Images Worth A Million Words

Vishal Sandesara and Karl Ni

October 2016

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, Günter, 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, Österreich, burano, island, insel, isola

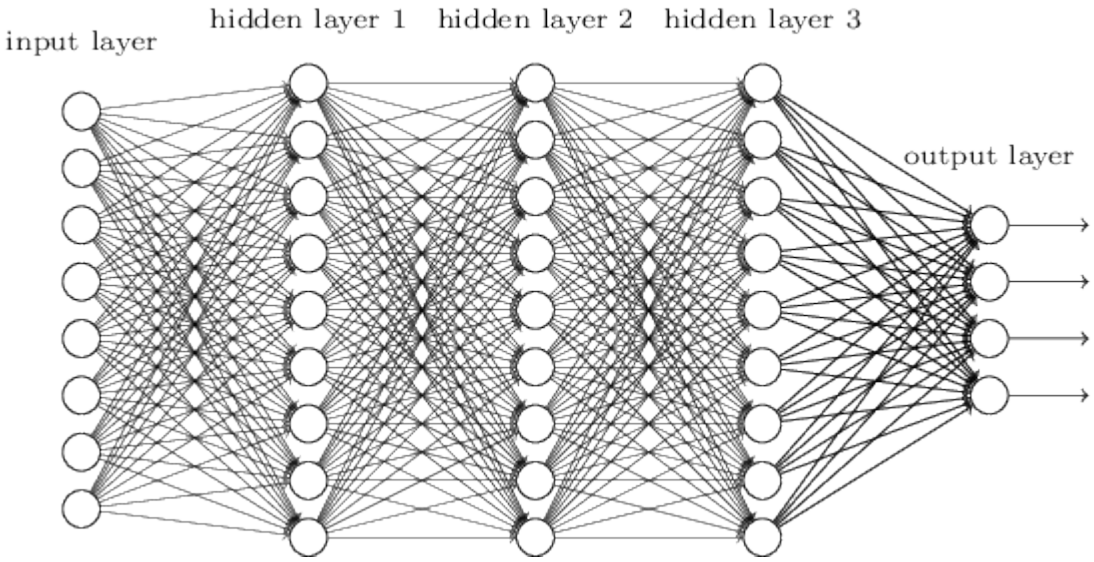
Metadata Tags: BlinkAgain, my_gear_and_me

