SAP Leonardo Machine Learning Research Retreat 2018 - Munich, Germany

Feedback Propagation in Deep Neural Networks

Vicente Ordóñez-Román

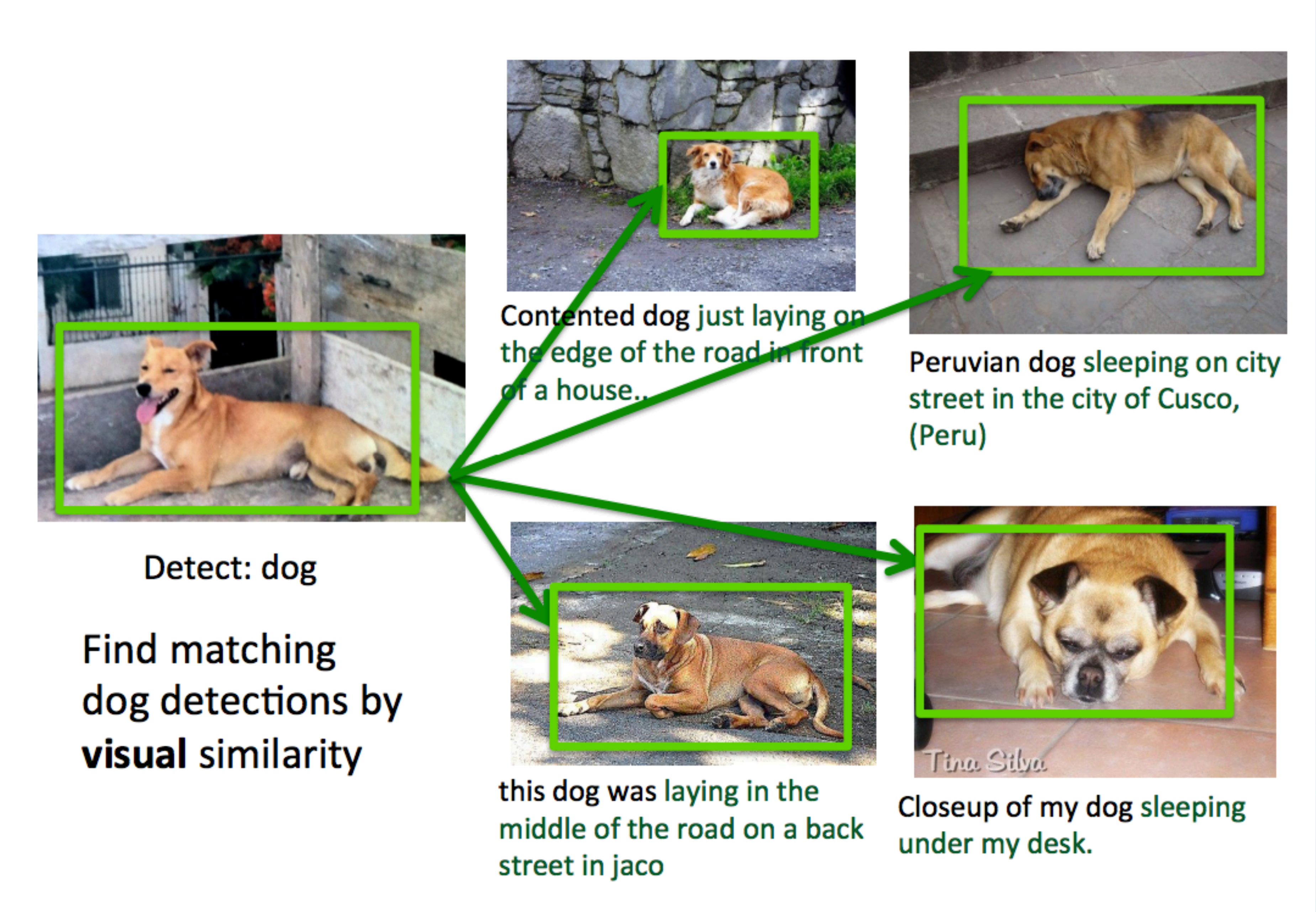
Assistant Professor
Department of Computer Science

University of Virginia

Past Work

Image Captioning

Large Scale Retrieval and Generation of Image Descriptions
V. Ordonez, X. Han, P. Kuznetsova, G. Kulkarni, M. Mitchell, K. Yamaguchi, K. Stratos, A. Goyal, J. Dodge, A. Mensch, H. Daume III, A.C. Berg, Y. Choi, T.L. Berg. International Journal of Computer Vision. **IJCV 2015**.



