Language and Vision: where we are and where we could go next

Raffaella Bernardi

University of Trento

June 9th, 2017

Language and Vision

Shared tasks

Image Captioning

a man is throwing a frisbee in a park

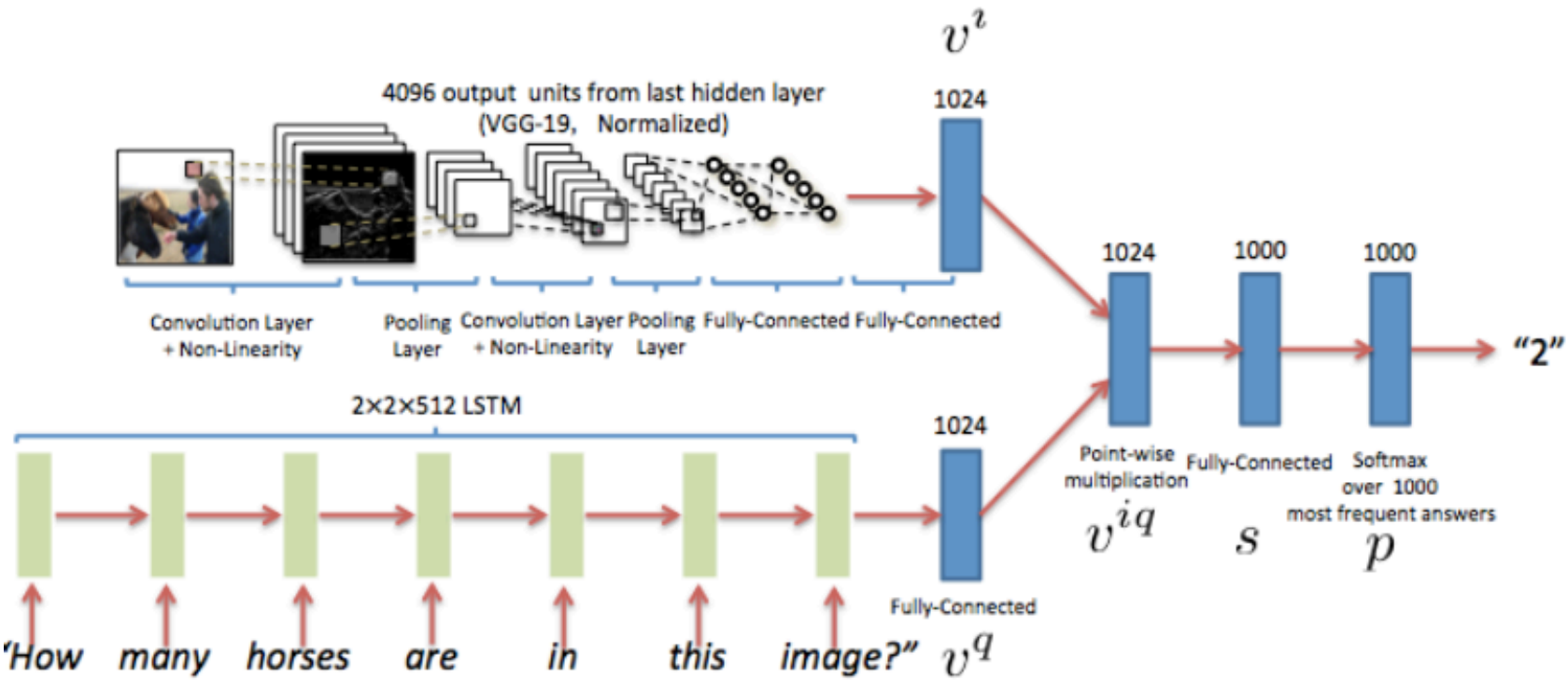
VQA

Yellow

What colour is the moustache made of?