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

Attalos

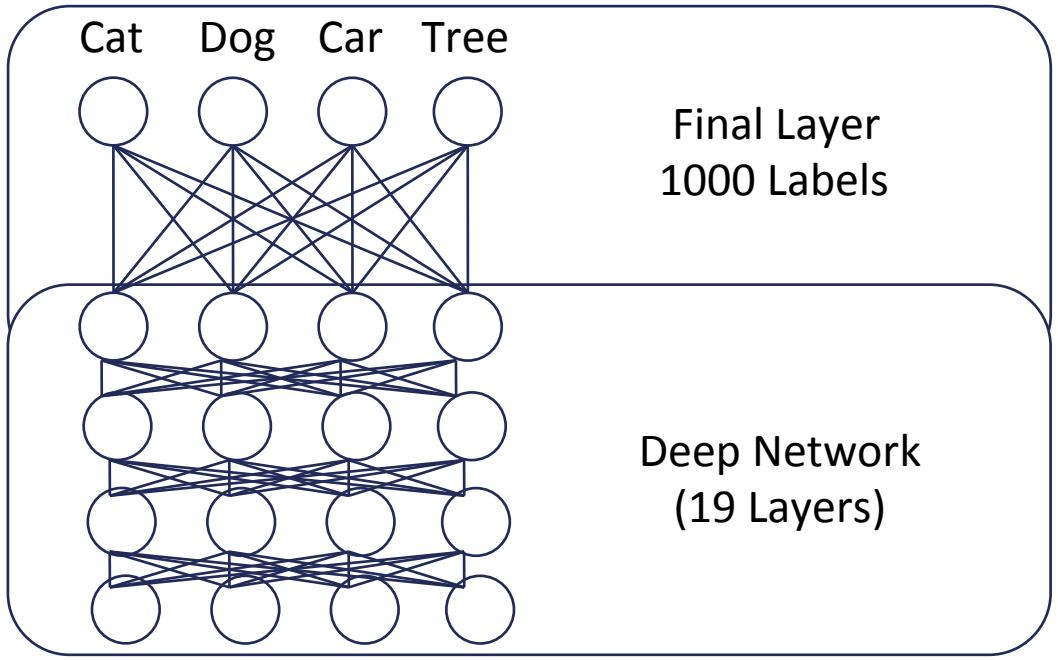
Searching images with words

LAB41

Traditional Optimization of Neural Networks

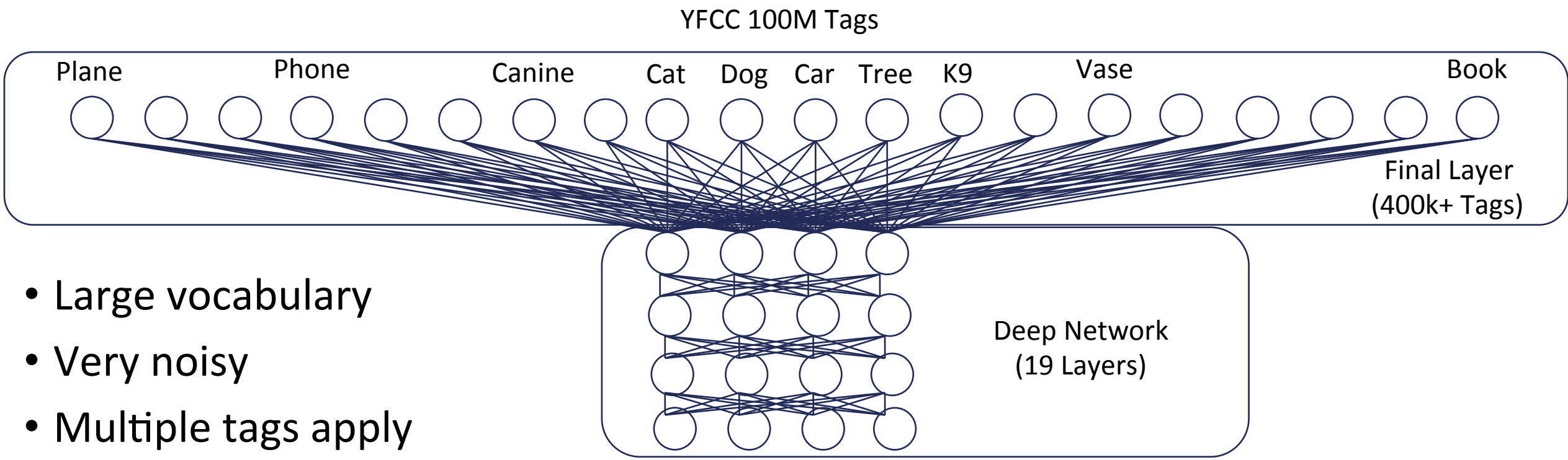
Final Layer is 2048 x 1000

ImageNet / Visual Genome



Large Number of Labels in YFCC

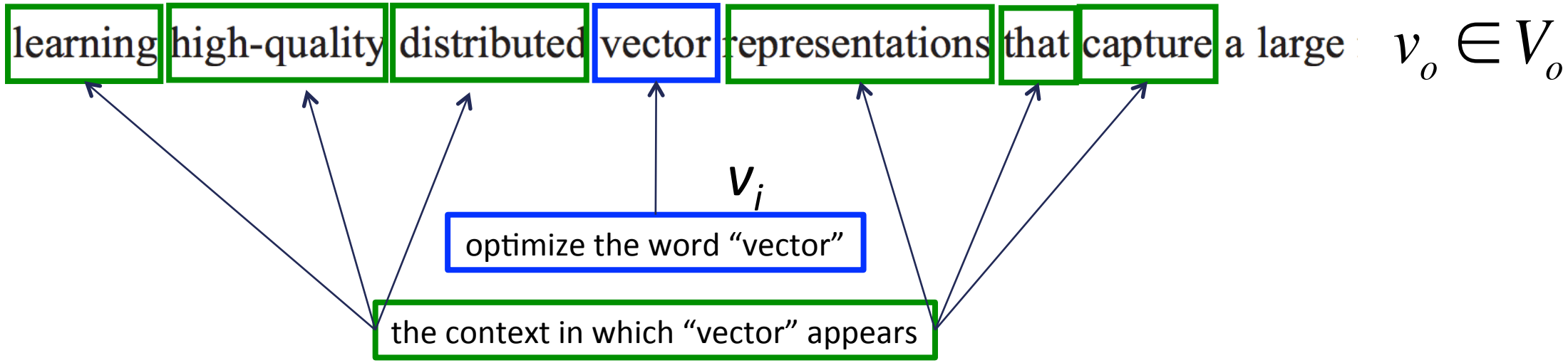
Final Layer is 2048 x 400,000+

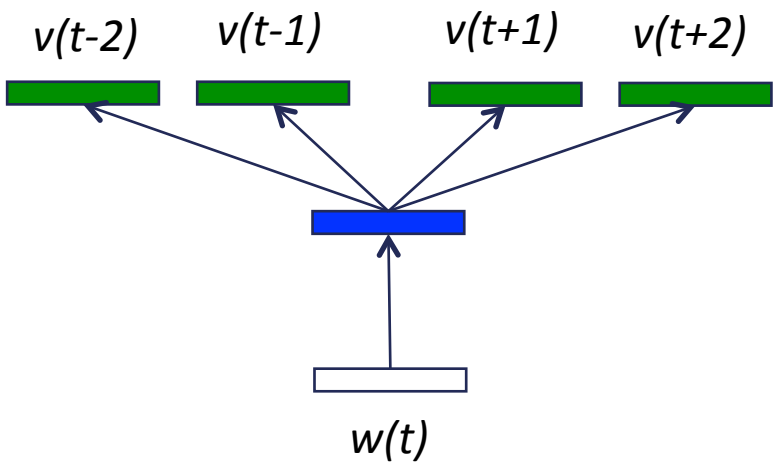


- Large vocabulary