Past Work

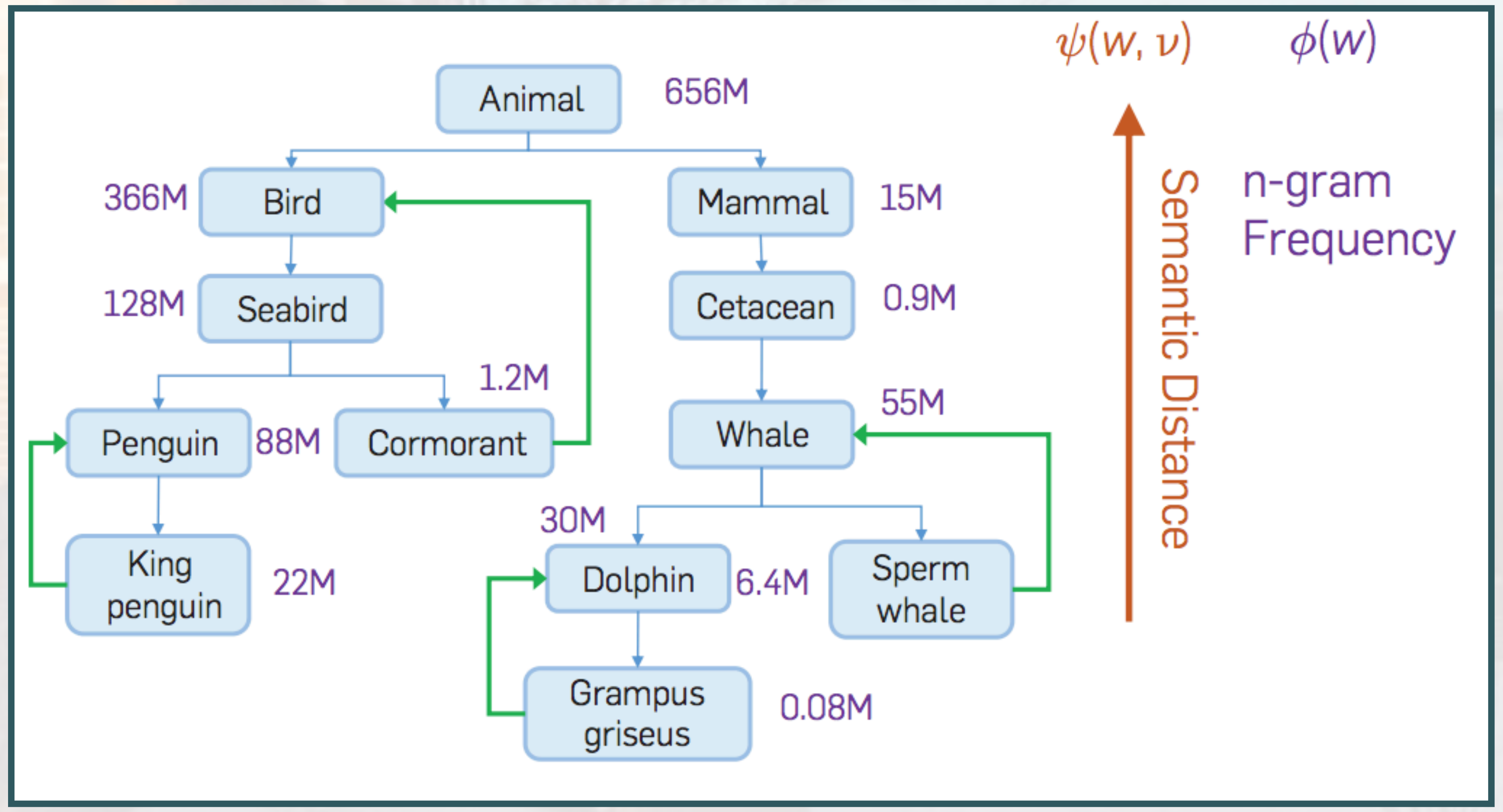
Image Captioning

Large Scale Retrieval and Generation of Image Descriptions
V. Ordonez, X. Han, P. Kuznetsova, G. Kulkarni, M. Mitchell, K. Yamaguchi, K. Stratos, A. Goyal, J. Dodge, A. Mensch, H. Daume III, A.C. Berg, Y. Choi, T.L. Berg. International Journal of Computer Vision. **IJCV 2015**.

Entry-level Categories

From Large Scale Image Categorization to Entry-Level Categories
Vicente Ordonez, Jia Deng, Yejin Choi, Alexander C. Berg, Tamara L. Berg. IEEE International Conference on Computer Vision. **ICCV 2013**.

Best Paper Award - Marr Prize



Past Work

Image Captioning

Large Scale Retrieval and Generation of Image Descriptions
V. Ordonez, X. Han, P. Kuznetsova, G. Kulkarni, M. Mitchell, K. Yamaguchi, K. Stratos, A. Goyal, J. Dodge, A. Mensch, H. Daume III, A.C. Berg, Y. Choi, T.L. Berg. International Journal of Computer Vision. **IJCV 2015**.

Entry-level Categories

From Large Scale Image Categorization to Entry-Level Categories
Vicente Ordonez, Jia Deng, Yejin Choi, Alexander C. Berg, Tamara L. Berg.
IEEE International Conference on Computer Vision. **ICCV 2013**.

Best Paper Award - Marr Prize

Referring Expressions

ReferItGame: Referring to Objects in Photographs of Natural Scenes
Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, Tamara L. Berg.
Empirical Methods on Natural Language Processing. **EMNLP 2014**.

Referit Game

Recent Work

Learning Beyond Pixel Data (e.g. Visual Layouts, Abstract Scenes)

Recent Work

Learning Beyond Pixel Data (e.g. Visual Layouts, Abstract Scenes)

Stating the Obvious: Extracting Visual Common Sense Knowledge
Mark Yatskar, Vicente Ordonez, Ali Farhadi.
North American Chapter of the Association for Computational Linguistics. **NAACL 2016**.

hold(people, umbrella**)**
wear(people, shoes**)**
hold(people, backpack**)**
covers(umbrella, people**)**

Quality

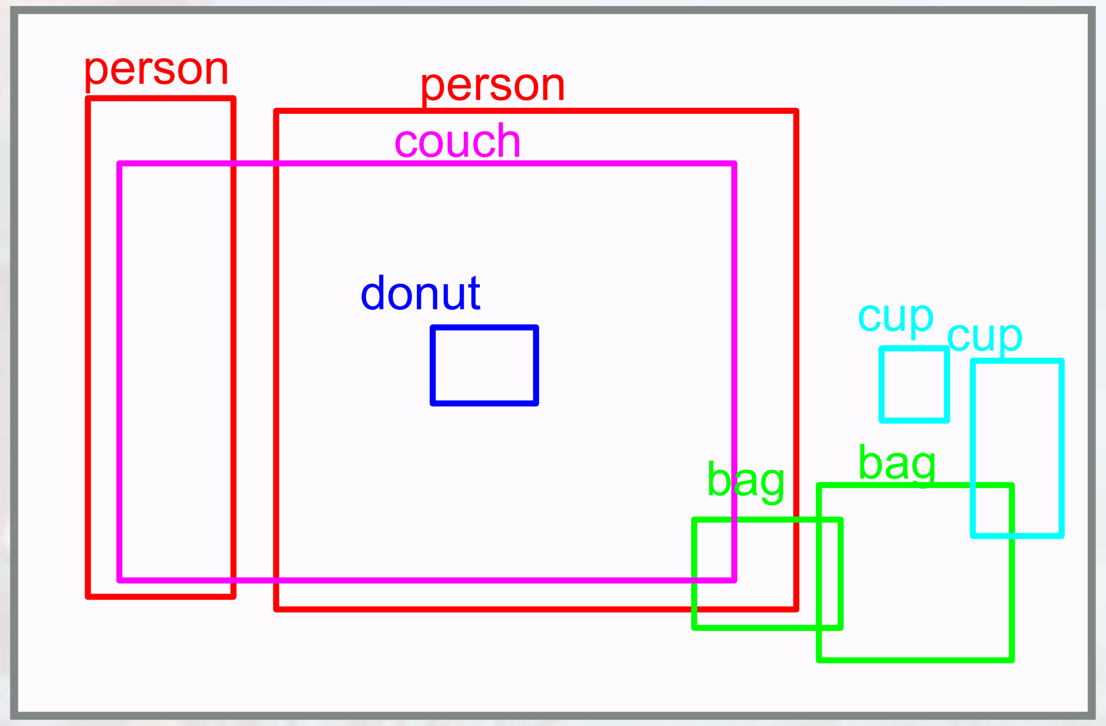
r(o1, o2)	holds(person, o2)
holds(pizza, broccoli)	holds(person, tie)
holds(person, tie)	holds(person, toothbrush)
holds(dining table, sandwich)	holds(person, cellphone)
holds(dining table, broccoli)	holds(person, baseball glove)
holds(dining table, pizza)	holds(person, remote)
…	…
holds(cell_phone, person)	holds(person, bench)
above(person, bus)	holds(person, dining table)
above(bicycle, car)	holds(person, car)

Recent Work

Learning Beyond Pixel Data (e.g. Visual Layouts, Abstract Scenes)

[Stating the Obvious: Extracting Visual Common Sense Knowledge]
[Mark Yatskar], [Vicente Ordonez], [Ali Farhadi].
North American Chapter of the Association for Computational Linguistics. **NAACL 2016**.

[Obj2Text: Generating Visually Descriptive Language from Object Layouts]
Xuwang Yin, [Vicente Ordonez].
Empirical Methods in Natural Language Processing. **EMNLP 2017**.



A woman sitting in a couch with a man holding a doughnut.

