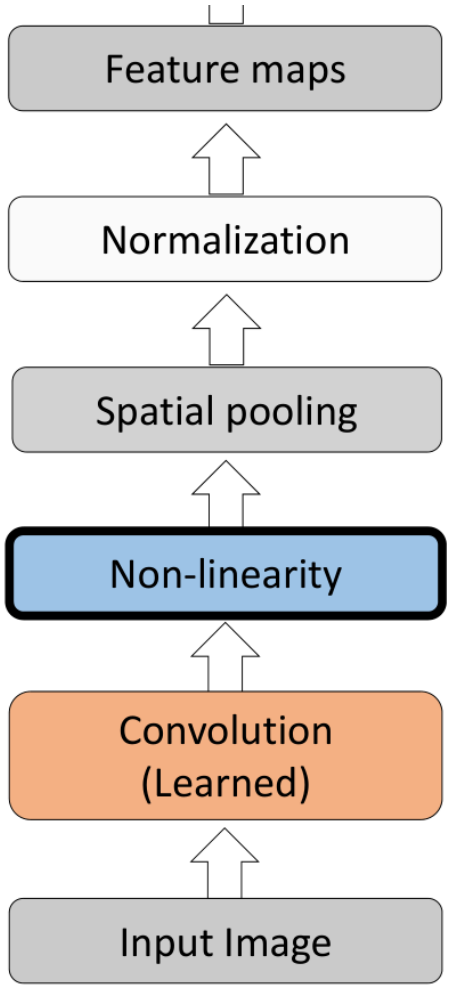


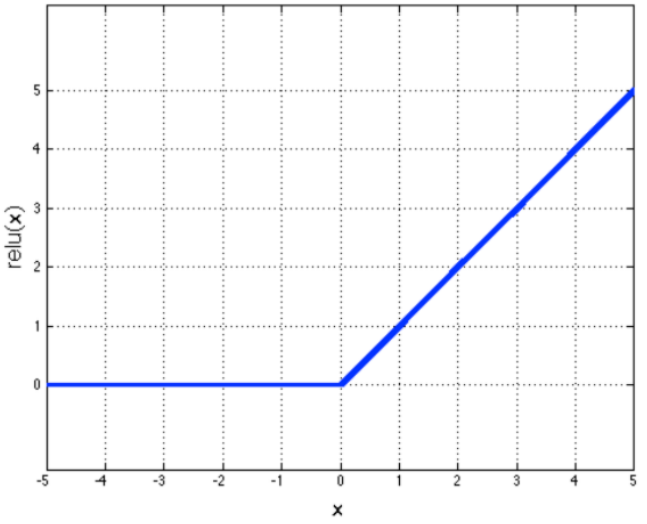
# Βαθιά Συνελικτικά Νευρωνικά Δίκτυα



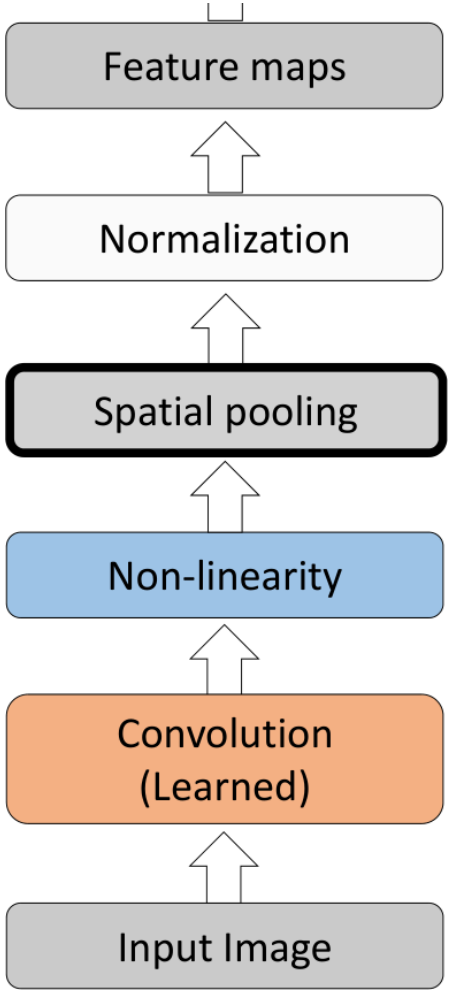



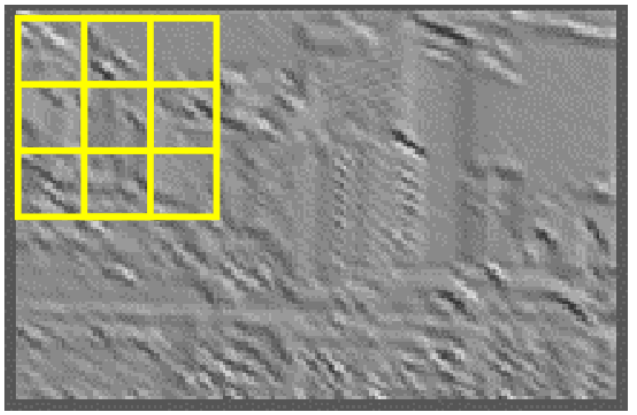
# Βαθιά Συνελικτικά Νευρωνικά Δίκτυα



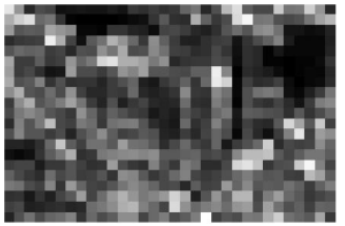



# Βαθιά Συνελικτικά Νευρωνικά Δίκτυα

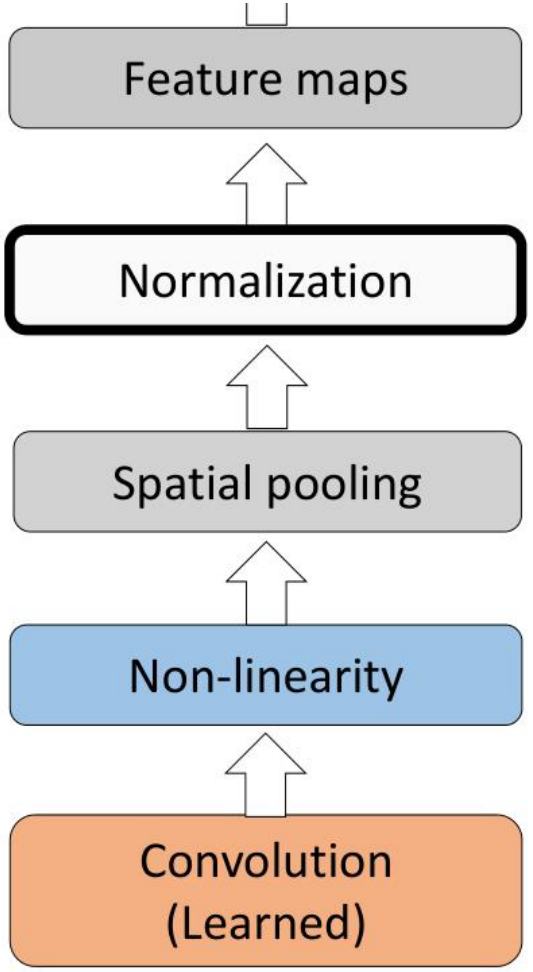