LaVi Models

Parikh et. ali



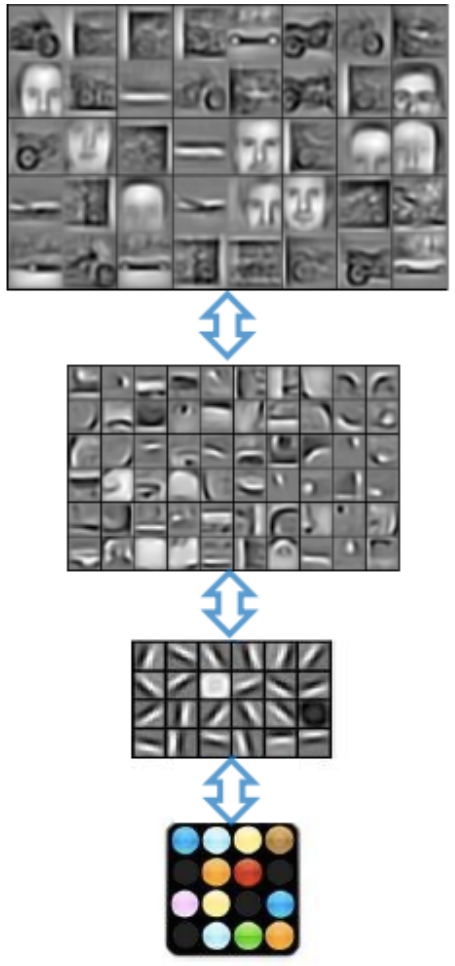
$$v^{iq} = v^i \circ v^q \qquad s = Wv^{iq} + b \qquad p_a = \frac{e^{s_a}}{\sum_{a'} e^{s_{a'}}}$$

CV Models

CNN: feature hierarchy

Deep Learning

- Deep architectures can be representationally efficient.
- Natural progression from low level to high level structures.
- Can share the lower-level representations for multiple tasks.



CV Models

Visualizing CNN layers

Aravindh Mahendran and Andrea Vedaldi 2015

Fig. 9 AlexNet inversions (all layers) from the representation of the "red fox" image obtained from each layer of AlexNet (Color figure online)

LaVi@UniTN

FOIL

task 1:
classification

People riding bicycles down the road approaching a dog.
FOIL

task 2:
foil word detection

People riding bicycles down the road approaching a **dog**.

task 3:
foil word correction

People riding bicycles down the road approaching a **bird**.

Original : A **narrow** room with various luggage and two men
FOIL : A **broad** room with various luggage and two men

Adjective

Original : A child wearing a very large and **loosely** tied necktie
FOIL : A child wearing a very large and **narrowly** tied necktie

Adverb

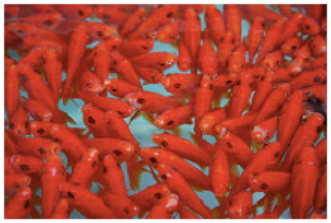
Original : A young boy **on** a couch holding two stuffed animals
FOIL : A young boy **beside** a couch holding two stuffed animals

Preposition

Original : A little girl trying to **push** a skateboard with other standing around
FOIL : A little girl trying to **pull** a skateboard with other standing around

Verb

Quantification

(a) 
(b) 
(c) 
(d) 
(e) 

Query: ___ fish are red.
Answers: (a) All, (b) Most, (c) Some, (d) Few, (e) No.

Most of the animals are dogs vs. *Three* of the animals are dogs.

Mass and Count nouns

a linguistic distinction

Stanford Encyclopedia of Philosophy.

Mass nouns: Examples: *milk*, *furniture* and *wisdom* .

they are invariable in grammatical number. Depending on the language [..] in English, mass nouns can be used with determiners like *much* and *a lot of*, but neither with one nor many.

Count nouns: Examples: *rabbit*, *table* and *idea*

they can be used in the singular and in the plural. [..] in English, count nouns can be employed with numerals like *one* and determiners like *many*, but not with *much*.

Mass and Count nouns

a linguistic distinction

Stanford Encyclopedia of Philosophy.

Mass nouns: Examples: *milk*, *furniture* and *wisdom* .

they are invariable in grammatical number. Depending on the language [..] in English, mass nouns can be used with determiners like *much* and *a lot of*, but neither with one nor many.