Recent Work

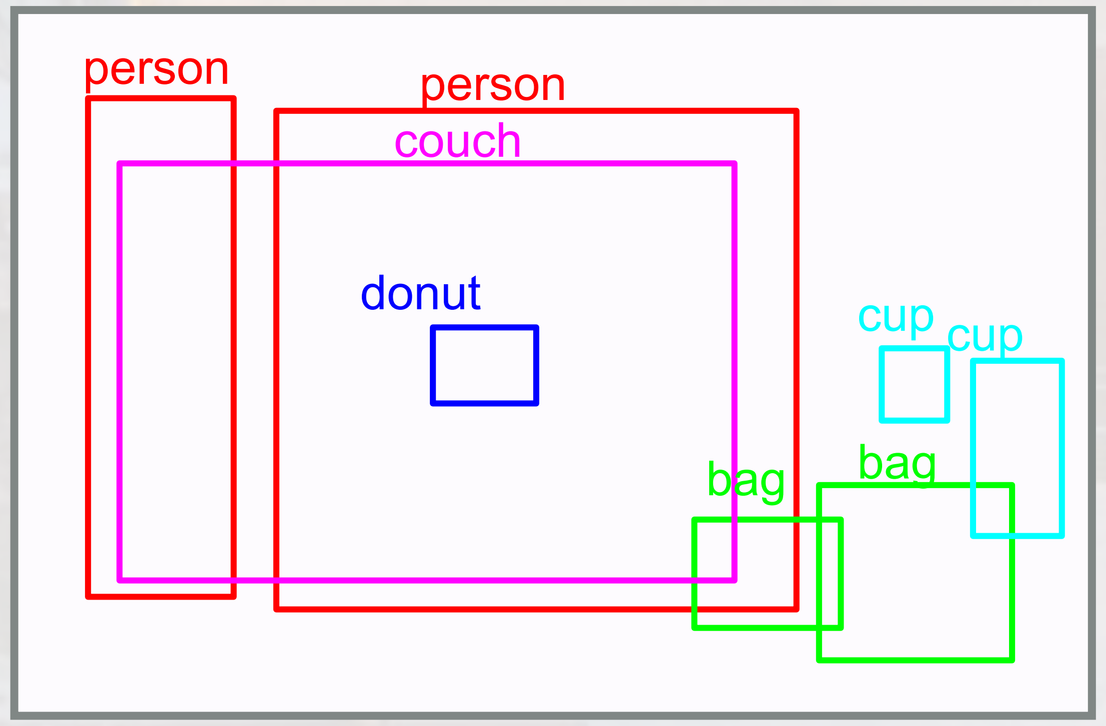
Learning Beyond Pixel Data (e.g. Visual Layouts, Abstract Scenes)

Stating the Obvious: Extracting Visual Common Sense Knowledge
Mark Yatskar, Vicente Ordonez, Ali Farhadi.
North American Chapter of the Association for Computational Linguistics. **NAACL 2016**.

Obj2Text: Generating Visually Descriptive Language from Object Layouts
Xuwang Yin, Vicente Ordonez.
Empirical Methods in Natural Language Processing. **EMNLP 2017**.

Text2Scene: Generating Abstract Scenes from Textual Descriptions
Fuwen Tan, Song Feng, Vicente Ordonez.
arXiv:1809.01110. September 2018.

A woman sitting in a couch with a man holding a doughnut.



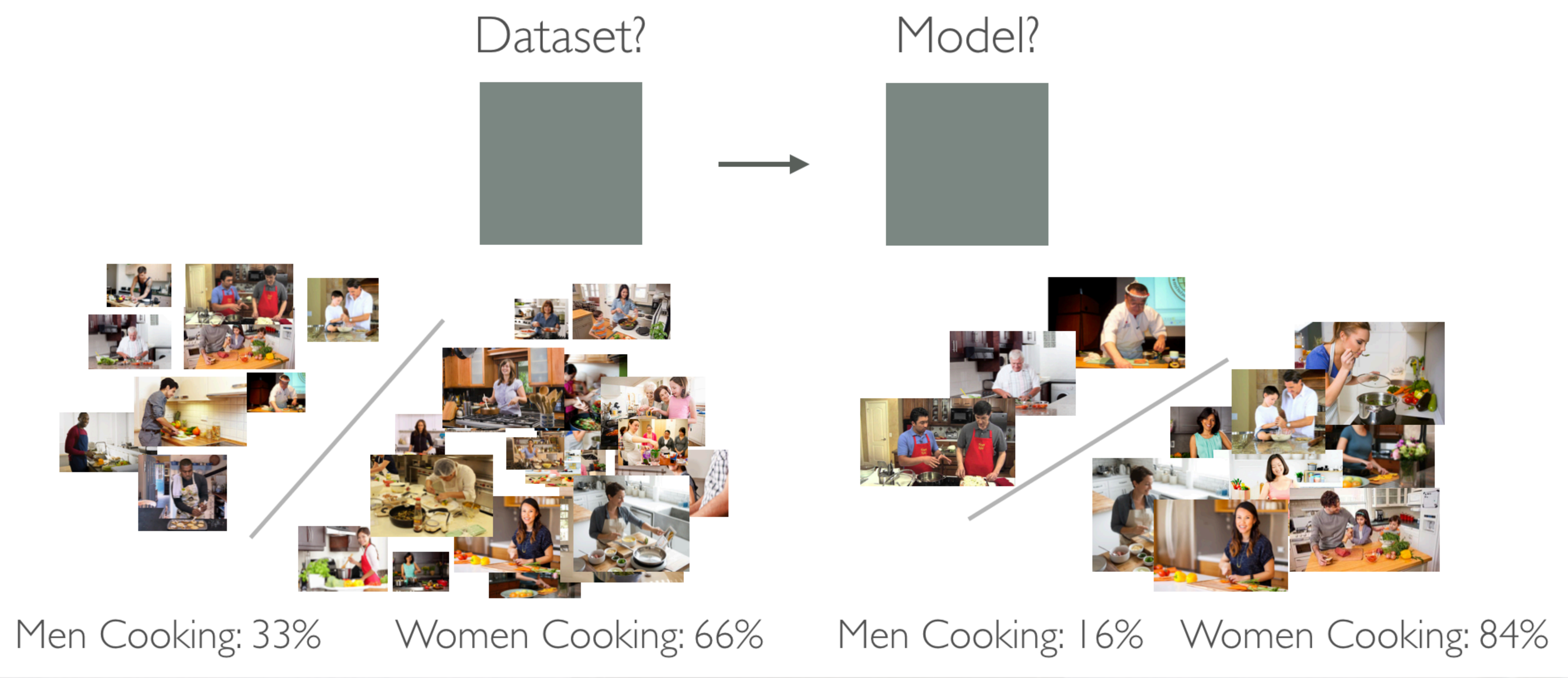
Recent Work

Fairness / Reducing and Addressing biases / Societal Stereotypes Effects in Vision and Language

Recent Work

Fairness / Reducing and Addressing biases / Societal Stereotypes Effects in Vision and Language

Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints
Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, Kai-Wei Chang.
Empirical Methods in Natural Language Processing. **EMNLP 2017**.
Best Long Paper Award!



Recent Work

Fairness / Reducing and Addressing biases / Societal Stereotypes Effects in Vision and Language

Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints
Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, Kai-Wei Chang.
Empirical Methods in Natural Language Processing. **EMNLP 2017**.
Best Long Paper Award!

Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods
Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, Kai-Wei Chang.
North American Chapter of the Association for Computational Linguistics. **NAACL 2018**. short.

[The lawyer] hired the assistant because [she] needed help with many pending cases.

The lawyer hired [the assistant] because [he] was unemployed.