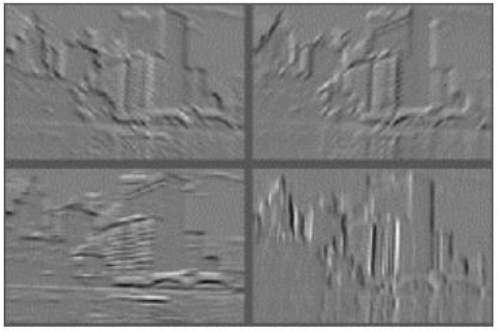



Max

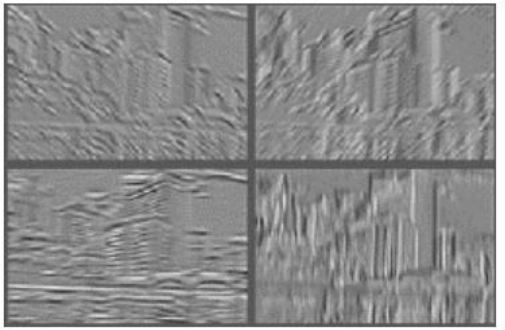


# Βαθιά Συνελικτικά Νευρωνικά Δίκτυα



Feature Maps

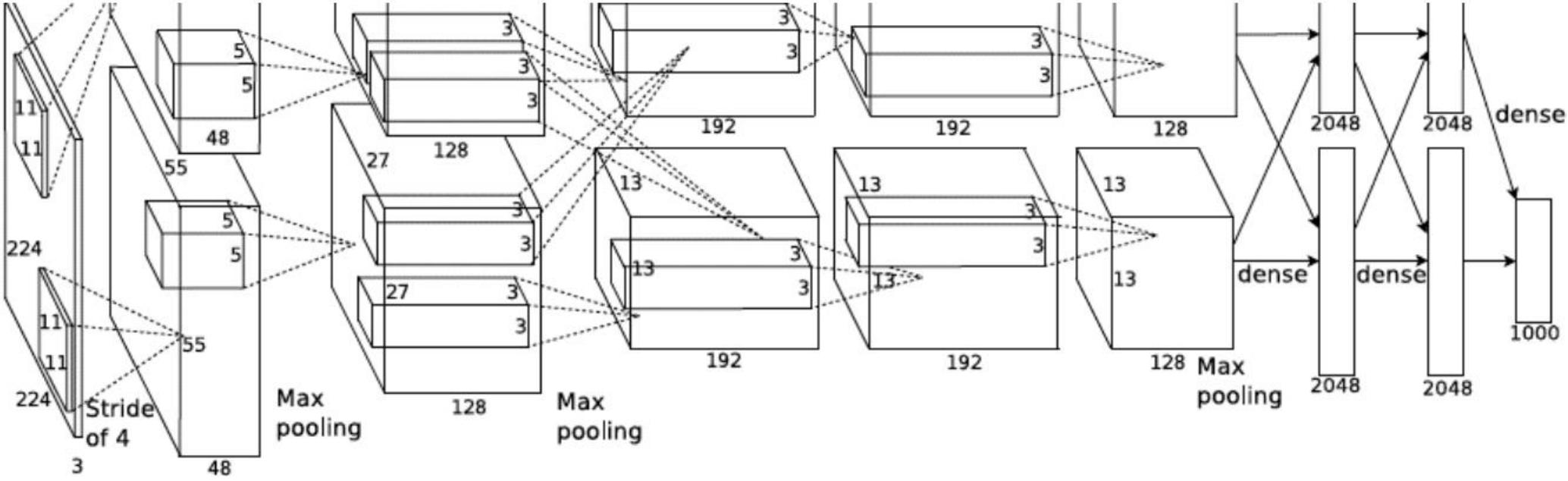


Feature Maps
After Contrast
Normalization

# Βαθιά Συνελικτικά Νευρωνικά Δίκτυα

Παρόμοιο πλαίσιο με το LeCun '98 αλλά:

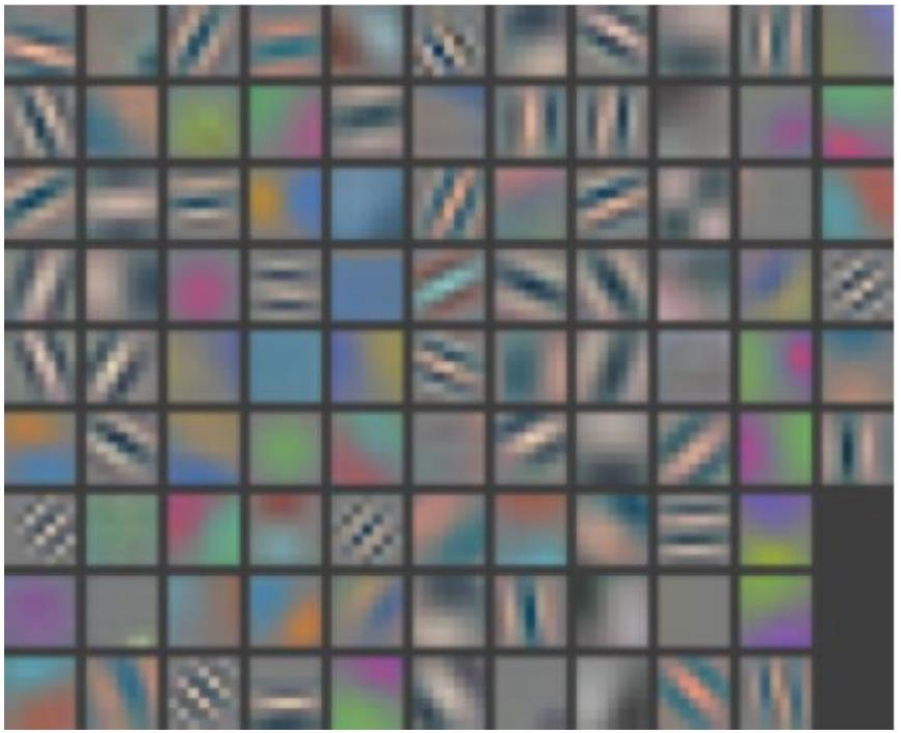
- Μεγαλύτερο μοντέλο (7 κρυφά επίπεδα, 650.000 μονάδες, 60.000.000 παράμετροι)
- Περισσότερα δεδομένα ($10^6$ έναντι $10^3$ εικόνων)
- Υλοποίηση σε GPU (50 φορές ταχύτερη από την CPU)
- Εκπαίδευση σε δύο GPU για μία εβδομάδα




