Im2Vec: Joint Semantic and Image Spaces

Data Sciences Presentation

January 30, 2015

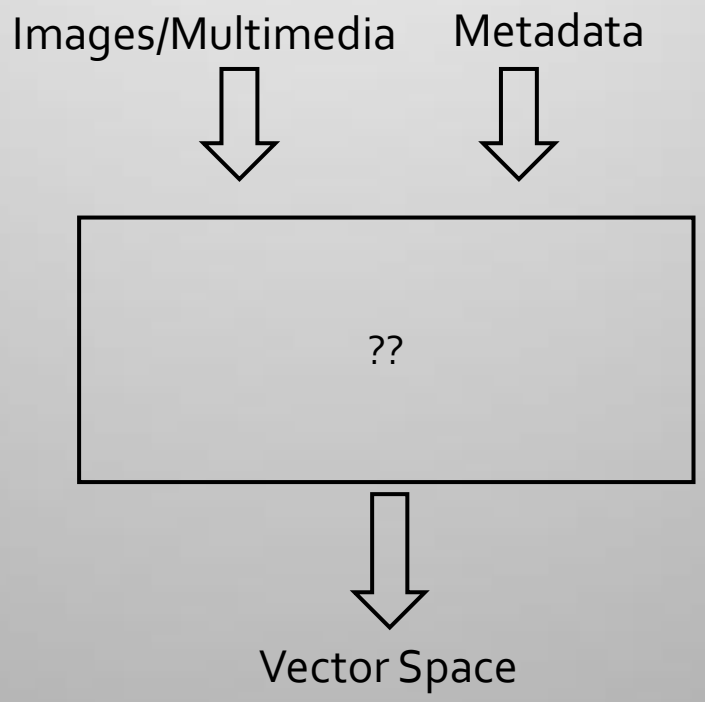
Lawrence Livermore National Laboratory

LLNL-PRES-657343

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. Lawrence Livermore National Security, LLC

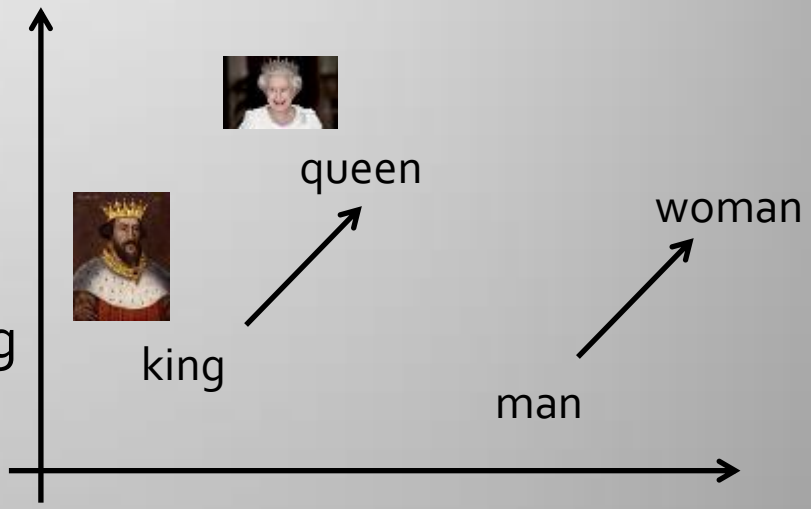
Learning from Unstructured Data

- Data: Large corpora of images & associated metadata that includes free text (open source)
- Problem: Given a multimedia corpus, can we create a joint vector space between multimodal elements?
  - Constraint 1: Unstructured text of any length, language, or relevance
  - Constraint 2: Minimal training guidance and minimal tuning
- Challenge: Very messy, lots of noise, heterogeneous, and sometimes irrelevant tags



Current Problem (ICCV 2015)

- Goal: Create a vector space to where multimodal elements can be mapped
- Advantages:
  - Similarities, distances, and differences between a diverse set of media make sense. For example:
    - Words to other Words: grammatical & contextual
    - Images to other images
    - Words to Images
    - Images to Words
  - Euclidean operations have meaning over diverse domain. For example:
    - Analogies: King is to queen as man is to woman
    - V("Woman") = V("King") –V("Queen") + V("Man")
    - V("Woman") ≈ V() – V() + V("Man")