- Very noisy
- Multiple tags apply

Unstructured data with large vocabulary?

Word Embeddings





- Optimize with respect to context
- Sample randomly to provide negative samples

Context of an image

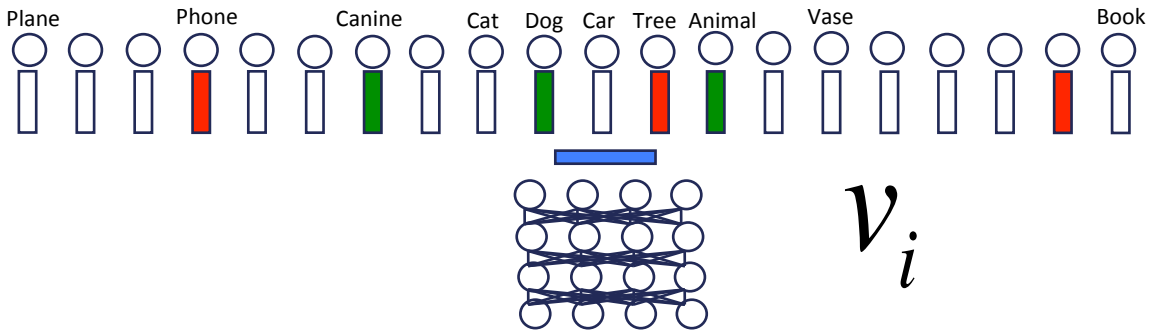
Use the context (i.e., metadata) of an image to define its place in vector space

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



$$v_o \in V_o$$

$$v_i$$

- Original word vector optimization:

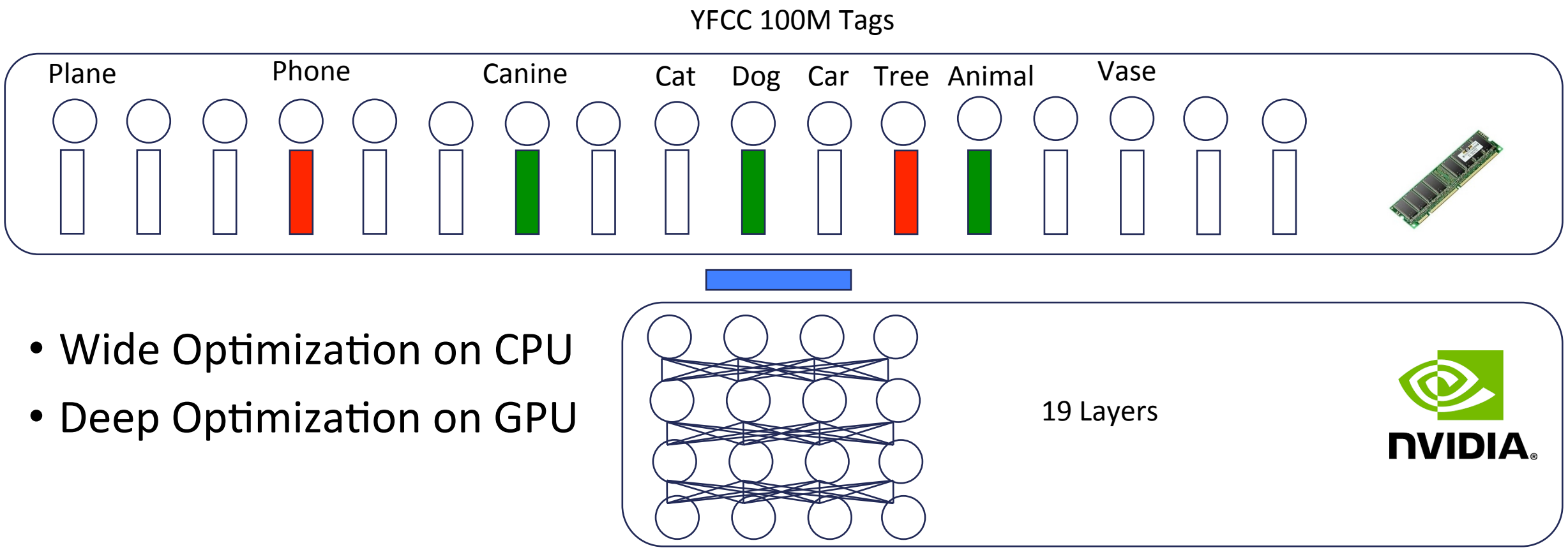
$$\max_{v_i, v_o} \sum_{o=1}^{P} \mathbb{E}_{w_I} \left[\log \sigma(v_o^T v_i)\right] + \sum_{k=1}^{N} \mathbb{E}_{w_k \sim P_n(w_k)} \left[\log(-v_k^T v_i)\right]$$

- Optimization with respect to images:

$$\max_{W,(v_o)} \sum_{o=1}^{P} \mathbb{E}_{w_I} \left[\log \sigma(v_o^T \boxed{f_W(v_i)})\right] + \sum_{k=1}^{N} \mathbb{E}_{w_k \sim P_n(w_k)} \left[\log(-v_k^T \boxed{f_W(v_i)})\right]$$