A. Krizhevsky, I. Sutskever, and G. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, NIPS 2012

# Οπτικοποίηση CNN

Επίπεδο 1



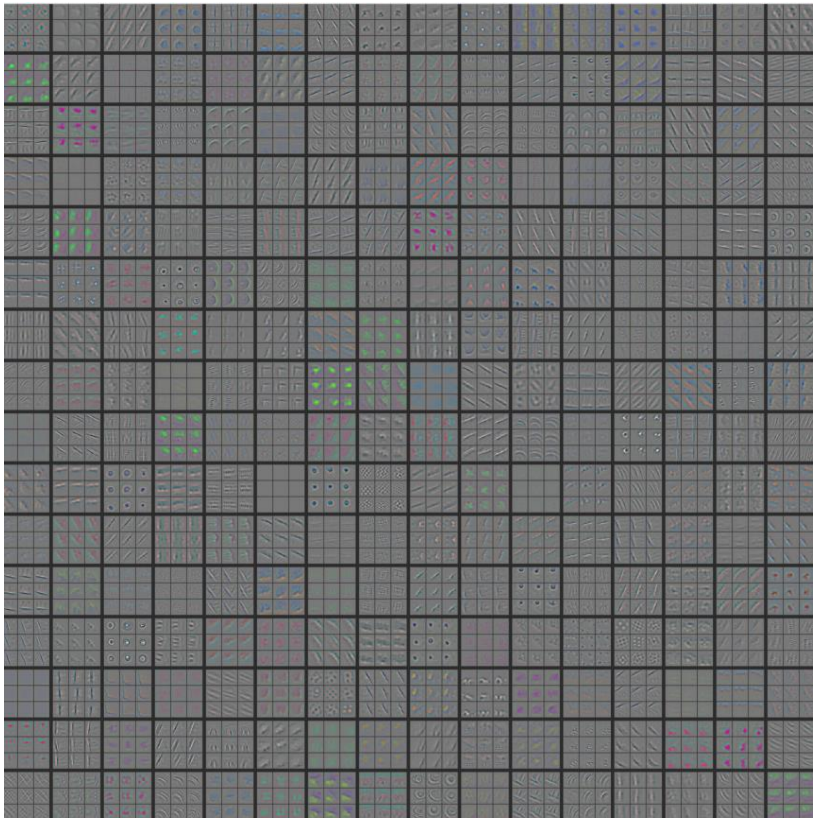
φίλτρα

9 κορυφαίες αποκρίσεις ανά φίλτρο

M. Zeiler and R. Fergus, Visualizing and Understanding Convolutional Networks, arXiv preprint, 2013

# Οπτικοποίηση CNN

Επίπεδο 2



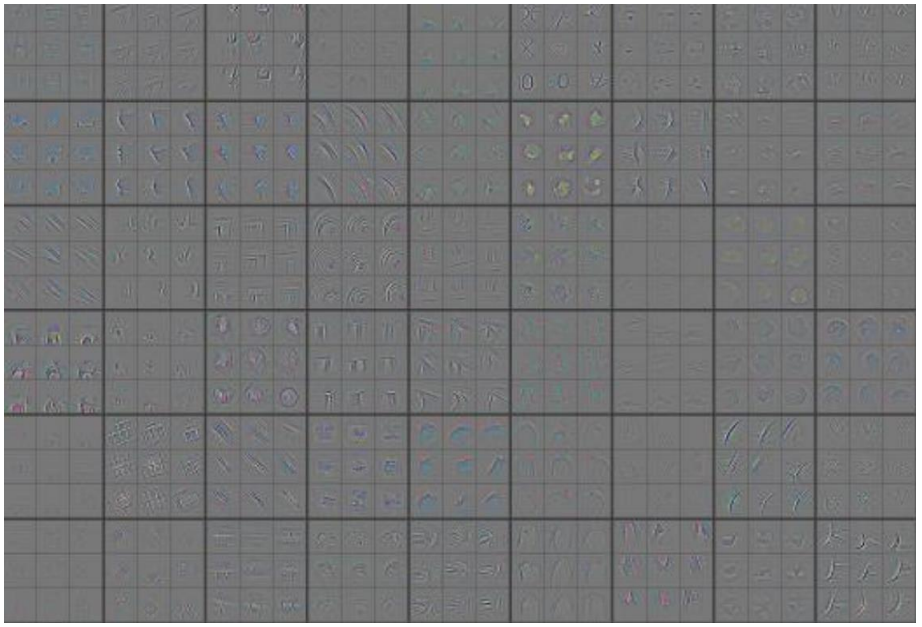
φίλτρα

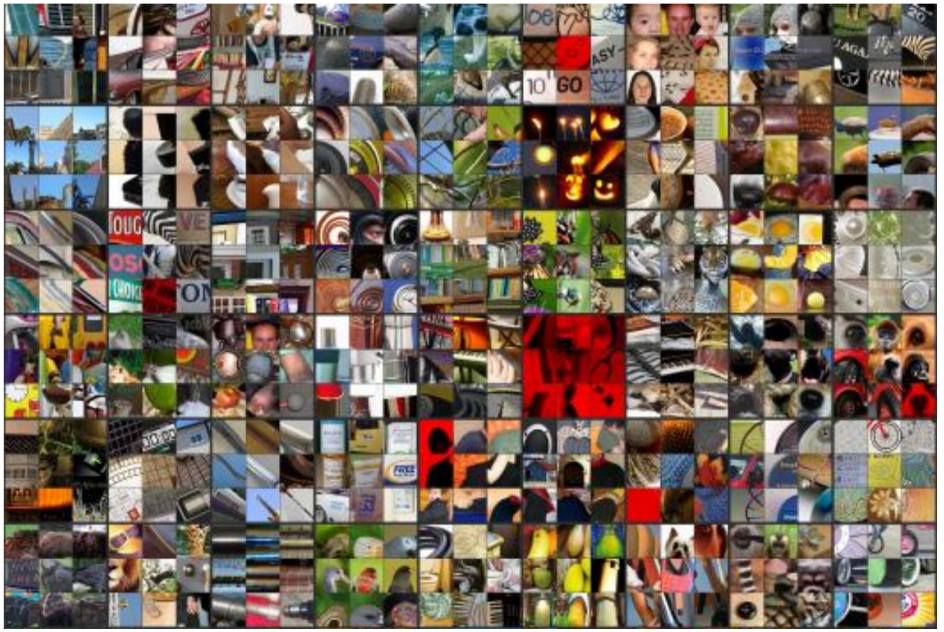
κορυφαίες αποκρίσεις ανά φίλτρο

M. Zeiler and R. Fergus, Visualizing and Understanding Convolutional Networks, arXiv preprint, 2013

# Οπτικοποίηση CNN

Επίπεδο 3





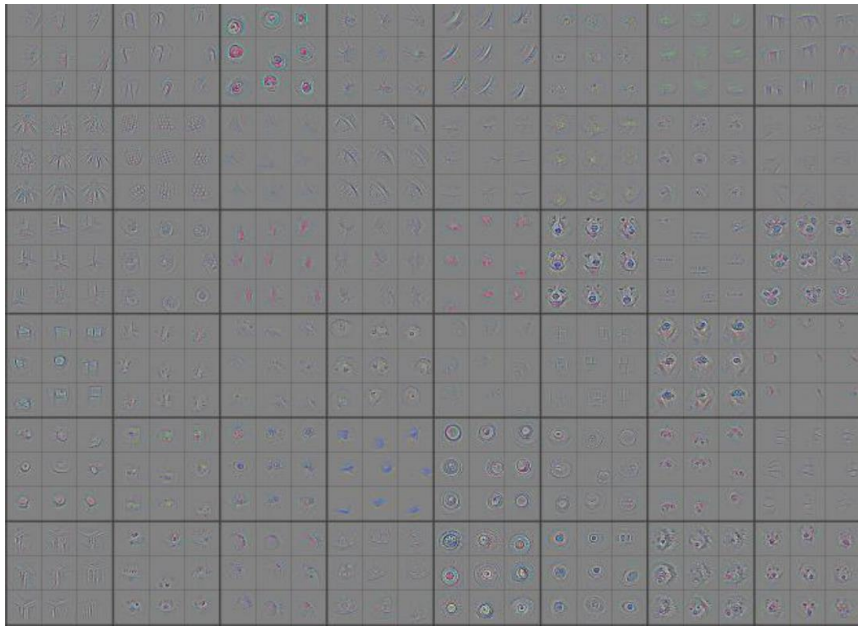
φίλτρα

κορυφαίες αποκρίσεις ανά φίλτρο

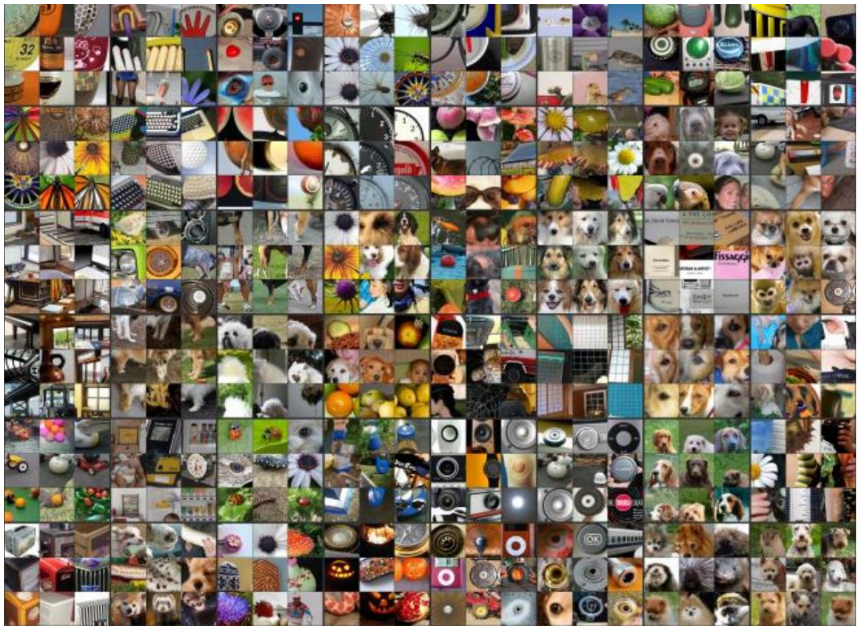
M. Zeiler and R. Fergus, Visualizing and Understanding Convolutional Networks, arXiv preprint, 2013