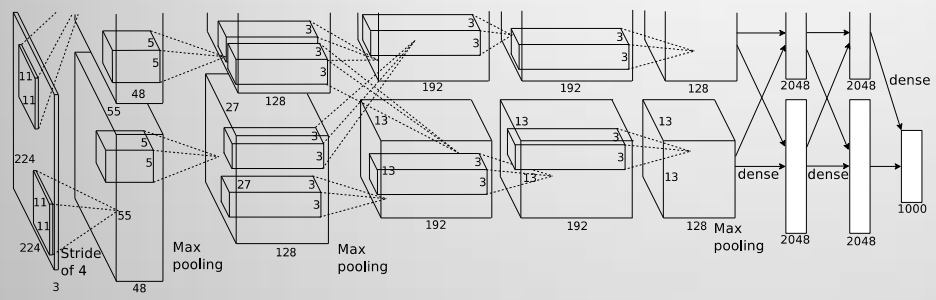


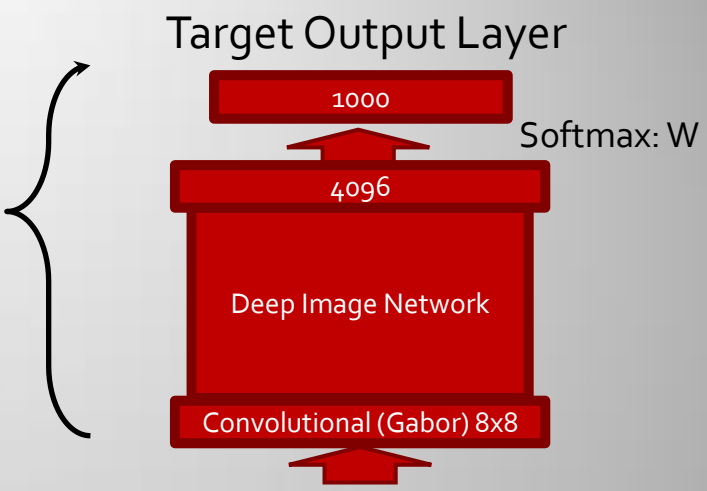
Roadmap of implementation

- Supervised Neural Network, Targeted Training [Krizhevsky et al, '12]
  - Supervised deep learning architecture
  - ILSVRC classification Task: 1000 "synset" classes, 150k images
  - YFCC 100M have no labels, but intermediate layers are useful
- Word2Vec multithreaded structure [Mikolov et al, NIPS, '13]
  - Distributed representation of semantics information
  - Not multi-modal, but it is extendable (codebase)
- CaffeNET [Y. Jia et al, '13]
  - Necessary for large-scale feature extraction and back propagation
  - Final layers need to be implemented
- Dual autoencoders with association labels [Vincent et al, '10, Feng '14]
  - Unsupervised cross-modal structure with autoencoders
  - Unprincipled and poor performance, especially at large scale.
  - Dimensionality is too large in both image and word space
  - However, it does offer a nonlinear solution for a highly complex solution

The AlexNET deep learning architecture

- ImageNET Competition (ILSVRC 2012) : 1000 Classes
- AlexNET Architecture:





- Final softmax layer, learning posteriors:

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{j=1}^{k} 1\left\{y^{(i)} = j\right\} \log \frac{e^{\theta_j^T x^{(i)}}}{\sum_{l=1}^{k} e^{\theta_l^T x^{(i)}}}\right] + \frac{\lambda}{2}\sum_{i=1}^{k}\sum_{j=0}^{n}\theta_{ij}^2$$

- Base structure is useful, but final layer is unsuitable for unstructured text at large scale
  - Manually intensive training
  - Inflexible targeted single labels
  - Concurrent definitions unsupported
  - Large dimensionality of vocabulary
  - Not "concept" driven