Count nouns: Examples: *rabbit*, *table* and *idea*

they can be used in the singular and in the plural. [..] in English, count nouns can be employed with numerals like *one* and determiners like *many*, but not with *much*.

Mass and Couns

Is there a perceptual distintincion?

Mass

milk

furniture

wisdom

Count

rabbit

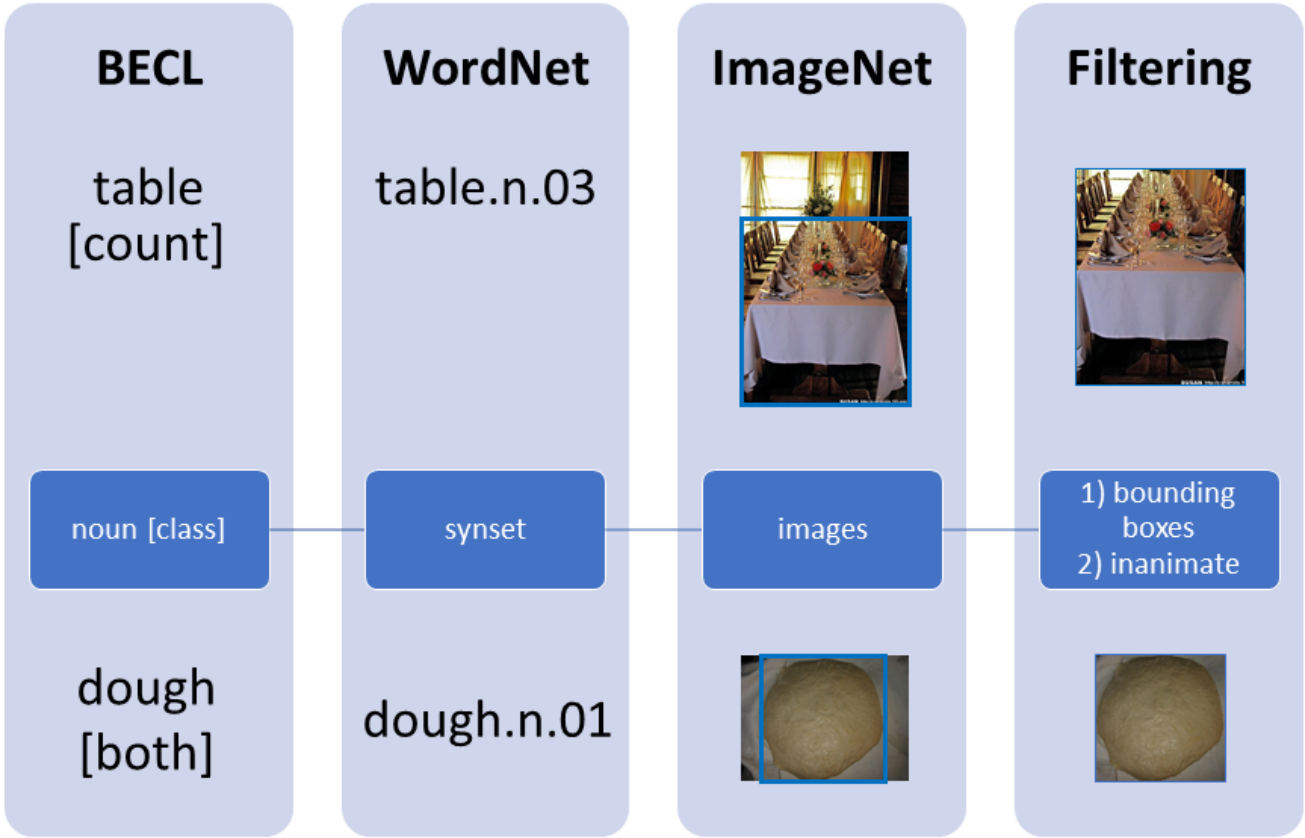
table

idea

Mass (Substance) and Count (objects)

Dataset: Construction

Starting point: Bochum English Countability Lexicon (BECL) Kiss et al. 2016



Mass (Substance) and Count (objects)