# Οπτικοποίηση CNN

Επίπεδο 4



φίλτρα

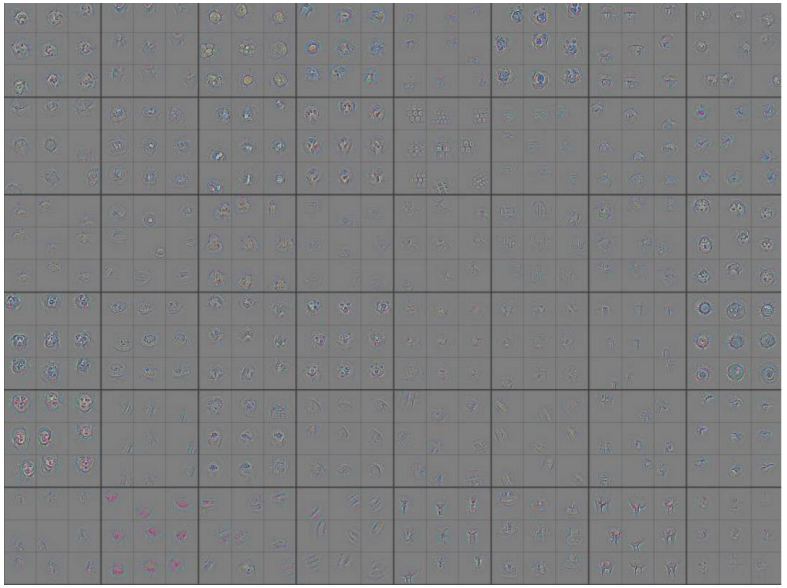


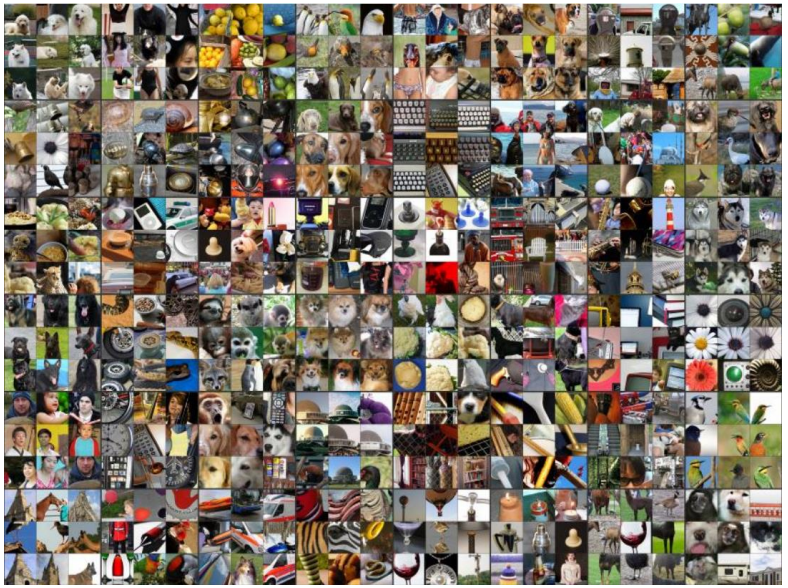
κορυφαίες αποκρίσεις ανά φίλτρο

M. Zeiler and R. Fergus, Visualizing and Understanding Convolutional Networks, arXiv preprint, 2013

# Οπτικοποίηση CNN

Επίπεδο 4





φίλτρα

κορυφαίες αποκρίσεις ανά φίλτρο

M. Zeiler and R. Fergus, Visualizing and Understanding Convolutional Networks, arXiv preprint, 2013

# Μοντέλα Όρασης-Γλώσσας

- Τα Μοντέλα Όρασης-Γλώσσας (VLMs) είναι μια αναδυόμενη κατηγορία μοντέλων Τεχνητής Νοημοσύνης σχεδιασμένα να κατανοούν και να παράγουν γλώσσα με βάση οπτικά δεδομένα.
- Συνδυάζουν την επεξεργασία φυσικής γλώσσας (NLP) με την υπολογιστική όραση για τη δημιουργία συστημάτων που μπορούν να αναλύσουν εικόνες και να παράγουν περιγραφές κειμένου.
- Μπορούν να απαντούν σε ερωτήσεις σχετικά με εικόνες ή να συμμετέχουν σε σύνθετους οπτικούς συλλογισμούς.
- Αυτή η συνέργεια επιτρέπει εφαρμογές σε διάφορους τομείς, όπως η προσβασιμότητα, η δημιουργία περιεχομένου και η εκπαίδευση.

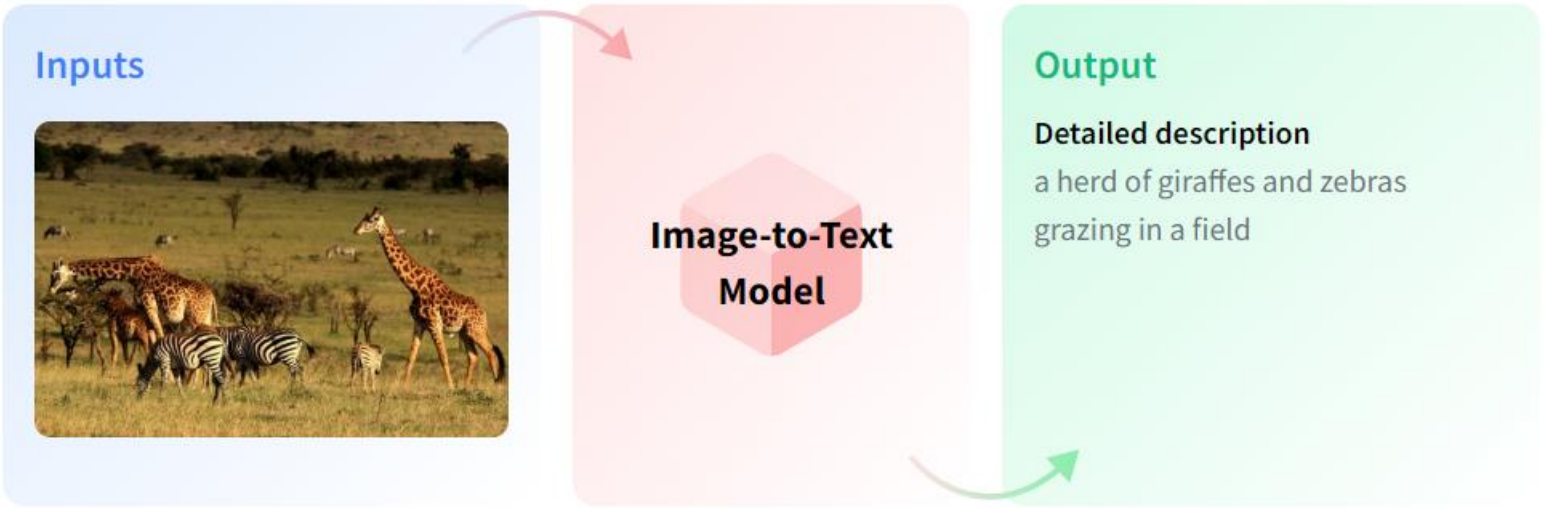
# Μοντέλα Όρασης-σε-Κείμενο

- Τα μοντέλα όρασης-σε-κείμενο επικεντρώνονται στη δημιουργία περιγραφών κειμένου ή στην απάντηση ερωτήσεων με βάση οπτικά δεδομένα.