00024a73d1a4c32fb29732d56a2:
Red Noel christmas electric signs noel

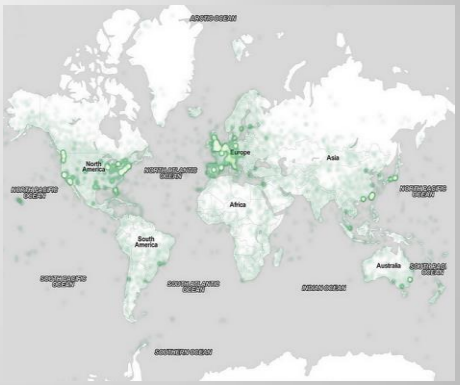
User description:
my black camaro

User description:
dat racing machine

Training with distributed representations

- Unstructured text is extremely noisy & varied

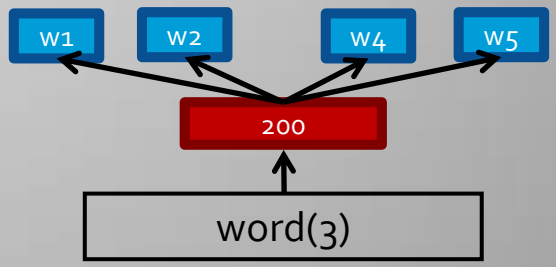
square	iphoneography	square format	instagram app	california	travel
nikon	usa	canon	london	japan	france
nature	art	music	europe	beach	united states
england	wedding	italy	new york	canada	city
vacation	germany	party	park	water	people
uk	spain	architecture	summer	festival	nyc
taiwan	paris	san francisco	australia	winter	sky
snow	concert	night	family	china	museum
food	street	live	washington	landscape	flower
sunset	photo	flowers	holiday	trip	photography



- Neural approaches are state of the art and perform surprisingly well [Baroni, '14]
- Mikolov et al: skip-gram distributed modeling

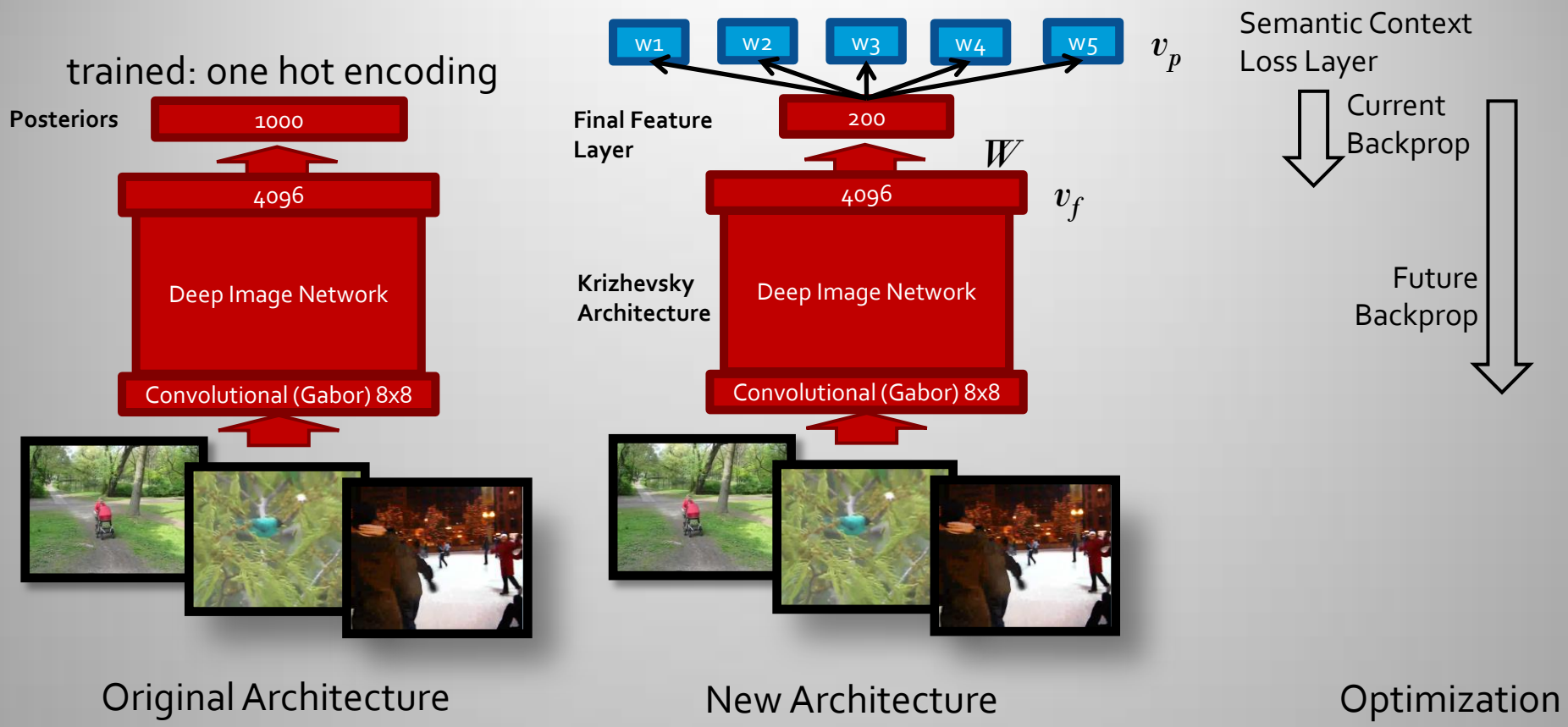
$$\mathcal{J}(v_w, v_o) = \log \sigma(v_o^T v_w) + \sum_n \mathbb{E}_{P_n}\left[\log \sigma(-v_w^T v_n)\right]$$