Dataset: Sample of mass nouns

Noun	Synset	Description	#occu.	#ima.
dough	dough.n.01	a flour mixture stiff enough to knead or roll	45	497
soil/dirt	soil.n.02	the part of the earth's surface consisting of humusand disintegrated rock	398/169	235
milk	milk.n.01	a white nutritious liquid secreted by mammals and used as food by human beings	386	196
coffee	coffee.n.01	a beverage consisting of an infusion of ground coffee beans;	356	159
coffee	coffee.n.02	any of several small trees and shrubs native to the tropical Old World yielding coffee beans	356	70

Mass and Count

Examples of images

dough.n.01

Mass and Count

Dataset: Numbers

Open American National Corpus (OANC) – metrics in BECL

	#syns	#uniq_N	#imgs (avg)	#imgs (range)	OANC freq (avg)	OANC freq (range)
mass	58	56	214.66	64 - 705	112.6	10 - 447
count	58	53	303.93	60 - 1467	1435.16	33 - 4121

Mass and Count

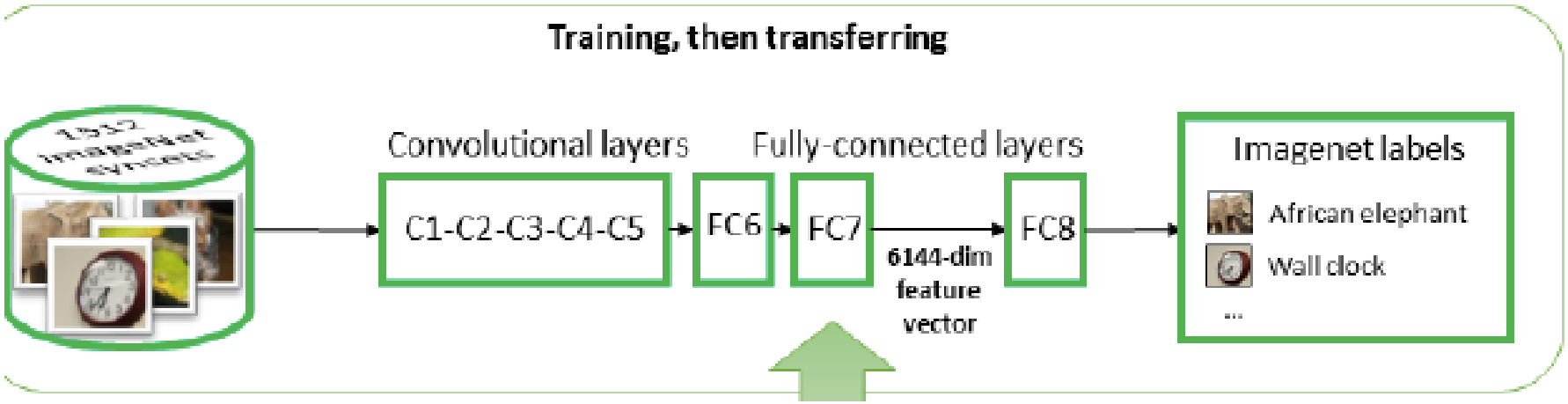
Variances

Synset feature vectors at given layer				
	<------ I N T R A ------>			
<---INTER--->				
[[0.3892	0.9384	2.8460	...]	*Image 1*
[1.9380	0.6930	1.2095	...]	*Image 2*
[0.3802	0.9830	-0.0293	...]	*Image 3*
[3.0939	3.5903	1.2093	...]	*Image 4*
[...	...	...	...]]	

