Training a Deep Learner on Open Source Imagery
Scale Issues

10 million images
Thousand classes
Curated data

100 million images
Millions of “classes”
User generated
Tagging Issues

Metadata Tags: 2015, April, Austria, Canon, Guenter, Gunter, Landscape, Leitenbauer, Wels, bild, bilder, canal, canale, city, flickr, foto, fotos, image, images, kanal, key, landschaft, photo, photos, picture, pictures, stadt, town, venedig, venetia, venezia, venice, wasser, water, www.leitenbauer.net, Osterreich, burano, island, insel, isola

Metadata Tags: BlinkAgain, my_gear_and_me

A Division of IQT Labs. Private - Do not disseminate without approval of Lab41 or In-Q-Tel
Multiple Object Issues

ImageNet iconic photos

Open source photos
A Need For Something More

Type in any search term and get back images that match

Large Vocabulary

Leverage Noisy Data

Multi-Concept Images
Traditional Optimization of Neural Networks

Final Layer is 2048 x 1000

ImageNet / Visual Genome

Cat      Dog      Car      Tree

Final Layer
1000 Labels

Deep Network
(19 Layers)
Large Number of Labels in YFCC

Final Layer is 2048 x 400,000+

YFCC 100M Tags

- Plane
- Phone
- Canine
- Cat
- Dog
- Car
- Tree
- K9
- Vase
- Book

Final Layer (400k+ Tags)

- Large vocabulary
- Very noisy
- Multiple tags apply
Unstructured data with large vocabulary?

Word Embeddings

- Optimize with respect to context
- Sample randomly to provide negative samples
Context of an image

Use the context (i.e., metadata) of an image to define its place in vector space

Metadata from Tags Field

2006 ana+capri capri goats italy mountains naples ocean pompeii rome ski+lift sorrento summer tour vacation venice yachts
Word2Vec vs Im2Vec

Final Layer is 2048 x 6,000,000

• Original word vector optimization:

\[
\max_{\mathbf{v}_i, \mathbf{v}_o} \sum_{o=1}^{P} \mathbb{E}_{w_I} \left[ \log \sigma (\mathbf{v}_o^T \mathbf{v}_i) \right] + \sum_{k=1}^{N} \mathbb{E}_{w_k \sim P_n(w_k)} \left[ \log (-\mathbf{v}_k^T \mathbf{v}_i) \right]
\]

• Optimization with respect to images:

\[
\max_{W, (\mathbf{v}_o)} \sum_{o=1}^{P} \mathbb{E}_{w_I} \left[ \log \sigma (\mathbf{v}_o^T f_W (\mathbf{v}_i)) \right] + \sum_{k=1}^{N} \mathbb{E}_{w_k \sim P_n(w_k)} \left[ \log (-\mathbf{v}_k^T f_W (\mathbf{v}_i)) \right]
\]
Implementation Specifics I

The GPU/CPU Split

- Wide Optimization on CPU
- Deep Optimization on GPU

YFCC 100M Tags

![Diagram showing YFCC 100M Tags and GPU/CPU split]
Pre-training with word vectors

- Pre-train on word corpus (YFCC / NYT / Wiki8B)
- Either optimize / don’t optimize the final layer
Implementation Specifics III

Sampling Methods

- Tensorflow implementation required positive sampling (uniform)
- Large distribution negative sampling requires a bit more thought
  - (Option I) Sample the next image’s labels
  - (Option II) Pre-sample each epoch

Metadata from Tags Field

2006 ana+capri capri goats italy mountains naples
ocean pompeii rome ski+lift sorrento summer
tour vacation venice yachts
• Words not in the vocabulary
• Update if optimizing last layer

\[
\log \sigma(v_O^TW_Fv_F) + \sum_{k=1}^{K} E_{w_k \sim P_n(w_k)} \left[ \log(-v_k^TW_Fv_F) \right] \\
+ C_{I,O} \cdot \log \sigma(v_O^Tv_I) + (1 - C_{I,O}) \cdot \log \sigma(1 - v_O^Tv_I)
\]
Single Corpus Performance

Small Vocab: IAPRTC

Large Vocab: Visual Genome
Multi-Corpora Performance

Train: Visual Genome, Test: ESP Game
http://attalos-demo.labs.internal/