Implementation Specifics I

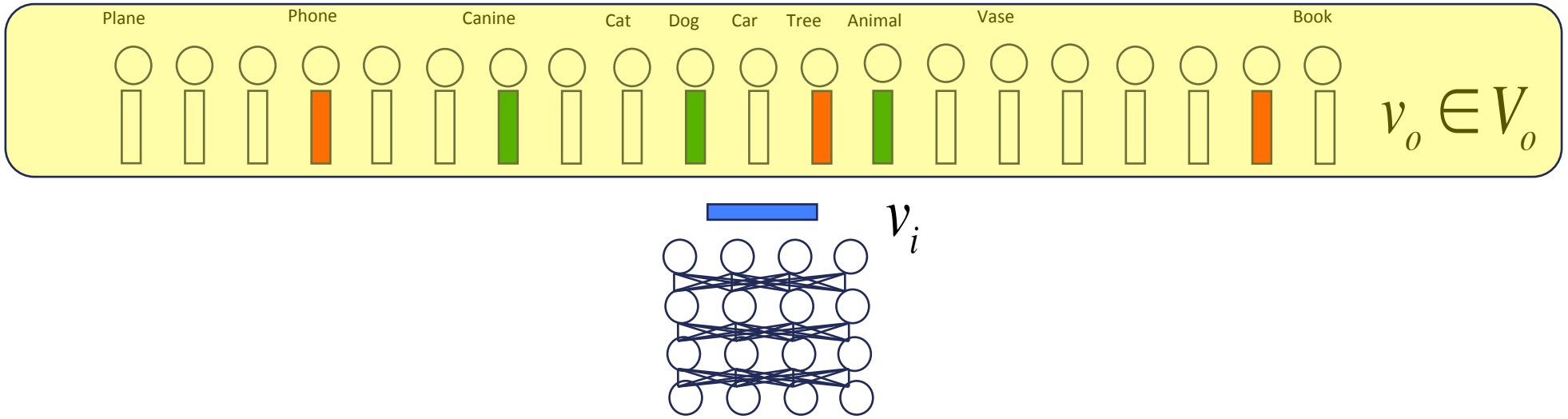
The GPU/CPU Split



- Wide Optimization on CPU
- Deep Optimization on GPU

Implementation Specifics II

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

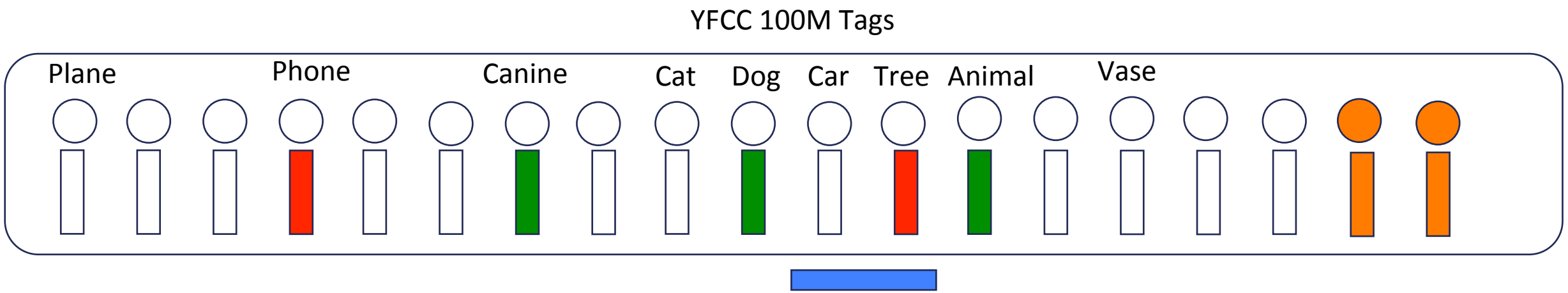
Metadata from Tags Field

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

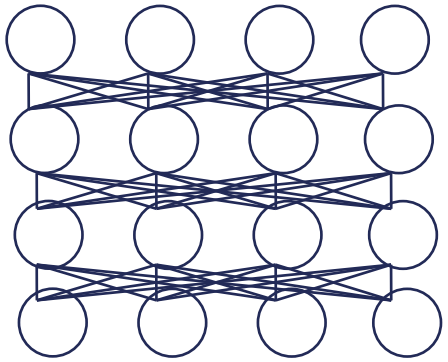
- Tensorflow implementation required positive sampling (uniform)
- Large distribution negative sampling requires a bit more thought
  - (Option I) Sample the next image's labels
  - (Option II) Pre-sample each epoch