CV Models

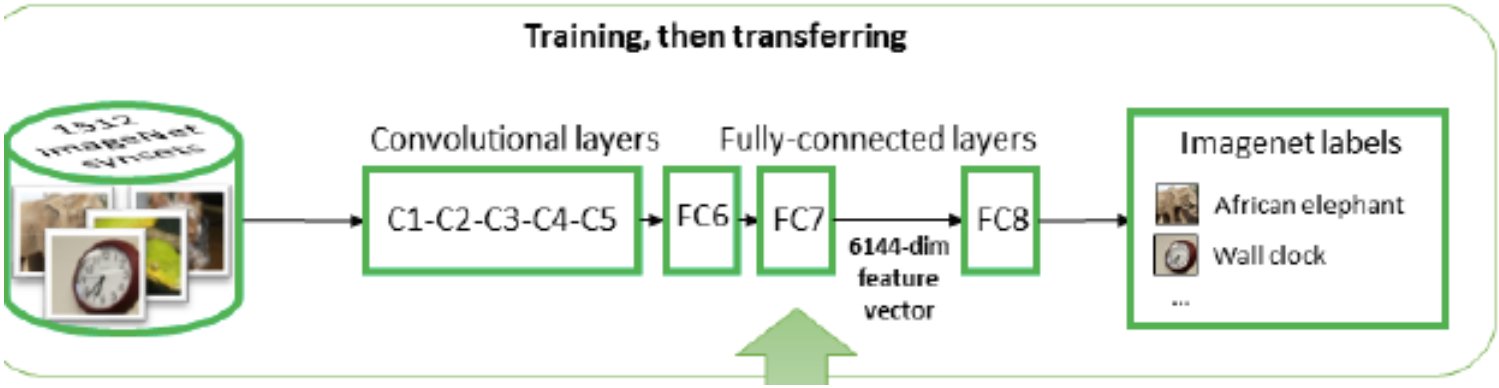
Convoutional Neural Network

We used the VGG-19 model (Simonyan and Zisserman (2014)), trained to classify objects.



Mass and Count

Feature layers of a CNN



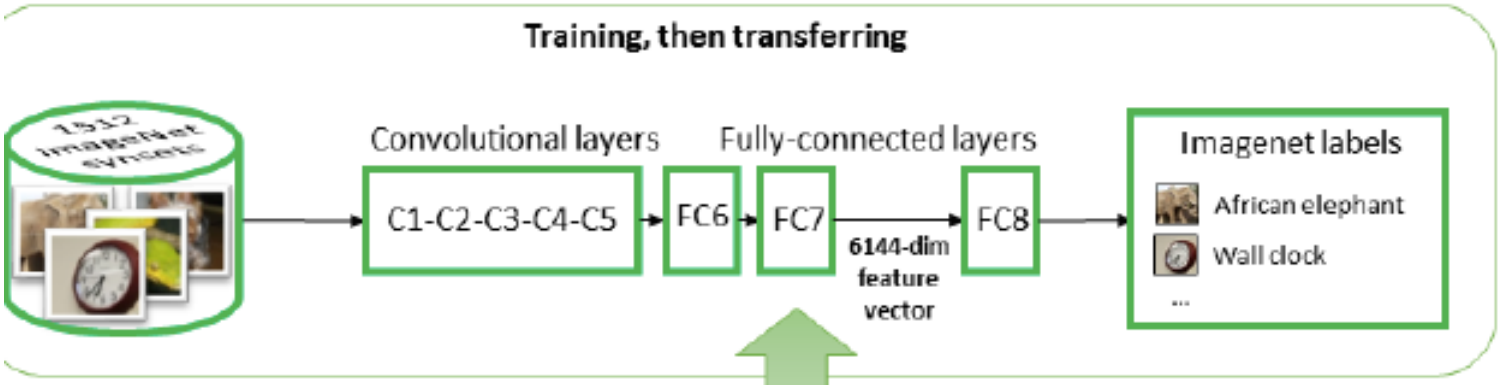
Each *Conv* consists of various hidden layers followed by a max pooling step which reduce the dimension by extracting salient features.

The *Conv* layers represent low-visual features (edges, texture, color) vs. the *fc* ones represent abstract features.

We compute the variances for the first and last $Conv2 - Conv5$ layers' outputs (low-features) and for the *fc* layers (abstract-features).