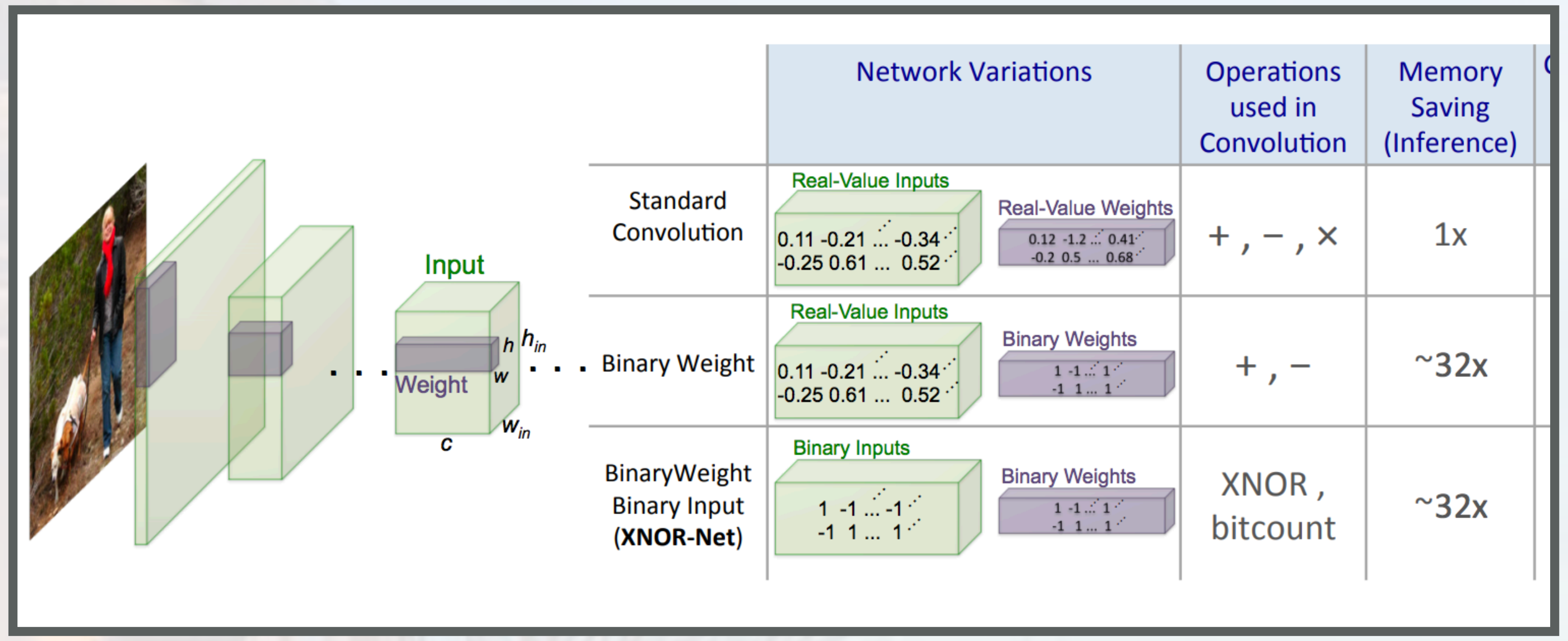
Recent Work

How to make our models Better! e.g Faster, Use Less Data or More Flexible.

Recent Work

How to make our models Better! e.g Faster, Use Less Data or More Flexible.

[XNOR-Net: ImageNet Classification Using Binary Convolutional Neural Networks](#)
Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, Ali Farhadi.
European Conference on Computer Vision. **ECCV 2016**. Amsterdam, The Netherlands. October 2016.



Recent Work

How to make our models Better! e.g Faster, Use Less Data or More Flexible.

XNOR-Net: ImageNet Classification Using Binary Convolutional Neural Networks
Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, Ali Farhadi.
European Conference on Computer Vision. **ECCV 2016**. Amsterdam, The Netherlands. October 2016.

Commonly Uncommon: Semantic Sparsity in Situation Recognition
Mark Yatskar, Vicente Ordonez, Luke Zettlemoyer, Ali Farhadi.
Intl. Conference on Computer Vision and Pattern Recognition. **CVPR 2017**.

Recent Work

How to make our models Better! e.g Faster, Use Less Data or More Flexible.

XNOR-N
Mohamm
Europea
Netherla

Commor
Mark Ya
Intl. Con

Recent Work

How to make our models Better! e.g Faster, Use Less Data or More Flexible.

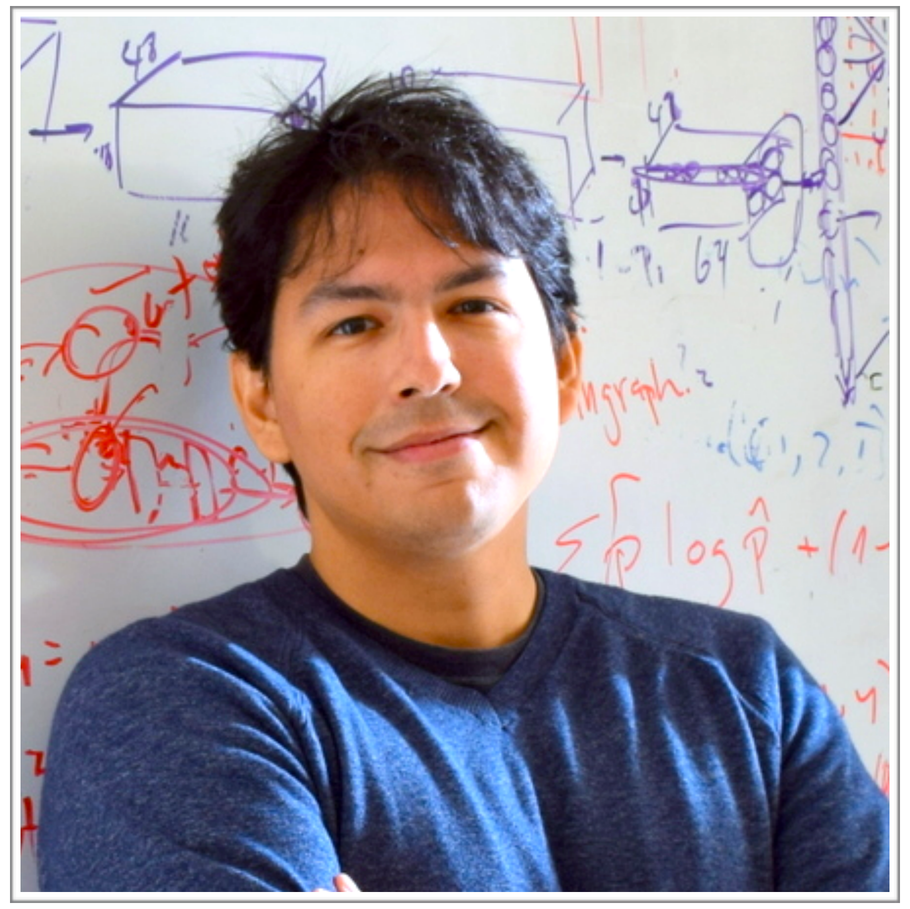
XNOR-Net: ImageNet Classification Using Binary Convolutional Neural Networks
Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, Ali Farhadi.
European Conference on Computer Vision. **ECCV 2016**. Amsterdam, The Netherlands. October 2016.