Implementation Specifics IV

Out of Vocabulary Words

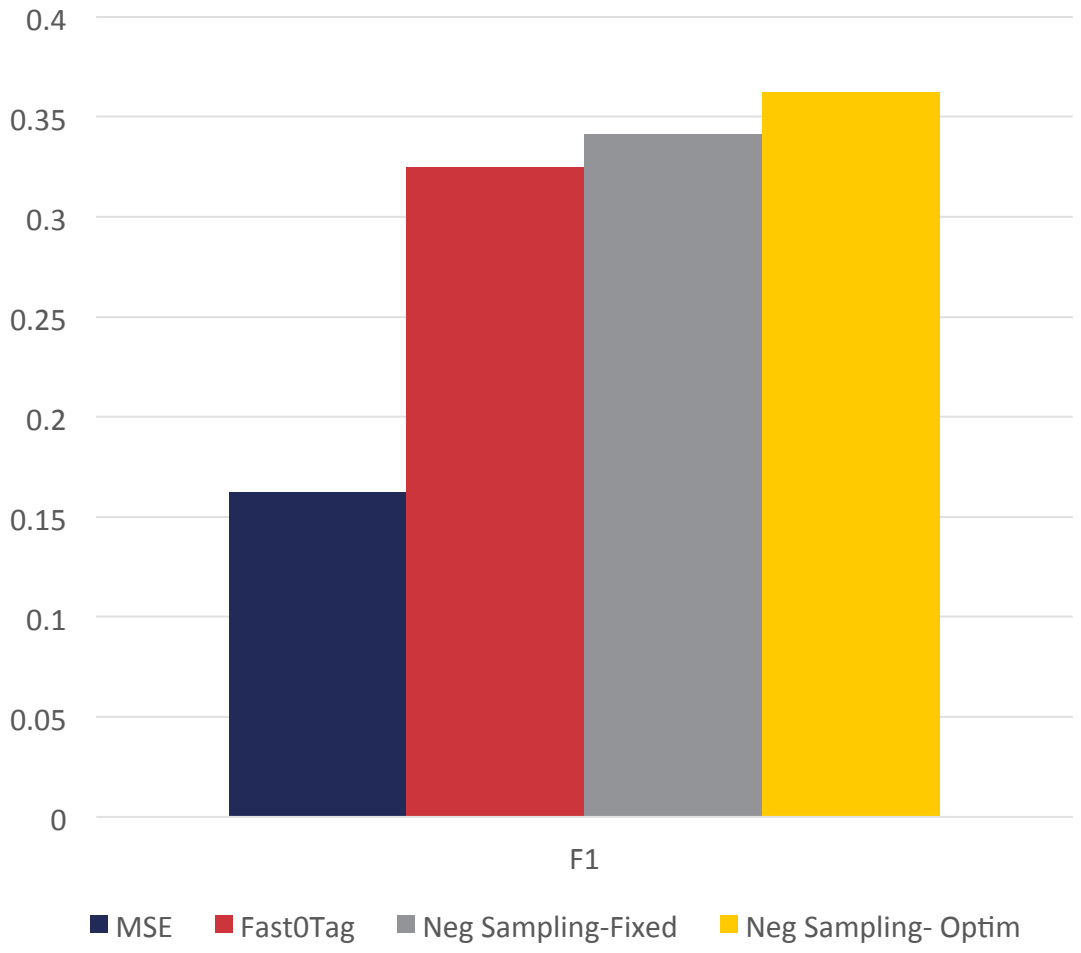


- Words not in the vocabulary
- Update if optimizing last layer

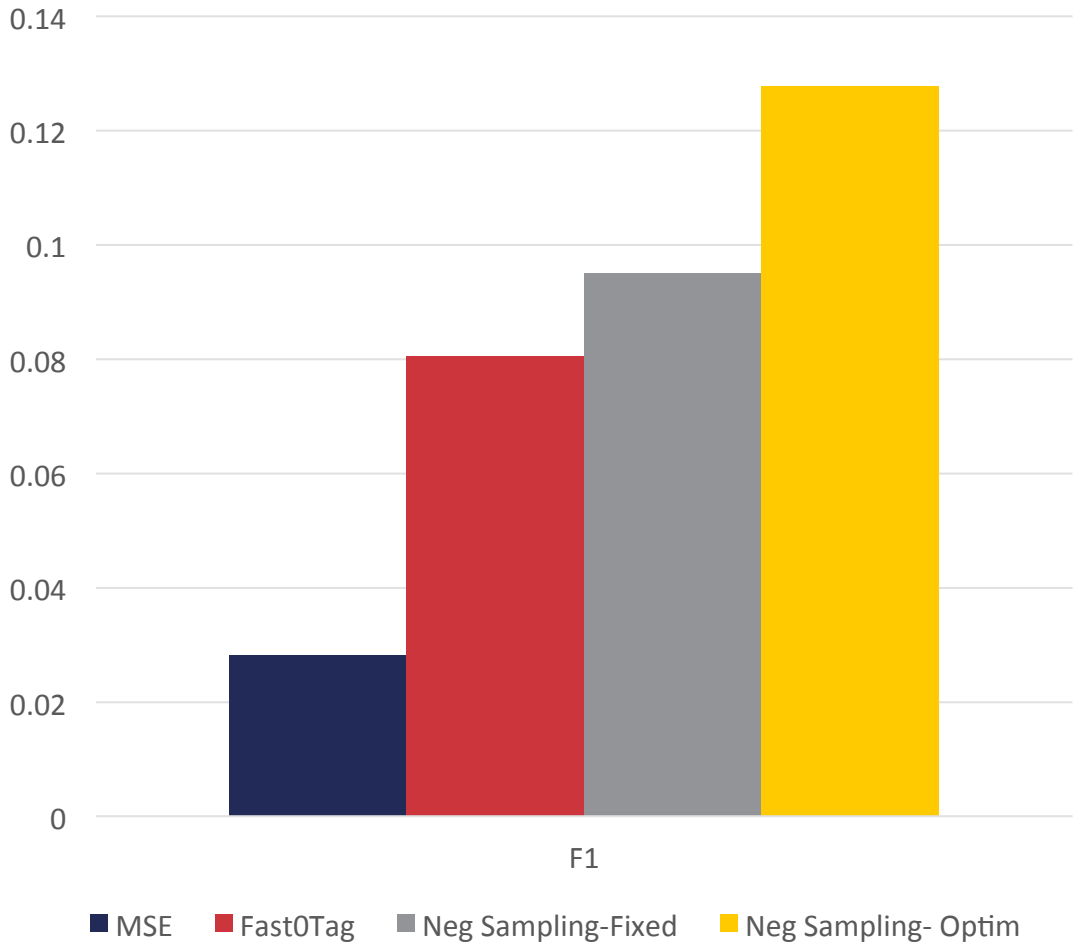


$$\log \sigma(v_O^T W v_F) + \sum_{k=1}^{K} \mathbb{E}_{w_k \sim P_n(w_k)} \left[\log(-v_k^T W v_F)\right]$$
$$+ C_{I,O} \cdot \log \sigma(v_O^T v_I) + (1 - C_{I,O}) \cdot \log \sigma(1 - v_O^T v_I)$$