  - Semantic-based vector representation of words
  - Context-based: taking a window around a word



- Solving the multiple label problem:
  - Robust to noisy metadata associated with imagery
  - Extendable a large corpus of "clean data"
  - Relate images to concepts (context) rather than labels

Context-based "skip" network



- Trained network places images and semantics in the same vector space
- Improvements: tune the network to use the negative gradient to back propagate

Positive association with negative sampling regularization

- Define the following vectors:
  - v_w: word vectors ($v_w \in \mathbb{R}^{200}$), v_f: image feature ($v_f \in \mathbb{R}^{4096}$)
  - v_o: output vector ($v_o \in \mathbb{R}^{200}$), v_p: positive sample ($v_p \in \mathbb{R}^{200}$) , v_n: negative sample ($v_n \in \mathbb{R}^{200}$)
- Mikolov et al., noise contrast estimation

$$\mathcal{J}(v_w, v_o) = \log \sigma(v_o^T v_w) + \sum_n \mathbb{E}_{P_n} \left[\log \sigma(-v_w^T v_n) \right]$$

- Added term to deal with related images:

$$\mathcal{J}(W^{(f)}) = \sum_p \log \sigma(v_p^T W v_f) + \sum_n \mathbb{E}_n \left[\log \sigma(-v_n^T W v_f) \right]$$

- Gradient update:

$$\begin{aligned} \nabla_W J(W^{(f)} &= \nabla J_+ + \nabla J_- \\ &= \left\{ \sum_p v_p \left[1 - \sigma(v_p^T W v_f) \right] - \sum_n v_n^T \sigma(v_n^T W v_f) \right\} v_f^T \end{aligned}$$

- Weighting matrix is the final layer of network

Implementation details

$$\mathcal{J}(v_w, v_o, W) = \sum_{w,f,o} \left\{ \log \sigma(v_o^T v_w) + \sum_p \log \sigma(v_p^T W v_f) + \sum_n \mathbb{E}_{P_n} \left[\log \sigma(-v_n^T W v_f) \sigma(-v_n^T v_w) \right] \right\}$$

- Joint optimization over v_w, v_o, and W
  - Mikolov does this with SGD over
  - Substitute v_p with v_o for joint training
  - The vocabulary is roughly 15% larger than necessary (meaningless and infrequent words / emoticons)

- Pre-training and vocabulary pruning
  - Lots of noise and unicode characters
  - Clean datasets: NY Times (20 years), Wikipedia (1st 9 billion characters)
  - If $v_o \neq v_p$ (i.e., joint training is not necessary)
    - Optimize word space first, then optimize W matrix
    - Better if we use a "clean" dataset first, and then optimize over the images based on the context that it sees.

Yahoo! FlickR Creative Commons 100M Dataset

- YFCC100M Offers Opportunity to Learn Semantic Space for Images, Videos, and Text
- One of the Largest Publicly Available Multimedia Datasets
  - 99.3 million images, 0.7 million videos
  - Metadata includes: description, camera type, gps location, tags, user
- Collaboration with ICSI Berkeley, Yahoo!, Amazon, and LLNL
- LLNL's Video Analytics LDRD provided speech and video features for the geo-location task in MediaEval2014, and ACM competition at ACM 2015

ICSI Works With Yahoo Labs and Lawrence Livermore Lab to Offer Analytics Tools for Over 100 Million Flickr Images and Videos

50TB Computing Program Runs Analysis on the Entire Flickr Creative Commons Dataset, One of the Largest Public Multimedia Datasets Ever Released to the Public

MARKET WIRED ICSI July 3, 2014 9:00 AM

BERKELEY, CA--(Marketwired - Jul 3, 2014) - The International Computer Science Institute (ICSI), a leading center for computer science research, today announced a collaboration with Yahoo Labs and Lawrence Livermore National Laboratory to process and analyze the recently released Yahoo Flickr Creative Commons 100 Million (YFCC100M) dataset, a publicly available corpus of user-generated content comprising more than 100 million images and videos.