Mass and Count

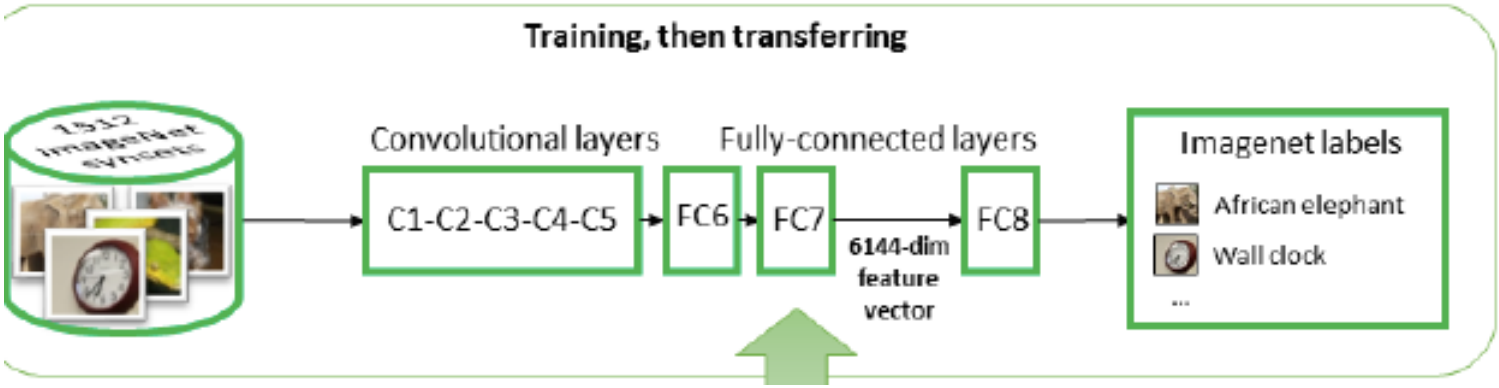
Feature layers of a CNN



Each *Conv* consists of various hidden layers followed by a max pooling step which reduce the dimension by extracting salient features.
The *Conv* layers represent low-visual features (edges, texture, color) vs. the *fc* ones represent abstract features.
We compute the variances for the first and last *Conv*2 – *Conv*5 layers' outputs (low-features) and for the *fc* layers (abstract-features).