Commonly Uncommon: Semantic Sparsity in Situation Recognition
Mark Yatskar, Vicente Ordonez, Luke Zettlemoyer, Ali Farhadi.
Intl. Conference on Computer Vision and Pattern Recognition. **CVPR 2017**.

Feedback-prop: Convolutional Neural Network Inference under Partial Evidence
Tianlu Wang, Kota Yamaguchi, Vicente Ordonez.
Intl. Conference on Computer Vision and Pattern Recognition. **CVPR 2018**.

Feedback-prop: Convolutional Neural Network Inference with Partial Evidence.
CVPR 2018





Tianlu Wang Kota Yamaguchi Vicente Ordonez

How do we model problems? Case in point: Vision and Language

P( , Two people playing with a kite on the beach)

If we had access to this:

P( , Two people playing with a kite on the beach)

A few things we might be able to do (in principle) by marginalizing variables:

P(Two people playing with a kite on the beach / )

Image Captioning

P( / Two people playing with a kite on the beach)

Image Retrieval (discrete)

Image Synthesis (continuous)

P( /  , The person holding the kite)

Referring Expressions

Let's take multilingual image captioning

P(Two people playing with a kite on the beach , Dos personas jugando en la playa con una cometa. / )

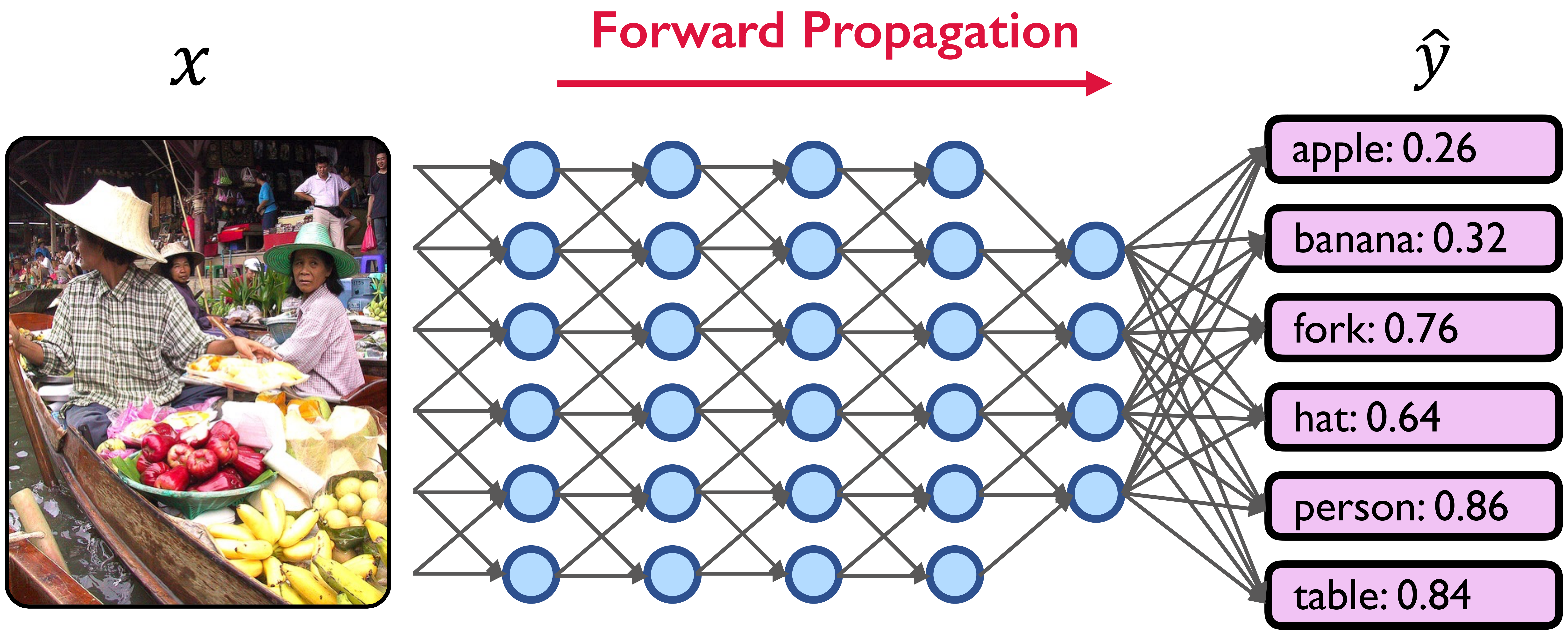
Then we can marginalize and only need one model to do translation both ways!

P(Two people playing with a kite on the beach / Dos personas jugando en la playa con una cometa. , )

P(Dos personas jugando en la playa con una cometa. / Two people playing with a kite on the beach , )

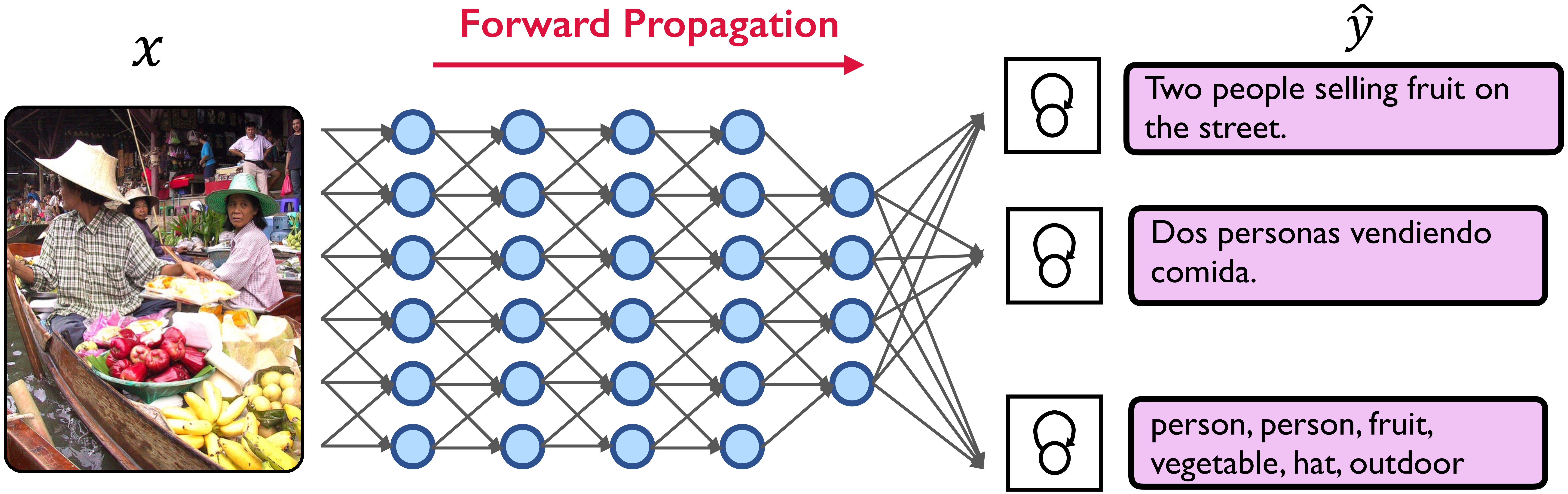
Neural Networks: Rigid Model

- [In most cases] once a model its trained, **input** and **output** variables are **fixed.**