Query Results on Text Alone

- Query “Red”, Metadata+NYTimes, Metadata Only

Word	Cosine distance
yellow	0.765134
pink	0.667090
purple	0.660480
orange	0.655323
white	0.647688
blue	0.646990
green	0.633826
stripes	0.549109
striped	0.539543
black	0.527789

Word	Cosine distance
yellow	0.766820
purple	0.690916
pink	0.690356
orange	0.659482
blue	0.624348
green	0.622471
white	0.615909
glowing	0.523828
black	0.518938
sky	0.517546
striped	0.516534

- Query “k9”, Metadata+NYTimes, Metadata Only

Word	Cosine distance
hund	0.452182
corgi	0.436650
canine	0.434721
udstillinger.	0.430973
gsp	0.424129
dachsunds	0.415784
shorthaired	0.415089
sennenhunde	0.408709
haustier	0.407449
10%2f2005	0.405896
kelpie	0.404788
k-9	0.403975
butcherwhite%2c	0.399475
pekingese	0.398497
kennels%0asitejoplin	0.396222
dog	0.395836
malinois	0.394518
hunde	0.393059
staffordshire_bull_terrier	0.389093
kennels	0.388342

Word	Cosine distance
hund	0.461504
australiancattledog	0.454843
germanshepherd	0.446335
canine	0.444728
shorthaired	0.443189
sennenhunde	0.432801
puppy	0.429908
neuter	0.429756
perro	0.426570
apbt	0.425053
bostonterrier	0.423600
mynoias	0.423025
k-9	0.420620
pekingese	0.417837
ciscc	0.415530
mastiff	0.413818
hundewelpen	0.412476
udstillinger.	0.410803
mutt	0.409479
dog	0.408160

Image analysis: concept driven

Terminal — ssh — 80×42

Word	Cosine distance
jeans	0.516866
purpleglory	0.510727
carangoides	0.473018
waistband	0.469072
zipper	0.468741
nanometres	0.468142
trousers	0.466580
placket	0.462134
eyes	0.453842
scrunched	0.446964
scyliorhinus	0.443418
turnbacks	0.442510
blue	0.441525
acuminate	0.441139
uraeginthus	0.441113
tomentose	0.440559
trevally	0.438237
hylexetastes	0.438134
gape	0.437152
waistcoat	0.433555
bodysuit	0.432041
starlike	0.431736
shirt	0.431235
margarornis	0.429040
suspenders	0.428517