Βασικά παραδείγματα:

- Υποτιτλισμός Εικόνων (Image Captioning): Το μοντέλο δημιουργεί περιγραφές φυσικής γλώσσας μιας εικόνας. Επεξεργάζεται τα οπτικά χαρακτηριστικά για να παράγει σχετικό κείμενο που περιγράφει τη σκηνή, τα αντικείμενα και τις σχέσεις τους μέσα στην εικόνα.
- Οπτική Απάντηση Ερωτήσεων (Visual Question Answering - VQA): Τα μοντέλα αυτά λαμβάνουν ως είσοδο μια εικόνα και μια ερώτηση σχετικά με αυτήν και παρέχουν απάντηση σε μορφή κειμένου.

# Μοντέλα Όρασης-σε-Κείμενο




Πηγη : huggingface.co

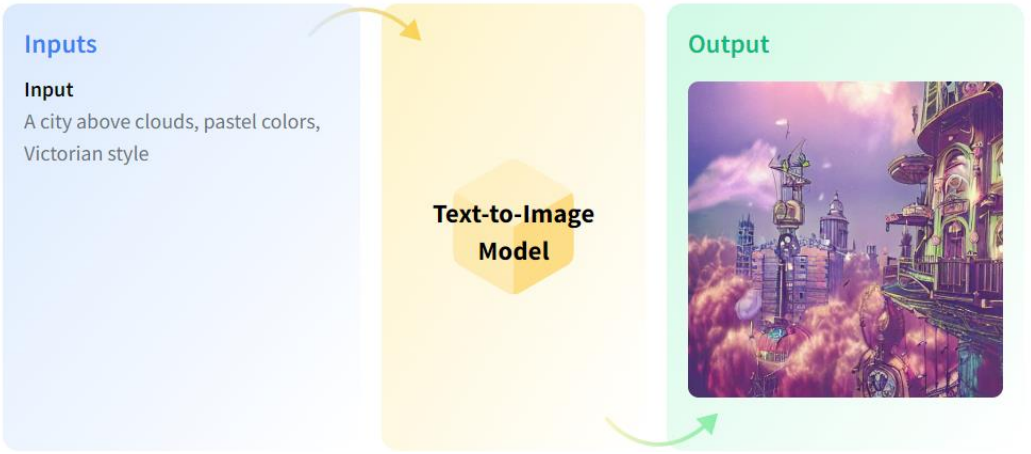
# Μοντέλα Κειμένου-σε-Όραση (Γεννήτριες Εικόνων από Κείμενο)

Τα μοντέλα κειμένου-σε-όραση λειτουργούν αντίστροφα, δημιουργώντας εικόνες από κειμενικές περιγραφές. Μεταφράζουν την είσοδο φυσικής γλώσσας σε οπτικές αναπαραστάσεις, οι οποίες μπορούν να χρησιμοποιηθούν σε διάφορες δημιουργικές και πρακτικές εφαρμογές.

Βασικές εφαρμογές:

- Γεννήτρια Εικόνων από Κείμενο (Text-to-Image Generation): Δεδομένης μιας κειμενικής περιγραφής (π.χ. «μια γάτα καθισμένη σε έναν κόκκινο καναπέ»), το μοντέλο δημιουργεί μια εικόνα που αντιστοιχεί στο κείμενο. Το DALL·E είναι ένα από τα πιο γνωστά μοντέλα αυτής της κατηγορίας.
- - Τροποποίηση Εικόνων με Βάση Κείμενο (Text-Driven Image Manipulation): Τα μοντέλα αυτά μπορούν να τροποποιήσουν υπάρχουσες εικόνες σύμφωνα με κειμενικές οδηγίες. Για παράδειγμα, μπορείτε να δώσετε εντολή στο μοντέλο να «αλλάξει το φόντο σε ηλιοβασίλεμα» για μια υπάρχουσα φωτογραφία.