Neural Networks: Rigid Model

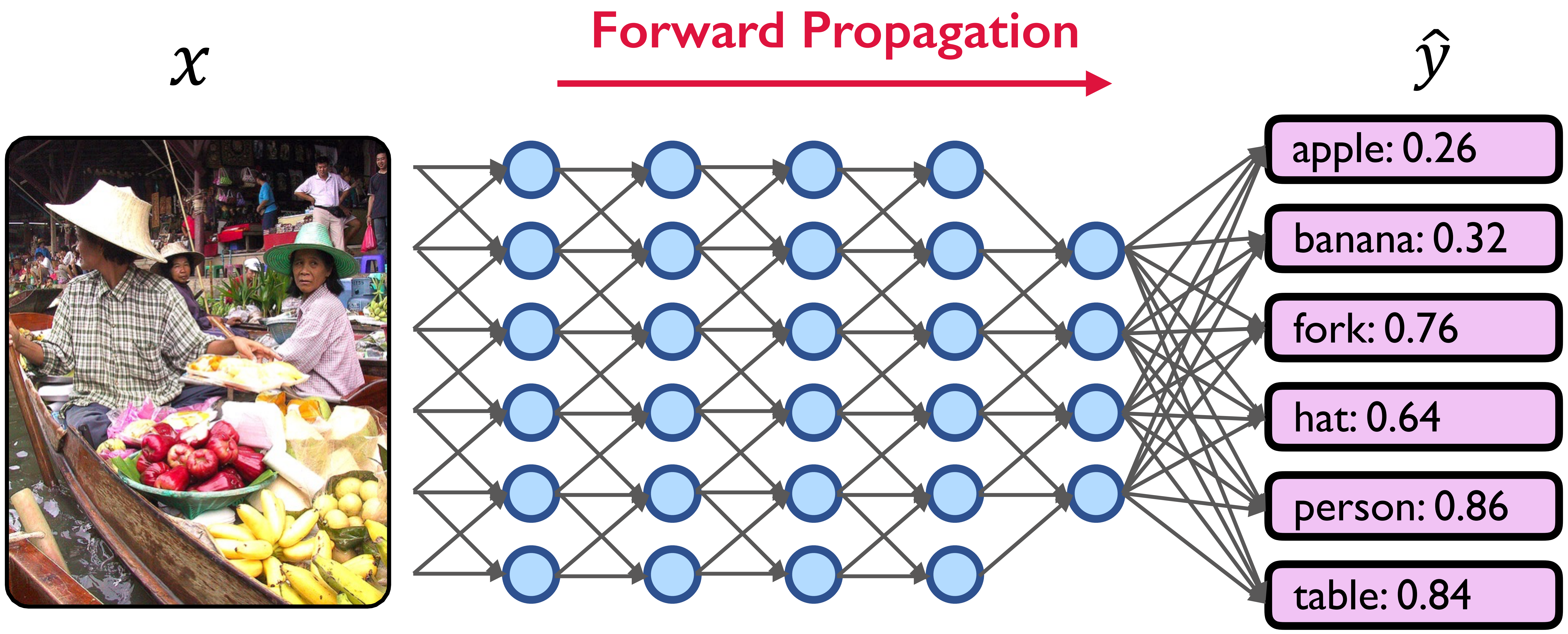
- [In most cases] once a model its trained, **input** and **output** variables are **fixed.**



individual outputs can be complex and structured

Neural Networks: Rigid Model

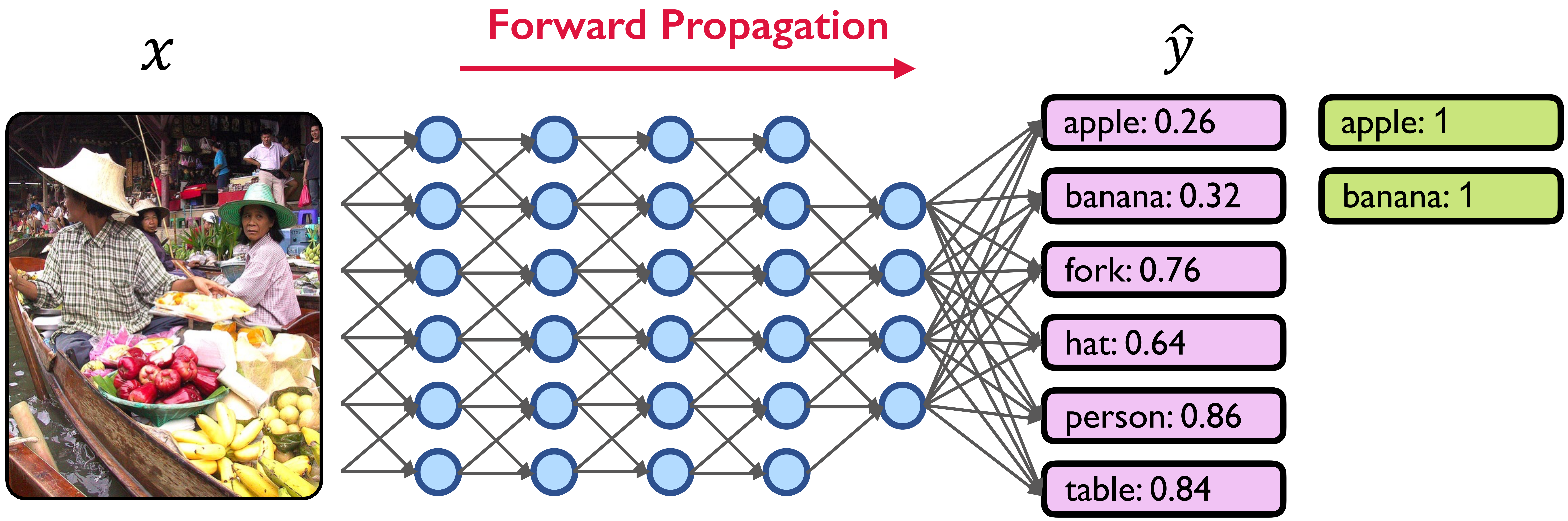
- [In most cases] once a model its trained, **input** and **output** variables are **fixed.**



But we will use this as our running example for simplicity

Neural Networks: Rigid Model

- [In most cases] once a model its trained, **input** and **output** variables are **fixed.**



What happens if we know the image has an apple and a banana?
How do we leverage that extra information?