Mass and Count

Feature layers of a CNN

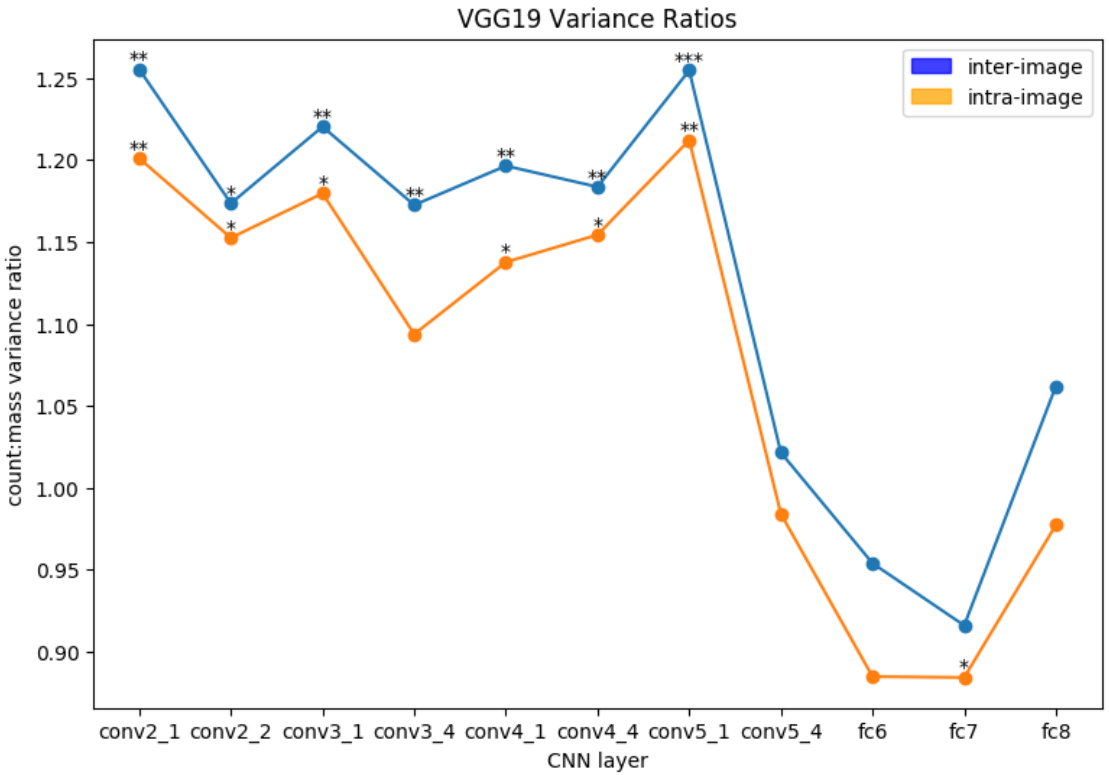


Each *Conv* consists of various hidden layers followed by a max pooling step which reduce the dimension by extracting salient features.
The *Conv* layers represent low-visual features (edges, texture, color) vs. the *fc* ones represent abstract features.
We compute the variances for the first and last $Conv2 - Conv5$ layers' outputs (low-features) and for the *fc* layers (abstract-features).

Mass and Count

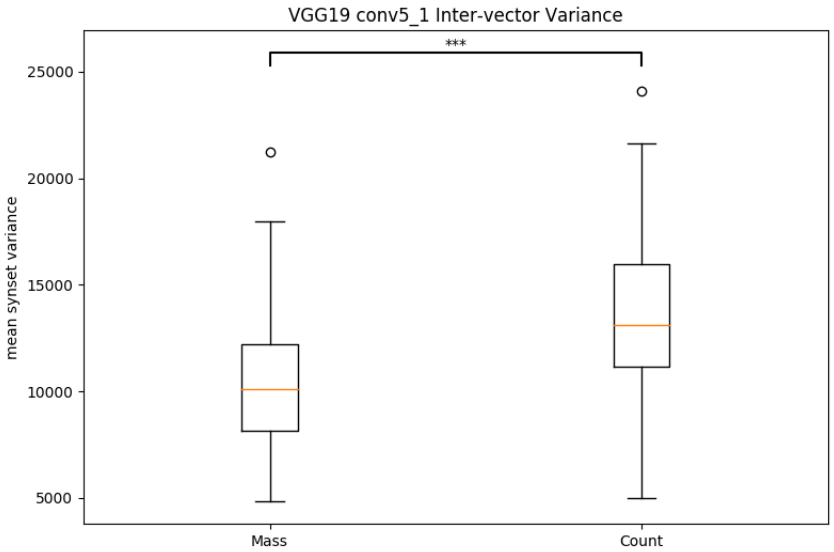
Variance: at which perceptual level?

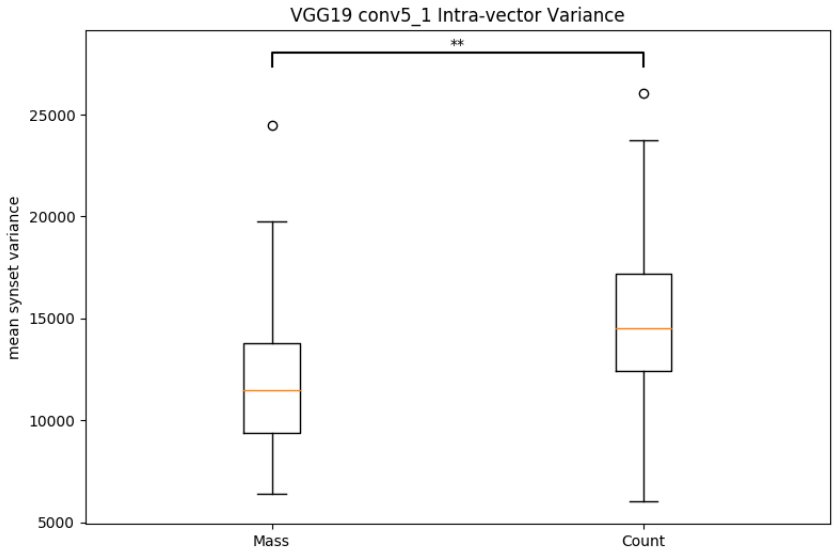
- ration(count/mass) $=1$ [variance of the two groups is equal]
- ration(count/mass) > 1 [mass's variance lower than count's]
- ration(count/mass) < 1 [mass's variance higher than count's]
- *** significant difference at $p < .001$; ** at $p < .01$; * at $p < .05$.