# Μοντέλα Κειμένου-σε-Όραση (Γεννήτριες Εικόνων από Κείμενο)





Πηγη : huggingface.co

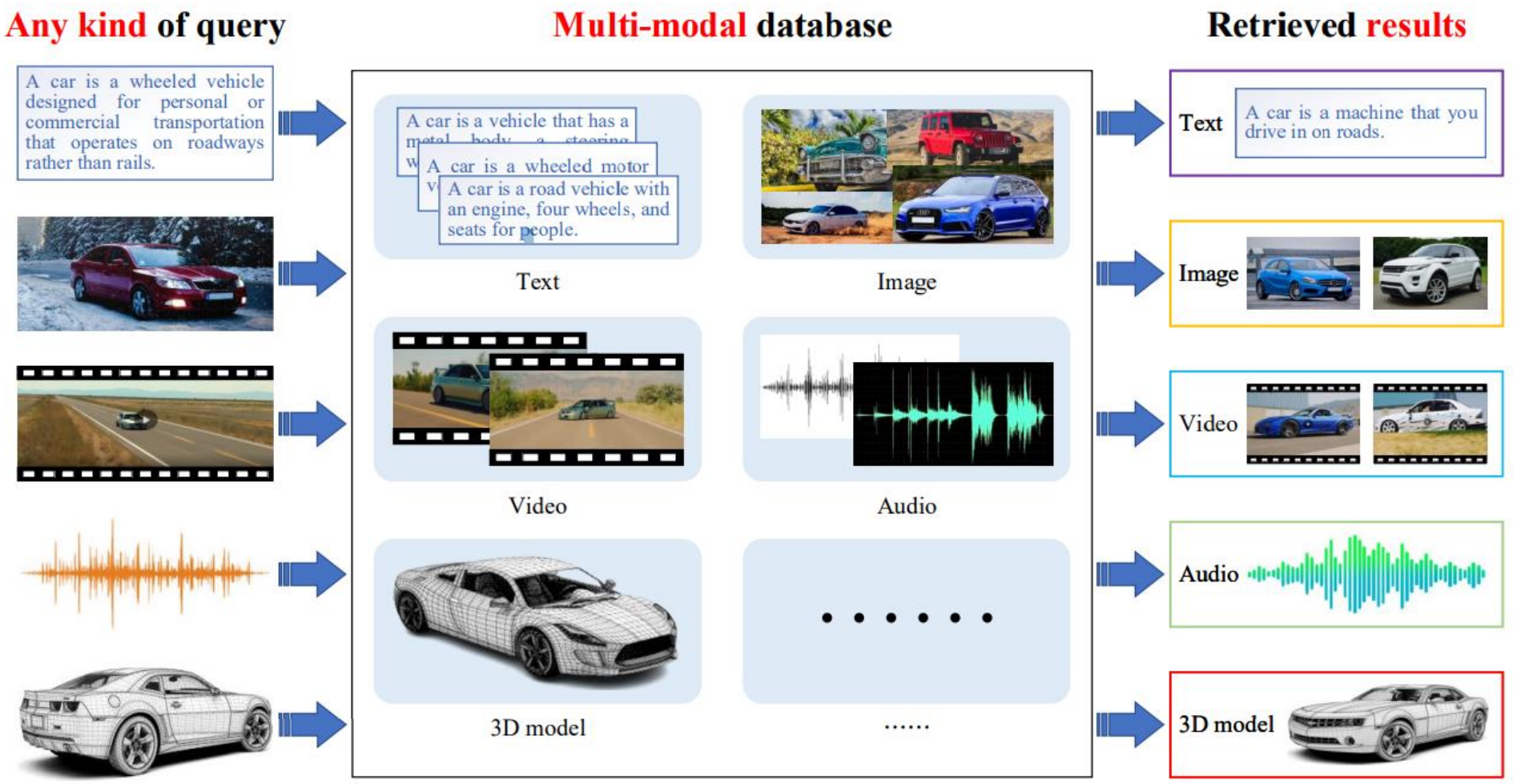
# Μοντέλα Διαμόρφωσης Διασταυρούμενων Τρόπων (Αναζήτηση Εικόνων με Κείμενο, Αναζήτηση Κειμένου με Εικόνες)

Τα μοντέλα διαμόρφωσης διασταυρούμενων τρόπων έχουν σχεδιαστεί για εργασίες όπου μία μορφή δεδομένων (π.χ. κείμενο ή εικόνα) χρησιμοποιείται για την αναζήτηση δεδομένων σε άλλη μορφή. Αυτά τα μοντέλα επιτρέπουν στους χρήστες να εκτελούν εργασίες όπως:

Βασικές εφαρμογές:

- Αναζήτηση Εικόνων με Κείμενο: Το μοντέλο ανακτά εικόνες που ταιριάζουν με μια δεδομένη κειμενική αναζήτηση. Αυτή η εφαρμογή χρησιμοποιείται συχνά σε μηχανές αναζήτησης, όπου οι χρήστες αναζητούν εικόνες πληκτρολογώντας λέξεις-κλειδιά ή φράσεις.
- Αναζήτηση Κειμένου με Εικόνες: Αντίστροφα, το μοντέλο μπορεί να λάβει μια εικόνα ως είσοδο και να ανακτήσει σχετικό κείμενο. Για παράδειγμα, οι χρήστες μπορούν να ανεβάσουν μια εικόνα ενός αντικειμένου και το μοντέλο να επιστρέψει περιγραφές, σχετικά άρθρα ή πληροφορίες προϊόντων.

# Μοντέλα Διαμόρφωσης Διασταυρούμενων Τρόπων (Αναζήτηση Εικόνων με Κείμενο, Αναζήτηση Κειμένου με Εικόνες)




Tianshi Wang, Fengling Li, Lei Zhu, Jingjing Li, Zheng Zhang, Heng Tao Shen, Cross-Modal Retrieval: A Systematic Review of Methods and Future Directions, 2023

# Λειτουργία μοντέλων όρασης – γλώσσας: Οπτική και Γλωσσική είσοδος

- Τα Μοντέλα Όρασης-Γλώσσας (VLMs) επεξεργάζονται δύο διαφορετικούς τύπους εισόδου: εικόνες και κείμενο. Αυτά τα μοντέλα μαθαίνουν να ευθυγραμμίζουν τα οπτικά χαρακτηριστικά (από εικόνες) με γλωσσικές αναπαραστάσεις (από κείμενο) ώστε να μπορούν να κατανοούν τόσο οπτικά όσο και γλωσσικά δεδομένα ταυτόχρονα.

Οπτική Είσοδος:
- Συνήθως, η εικόνα επεξεργάζεται μέσω ενός βαθιού συνελικτικού νευρωνικού δικτύου (CNN), όπως το ResNet, Vision Transformers (ViT) ή παρόμοιες αρχιτεκτονικές.
- Αυτά τα μοντέλα μετατρέπουν την εικόνα σε μια πυκνή αναπαράσταση χαρακτηριστικών που αποτυπώνει τα βασικά της στοιχεία, όπως σχήματα, αντικείμενα και υφές.

Γλωσσική Είσοδος:
- Το κείμενο επεξεργάζεται μέσω γλωσσικών μοντέλων, όπως τα Transformers (π.χ. BERT, GPT).
- Το κείμενο χωρίζεται σε tokens και κάθε token μετατρέπεται σε ενσωμάτωση (embedding), η οποία αποτυπώνει τη σημασιολογική έννοια των λέξεων.

# Λειτουργία μοντέλων όρασης – γλώσσας: Εξαγωγή Χαρακτηριστικών και Αναπαράσταση

Τόσο οι οπτικές όσο και οι γλωσσικές είσοδοι μετατρέπονται σε έναν ενοποιημένο χώρο μέσω μιας διαδικασίας που ονομάζεται εξαγωγή χαρακτηριστικών.

Οπτικά Χαρακτηριστικά:
- Πρόκειται για διανύσματα υψηλών διαστάσεων που αναπαριστούν συγκεκριμένα στοιχεία της εικόνας (όπως αντικείμενα, φόντα ή υφές).