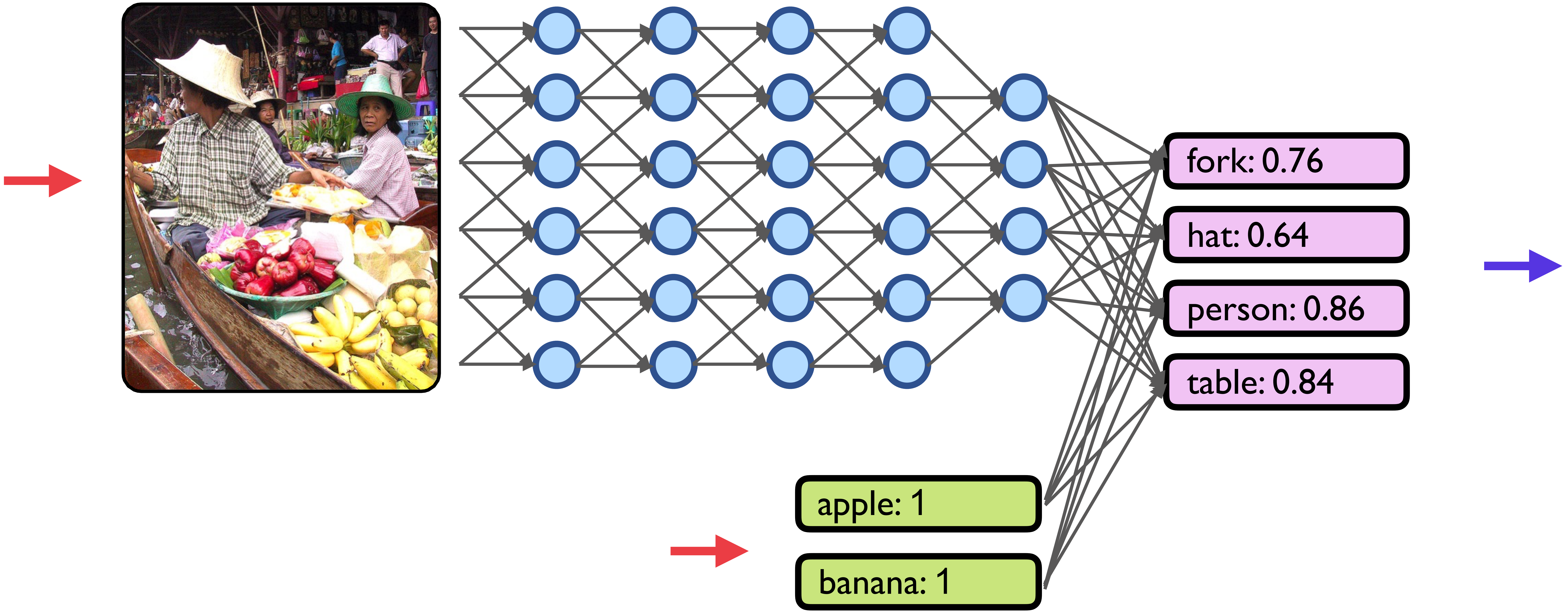
Neural Networks: Rigid Model

- [In most cases] once a model its trained, **input** and **output** variables are **fixed.**

We have: P(apple banana hat fork person table / )

But we need: P(fork hat person table /  , apple banana)

A simple (naive?) solution



But breaks if for a given example we are given other subset of labels as known.

Feedback-prop: Convolutional Neural Network Inference with Partial Evidence.

Tianlu Wang, Kota Yamaguchi, Vicente Ordonez. **CVPR 2018**

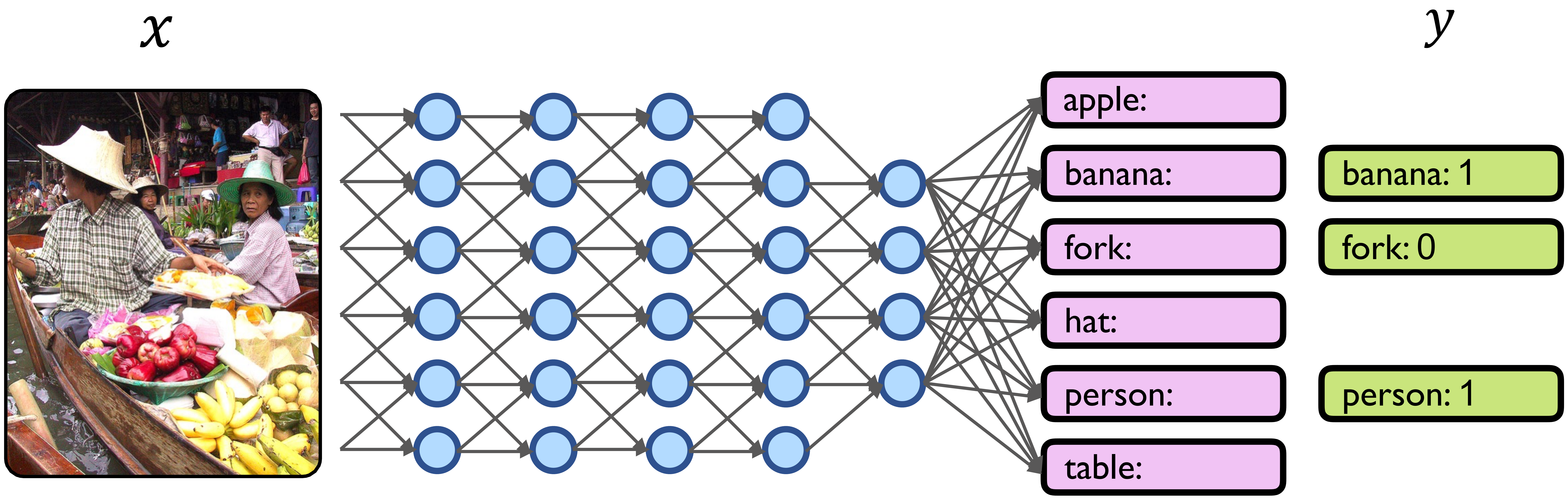
Main Contribution

Train this: P(apple banana fork hat person table / )

But be able to execute this at test time: (for any arbitrary subset of known variables)

P(fork hat person table /  , apple banana)

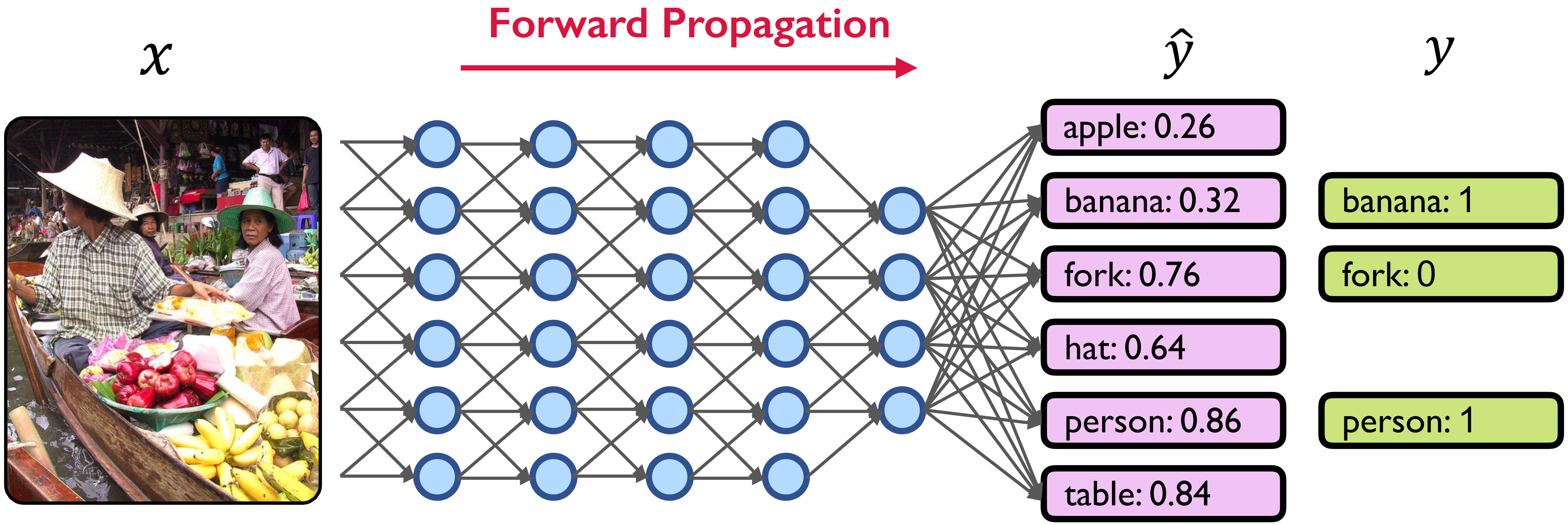
Feedback-prop: Convolutional Neural Network Inference with Partial Evidence.
CVPR 2018