Descriptive adjectives and nouns

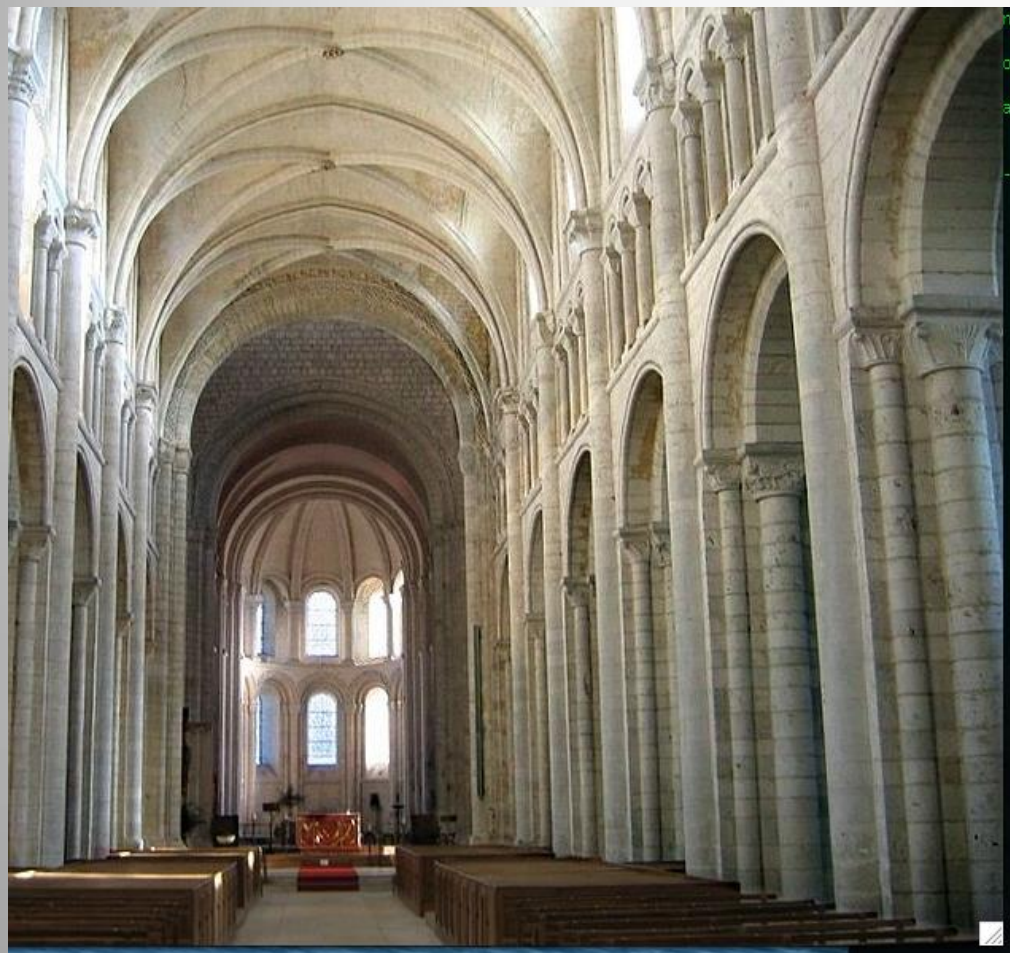
Terminal — ssh — 80×41

Word	Cosine distance
panteleimon	0.616204
ierarhi	0.613494
style	0.610755
romanesque	0.593696
rayonnant	0.584500
roman	0.574076
gothic	0.569602
everilda	0.569123
remacle	0.567654
pentarchy	0.560716
sernin	0.552525
beguinage	0.550396
procula	0.542281
malankar	0.540048
paternus	0.539920
servatius	0.536393
stiftskirche	0.535514
agone	0.534483
basilika	0.531380
frescoed	0.528966
romanides	0.527716
trophime	0.526481
paleologan	0.525358
mugnano	0.525105
haymanout	0.522536
saint	0.522463
catholic	0.522242
kilmacduagh	0.522187
eremitani	0.521595
alacoque	0.521526
hymnodist	0.520131
concezione	0.517924
baroque	0.517868
tempietto	0.515922
patron	0.512890
skiti	0.512504
basilique	0.512321
motherchurch	0.511282

Descriptive verbs and nouns

Word	Cosine distance
leaves	0.672294
eating	0.653425
eucalyptus	0.618348
cupressinum	0.609821
dacrydium	0.607026
gumnut	0.605137
sophora	0.594255
pinecones	0.585673
perfoliata	0.578496
cichorium	0.577513
eat	0.576809
intybus	0.575559
jackfruit	0.575058
saprophytes	0.574339
bipinnate	0.572866
guajava	0.571780
ribwort	0.570426
citriodora	0.569942

General and specific terms



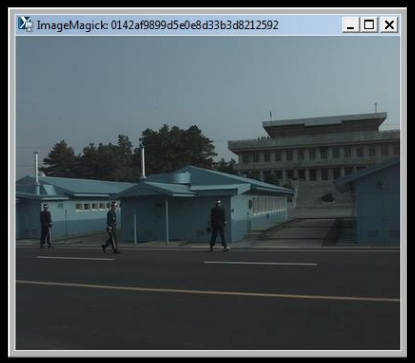
```
nary word!
 sentence (EXIT to break): /p/lscratchf/mayhew5/ImageCLEF/images/1
o OS: display /p/lscratchf/mayhew5/ImageCLEF/images/1/71.jpg&

atchf/mayhew5/ImageCLEF/images/1/71.jpg  Position in vocabulary: 2

                                   Word       Cosine distance
-------------------------------------------------------------
                                  abbey              0.926146
                                pointed              0.920063
                                   nave              0.895640
                                    has              0.872950
                              transverse             0.867482
                                   ribs              0.862108
                                 church              0.803468
                                     de              0.721272
                                  style              0.603996
                                chancel              0.597058
                                  roman              0.596090
                                  third              0.588714
                                  cover              0.587076
                                     la              0.584817
                                mugnano              0.578046
                              cathedral              0.574122
                              rayonnant              0.572280
                                  under              0.568329
                                velicat              0.566361
                                transept             0.560859
                               triforium             0.558442
                              groundplan             0.557511
                                 achonry             0.553789
                                coronati             0.552653
                               obradoiro             0.552076
                                 largest             0.551325
                                reparata             0.550085
                               transepts             0.549157
                             monterrubio             0.548030
                                   right             0.547457
                                  aisled             0.547191
                              hammerbeam             0.546126
                                    apse             0.546108
                                 chapels             0.546059
                                  chevet             0.545267
                                cloister             0.544374
                                    ancud            0.543130
                                vallalta             0.541016
                                  indeed             0.540135
                                  sernin             0.538364
Enter word or sentence (EXIT to break):
```