Γλωσσικά Χαρακτηριστικά:
- Αυτά τα διανύσματα αναπαριστούν τις έννοιες των λέξεων ή φράσεων στο πλαίσιο του κειμένου εισόδου.

# Λειτουργία μοντέλων όρασης – γλώσσας: Ευθυγράμμιση Διασταυρούμενων Τρόπων (Cross-Modal Alignment)

Για να κατανοήσει και τις δύο μορφές δεδομένων (οπτικά και γλωσσικά), το μοντέλο πραγματοποιεί ευθυγράμμιση διασταυρούμενων τρόπων. Σε αυτό το στάδιο, τα οπτικά και γλωσσικά χαρακτηριστικά αντιστοιχίζονται και ευθυγραμμίζονται σε έναν κοινό χώρο ενσωμάτωσης (embedding). Για παράδειγμα, η λέξη «γάτα» στο κείμενο ευθυγραμμίζεται με τα οπτικά χαρακτηριστικά που αναπαριστούν μια γάτα στην εικόνα.

**Τεχνικές Ευθυγράμμισης:**

- **Αντιθετική Μάθηση (Contrastive Learning):**
  - Εκπαιδεύει το μοντέλο να ελαχιστοποιεί την απόσταση μεταξύ σχετικών ζευγών κειμένου-εικόνας και να τη μεγιστοποιεί για μη σχετιζόμενα ζευγάρια.
- **Μηχανισμοί Προσοχής (Attention Mechanisms):**
  - Η προσοχή διασταυρούμενων τρόπων (cross-modal attention) χρησιμοποιείται συχνά για να εστιάζει επιλεκτικά σε σχετικά μέρη της εικόνας ή του κειμένου, δημιουργώντας αντιστοιχίες μεταξύ οπτικών στοιχείων και λέξεων.

.

# Λειτουργία μοντέλων όρασης – γλώσσας: Στρώματα Σύντηξης (Fusion Layers)

Αφού εξαχθούν και ευθυγραμμιστούν τα χαρακτηριστικά από τα οπτικά και γλωσσικά δεδομένα, συχνά συνδυάζονται σε έναν κοινό χώρο αναπαράστασης. Αυτή η σύντηξη επιτρέπει στο μοντέλο να λαμβάνει αποφάσεις βασισμένες και στις δύο μορφές δεδομένων ταυτόχρονα.

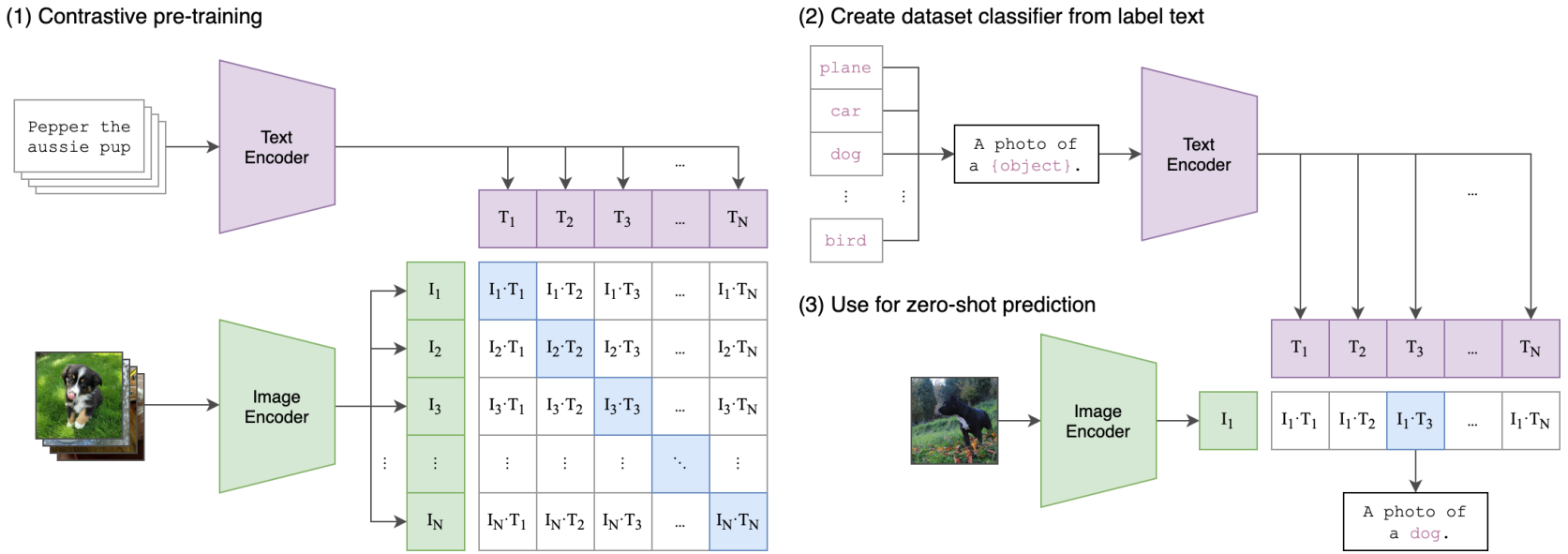
Στρατηγικές Σύντηξης:

- Καθυστερημένη Σύντηξη (Late Fusion):
  - Τα χαρακτηριστικά επεξεργάζονται ανεξάρτητα και οι τελικές αναπαραστάσεις τους συνδυάζονται στο τέλος της διαδικασίας.

- Πρώιμη Σύντηξη (Early Fusion):
  - Τα οπτικά και γλωσσικά χαρακτηριστικά συνδυάζονται νωρίς στο μοντέλο και επεξεργάζονται μαζί από τα αρχικά στάδια.

- Σύντηξη με Διασταυρούμενη Προσοχή (Cross-attention Fusion):
  - Τα χαρακτηριστικά κάθε τρόπου ενημερώνουν το άλλο κατά τη διάρκεια της επεξεργασίας, εστιάζοντας επιλεκτικά σε σχετικά χαρακτηριστικά και από τις δύο εισόδους.

# Το μοντέλο CLIP




Learning Transferable Visual Models From Natural Language Supervision
Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever, 2021

# Πηγές

1. Andrew NG, Introduction to computer vision
2. A. Efros, Introduction to computer vision
3. https://www.geeksforgeeks.org/computer-vision/