Mass and Count

Variance: $Conv5_1$?





Mass and Count

Synset with highest vs. lowest variances

Conv5_1 intra- variance		*Conv5_1 inter- variance*	
top-10	bottom-10	top-10	bottom-10
magazine_01 (c)	range_04 (c)	magazine_01 (c)	egg_yolk_01 (m)
salad_01 (m)	dough_01 (m)	shop_01 (c)	range_04 (c)
shop_01 (c)	mountain_01 (c)	salad_01 (m)	dough_01 (m)
church_02 (c)	mesa_01 (c)	machine_01 (c)	mountain_01 (c)
machine_01 (c)	flour_01 (m)	church_02 (c)	mesa_01 (c)
floor_02 (c)	milk_01 (m)	stage_03 (c)	milk_01 (m)
press_03 (c)	glacier_01 (m)	press_03 (c)	flour_01 (m)
stage_03 (c)	butter_01 (m)	floor_02 (c)	butter_01 (m)
pasta_01 (m)	egg_yolk_01 (m)	brunch_01 (m)	glacier_01 (m)
brunch_01 (m)	floor_04 (c)	building_01 (c)	sugar_01 (m)

Mass and Count: next step

Can CNN learn to quantify both objects and substance?

Most of the animals are dogs.

Most of the sand is dirty.

Mass and Count: next step

Can CNN learn that mass nouns (substance/liquid) are uncountable?






UniTN Team

Ionut

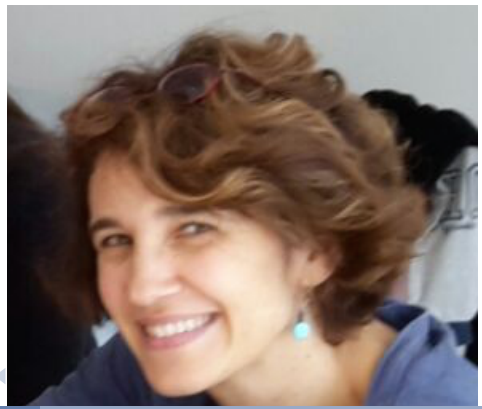
Sandro

Ravi

Addisson

Aurelie

me



FOIL

Word Pairs

Resources: For ADV, ADJ and V: VerbOcean, Computing Lexical Contrast, SimLex999. For PP, Berry et al. (1995), Examples:

PP	ADV	ADJ	VERBS
across at	actively passively	able unable	add divide
across behind	altogether partly	ancient modern	allow ban
across near	broadly narrowly	asleep awake	attack defend
across on	carefully carelessly	funny dull	begin end
across under	carelessly cautiously	huge tiny	catch miss
at below	comfortably uncomfortably	safe dangerous	deliver accept
at in	completely partially	ugly beautiful	drop add
at on	entirely partly	urban rural	knock beat
at under	formally informally	vertical horizontal	merge sell

FOIL

Dataset

	no. of unique ima.		no. of unique datap.		no. of unique pairs	
	Train	Test	Train	Test	Train	Test
Noun*	22,101	15,435	73,076	37,381	236	194
Verb	6314	2788	7925	3353	268	219
Adjective	15,640	9009	20,720	11,900	80	62
Adverb	1011	451	1044	475	38	36
Preposition	8733	5551	24,665	15,755	101	89
TOT	22,101	15,435	127,430	68,864	723	600