Single Corpus Performance



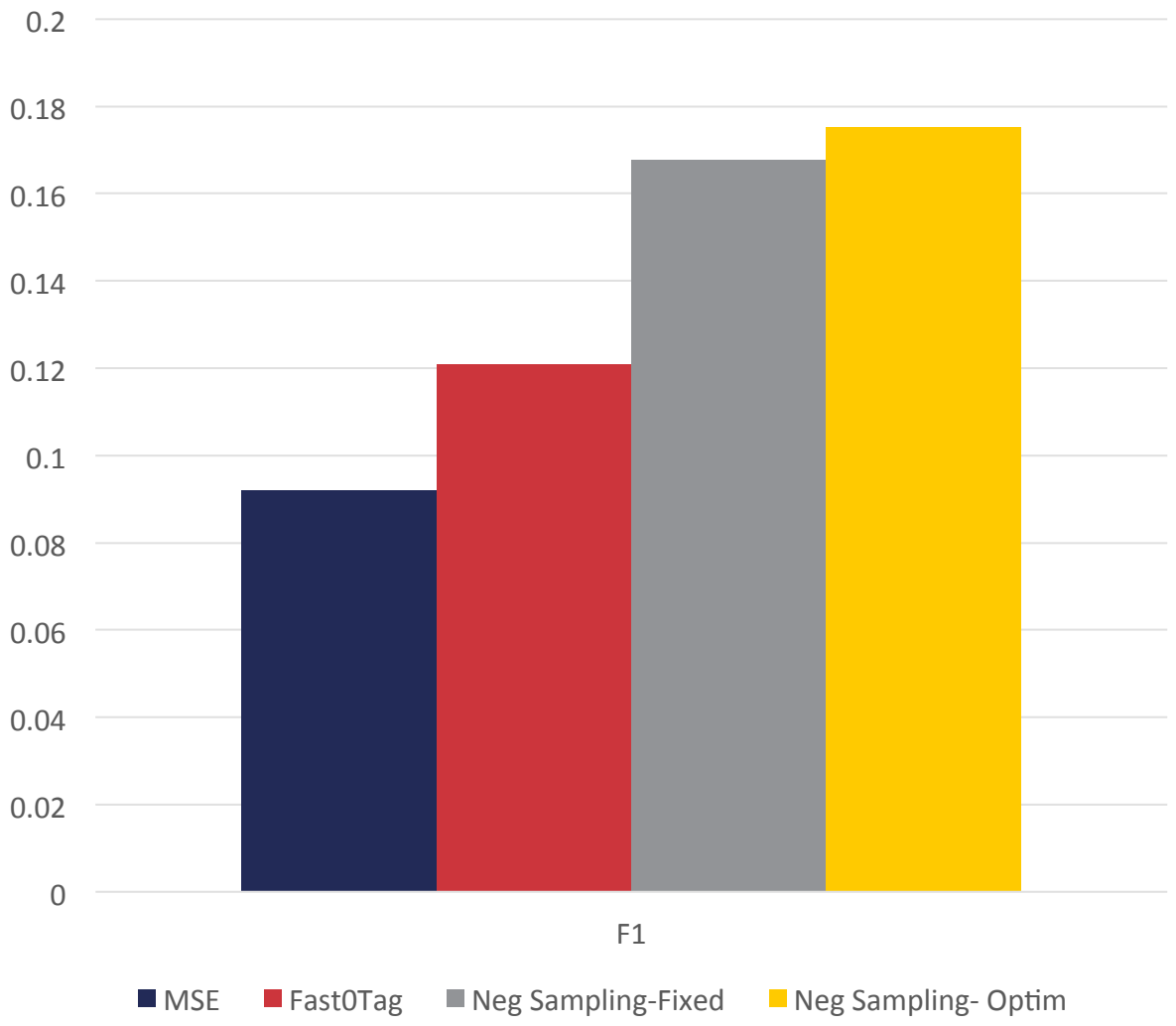
Small Vocab: IAPRTC



Large Vocab: Visual Genome

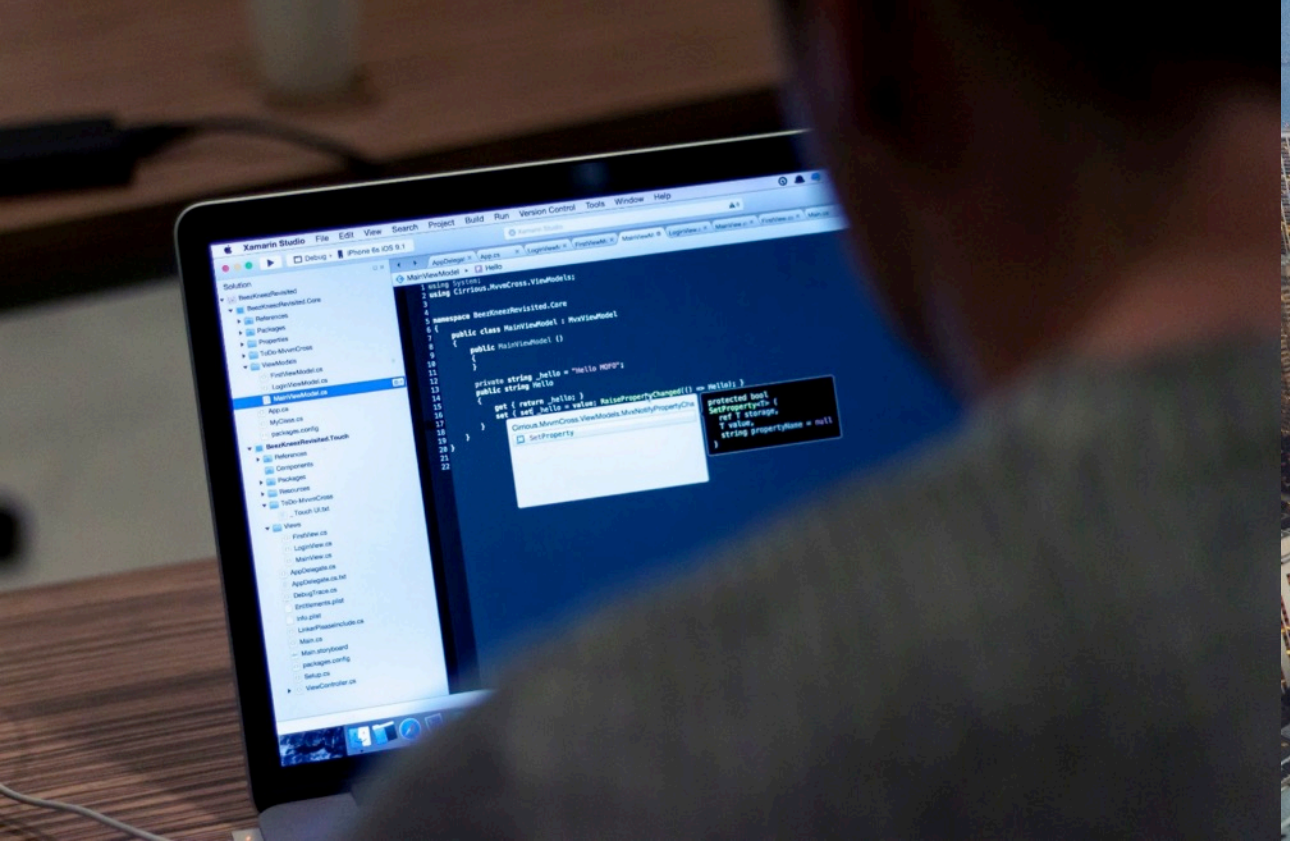
Multi-Corpora Performance

Train: Visual Genome, Test: ESP Game



DEMONSTRATION

http://attalos-demo.labs.internal/



EMAIL	TWITTER	WEBSITE	BLOG
kni@iqt.org vsandesara@iqt.org	@_lab41	www.lab41.org	gab41.lab41.org