Semantic-based examples

```
System call to OS: ./eogy.sh 01033aa55eb945b174d5d73ea1cd91&
Opening image /p/lscratchf/videousr/YLI/raw_data/pearce7/2nd/yfcc100m_images/0000/01033aa55eb945b174d5d73ea1cd91

Word: 01033aa55eb945b174d5d73ea1cd91  Position in vocabulary: 7915

                                   Word       Cosine distance
------------------------------------------------------------------------
                                  wooden              0.999018
                                falcon%3f             0.998688
                                 eagle%3f             0.998636
                                   sachi              0.998590
                                  carved              0.998049
                                   sushi              0.946142
                              2009.12.11              0.919562
                                 turning              0.888647
                               andromeda              0.848497
                                 japonica             0.843746
                                   pieris             0.841031
                                 banshees             0.792362
                                 thrashers            0.791142
                                      96              0.787241
                                  ikebana             0.784596
                                 railroads            0.781760
                            flowerarrangement         0.780012
                                 touch%2c             0.777637
                                   nordic             0.771064
                             copyright%3a             0.769514
                                 japanese             0.768818
                        %e9%bb%83%e5%b1%b1            0.768597
                                   saipan             0.753743
                            4%2f22%2f2007             0.753694
                                    shen              0.752610
                                  brixton             0.752599
                                11.2.13-64            0.752433
                                 garapan%2c           0.749263
                                   suede              0.747911
                              festival-9177           0.747100
                                chinasquare           0.741353
                                 satellite            0.739753
                                  jiuzhou             0.739832
```

wooden
falcon%3f
eagle%3f
sachi
carved

```
System call to OS: ./eogy.sh 0142af9899d5e0e8d33b3d8212592&
Opening image /p/lscratchf/videousr/YLI/raw_data/pearce7/2nd/yfcc100m_images/0000/0142af9899d5e0e8d33b3d8212592

Word: 0142af9899d5e0e8d33b3d8212592  Position in vocabulary: 24503

                                   Word       Cosine distance
------------------------------------------------------------------------
                                    joint              0.884869
                                 security              0.847315
                                 soldiers              0.831235
                                battalion              0.822842
                                     eeas              0.794635
                                     csdp              0.794508
                                demarcation            0.789777
                                  somalia              0.787522
                                 9-18-06               0.774027
                               panmunjom%2c            0.762576
                                  sicilia              0.762118
                               gnu%2flinux             0.751620
                                   uganda              0.748957
                                    stand              0.745661
                                      lan              0.742777
                                barton%27s             0.733356
                               assignment              0.731707
                              honolulu%2c              0.728423
                                 aerodrome             0.728140
                                   waikiki             0.727896
                                  airfield             0.727810
                                 republika             0.725747
                                   heights             0.725732
                                   missing             0.725317
                                     ncsoft            0.724511
                              d%27affaires             0.724282
                                 republica             0.720826
                                    wallop             0.716066
                                       aac             0.716034
                                    toasts             0.715368
                                 charg%c3%a9           0.713974
                                     honor             0.713221
                                   choice,             0.711438
                                  oahu%2c              0.711373
                                  portions             0.711172
                                    viewed             0.710991
                                    battle             0.710911
                                     guest             0.709658
                                angleterre             0.709283
                                  queenston            0.709230
```

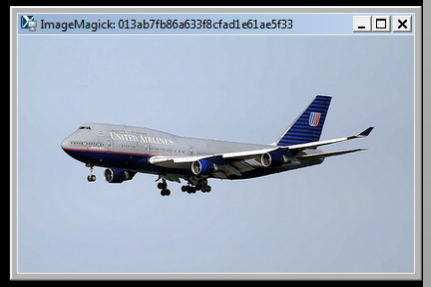


joint
security
soldiers
batallion

```
System call to OS: ./eogy.sh 01267f59f5a2ef449dbfbf30d080fb&
Opening image /p/lscratchf/videousr/YLI/raw_data/pearce7/2nd/yfcc100m_images/0000/01267f59f5a2ef449dbfbf30d080fb

Word: 01267f59f5a2ef449dbfbf30d080fb  Position in vocabulary: 16838

                                   Word       Cosine distance
------------------------------------------------------------------------
                                   salute              0.997519
                                 stickmen              0.997018
                                    dubya              0.996151
                                     noel              0.919468
                                 %28usa%29             0.830412
                                %28guitar%29           0.829135
                                   venter              0.821679
                                darcy%27s              0.813508
                                 shelter,              0.810012
                                  notices              0.809130
                                     chhs              0.806453
                                     curb              0.782679
                                 duncanville           0.781747
                                       co              0.778819
                                 sidewalk              0.774304
                                  pastels              0.764487
                                   tailed              0.759748
                                  sisters              0.758384
                                    molly              0.748624
                          rays_20131005_193702         0.746620
                                      sox              0.738984
                                  smarties             0.737818
                                  marbles              0.736062
                                  treetop              0.735861
                                    candy              0.735677
                                schlenkerla            0.735609
                               anticipation            0.734956
                                     scan              0.733751
                                     alds              0.732265
                                 today%21              0.725487
                                     bean              0.712429
```

salute
stickmen
dubya
noel

```
System call to OS: ./eogy.sh 013ab7fb86a633f8cfad1e61ae5f33&
Opening image /p/lscratchf/videousr/YLI/raw_data/pearce7/2nd/yfcc100m_images/0000/013ab7fb86a633f8cfad1e61ae5f33

Word: 013ab7fb86a633f8cfad1e61ae5f33  Position in vocabulary: 22623

                                   Word       Cosine distance
------------------------------------------------------------------------
                                  aircraft             0.998003
                                   airline             0.996767
                                  airplane             0.941008
                                       747             0.934217
                                  aviation             0.850612
                              a%c3%a9rodrome           0.849825
                                 %23avgeeks            0.844182
                                 %23avgeek             0.841904
                                    n194ua             0.832353
                                      solo             0.827937
                                      ksfo             0.821957
                                picturesque            0.821452
                                    32000ft            0.818409
                              highway%0aeast           0.813534
                                   cdg-ord             0.813476
                                       lan             0.811969
                                   f-giua              0.806512
                                    ailes              0.804225
                                  mandalay             0.804199
                                   waikiki             0.798282
                                     ncsoft            0.798160
                              honolulu%2c              0.793537
                            transmiss%c3%a3o           0.792720
                                az%0abarraza           0.792516
                                   notower             0.792107
                                 hawaii%2c             0.782302
                                      ebnm             0.782122
                                   oahu%2c             0.781455
                                     cargo             0.780381
                                     namur             0.779466
                                    boeing             0.779284
                                   belgien             0.771630
                                    phulay             0.768843
                                       cdg             0.759586
                            county%0atucson%2c         0.759392
                                       jet             0.753025
                                     sands             0.751161
                                    flight             0.750508
                                  par%c3%ads           0.741097
                                   bubbles             0.739848
```



aircraft
airline
airplane
747
aviation

Summary

- Multimodal vector space
  - Deep learning to understand image space
  - Final layer replacement with semantic methodologies
  - Promising results
    - Wikipedia Dataset
    - YFCC100M Dataset
- Future Work
  - Integration with UC Berkeley's Caffe
  - Use a better learner (e.g., GoogleNet)
  - Full back-propagation for final layer
  - Additional layers to be added for more complexity

References

- [1] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet Classification of Deep Convolutional Neural Networks," NIPS 2012
- [2] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P. Manzagol, "Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Denoising Criterion", JMLR 2010
- [3] F. Feng, X. Wang, and R. Li, "Cross-modal Retrieval with Correspondence Autoencoder", ACM-MM 2014
- [4] R. Socher, M. Ganjoo, C. Manning, and A. Ng, "Zero-shot Learning Through Cross-Modal Transfer," NIP 2013
- [5] Y. Jia, "Caffe: An Open Source Convolutional Architecture for Fast Feature Embedding," UC Berkeley Vision Website 2013
- [6] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," NIPS 2013
- [7] K. Ni, R. Pearce, K. Boakye, B. Van Essen, B. Chen, E. Wang, "Large-scale Deep Learning on the YFCC Dataset," On Archives, 2015
- [8] M. Mahoney, "Large Text Compression Benchmark," March 2006
- [9] M. Baroni, G. Dinu, German Kruszewski, "Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors," ACL-2014