Initial Condition:

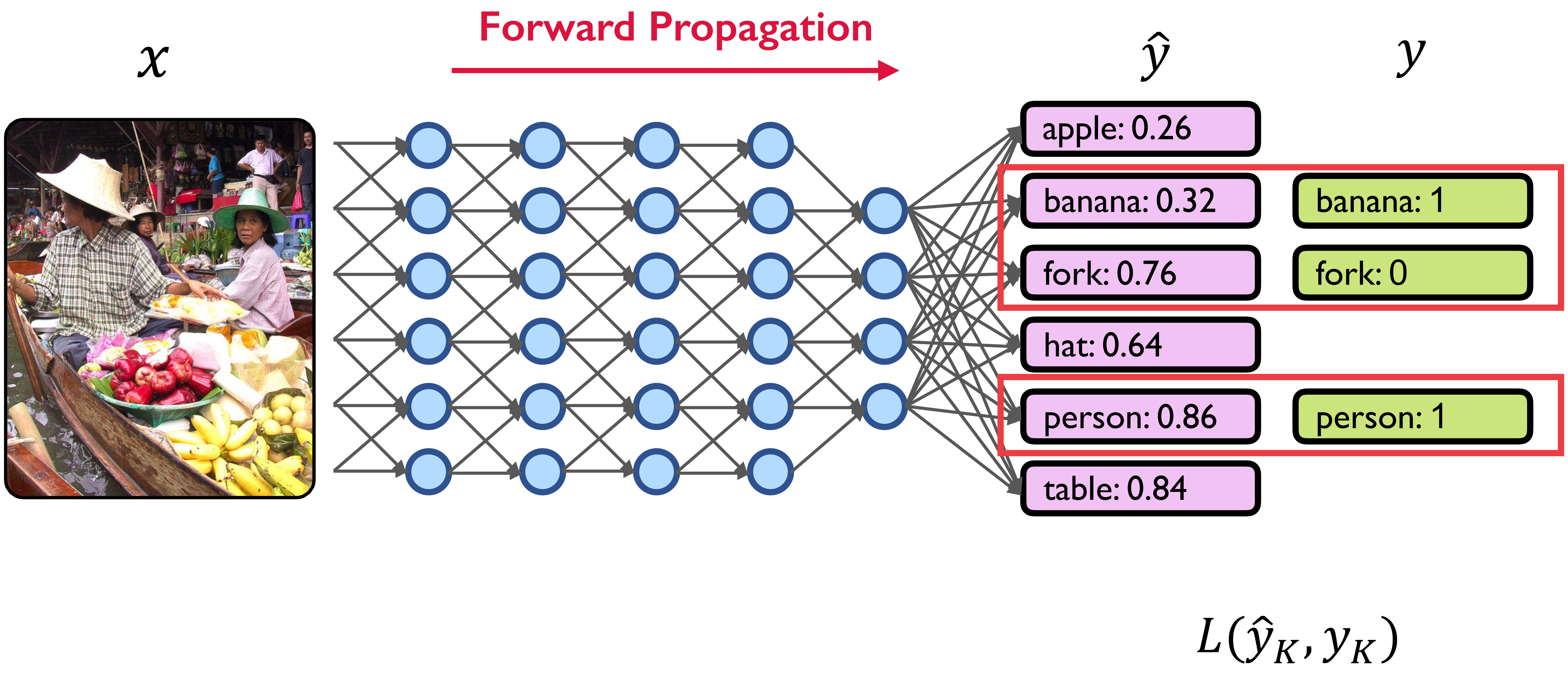
* Multi-task network trained.
* Input Image + Input partial evidence.

Feedback-prop: Convolutional Neural Network Inference with Partial Evidence. CVPR 2018



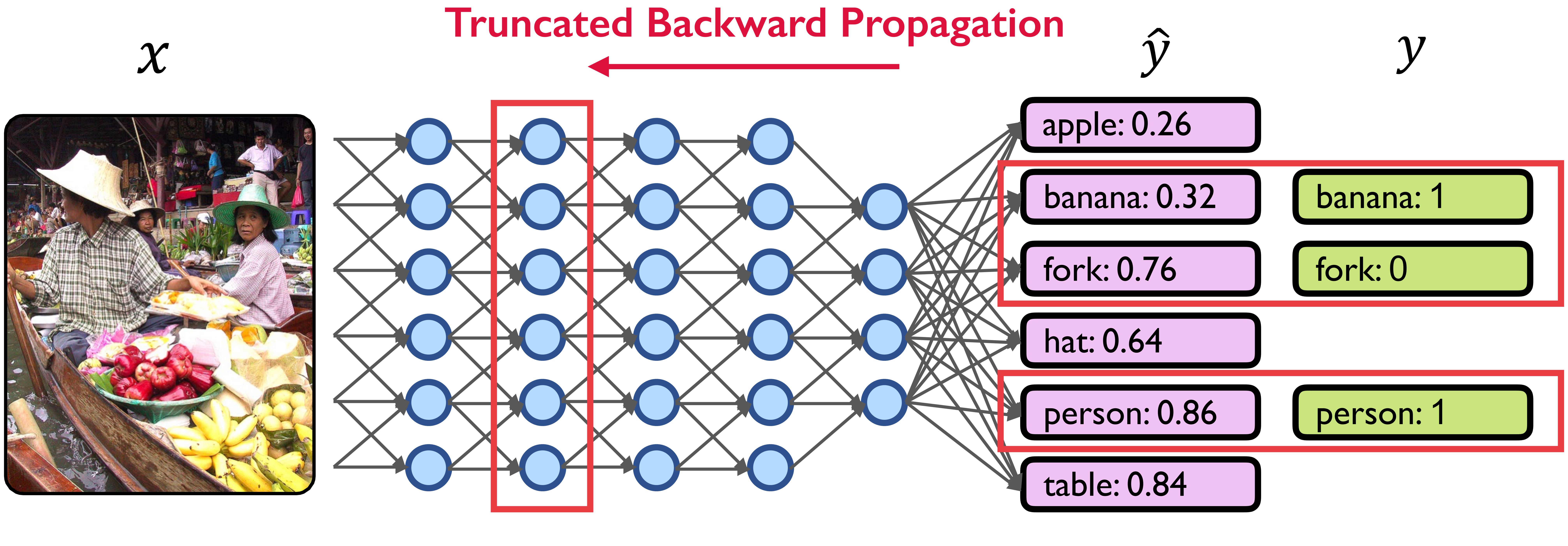
Step 1: Forward-propagate and estimate jointly the scores for all variables.

Feedback-prop: Convolutional Neural Network Inference with Partial Evidence. CVPR 2018



Step 2: Compute partial loss between known labels and their current scores.

Feedback-prop: Convolutional Neural Network Inference with Partial Evidence.
CVPR 2018



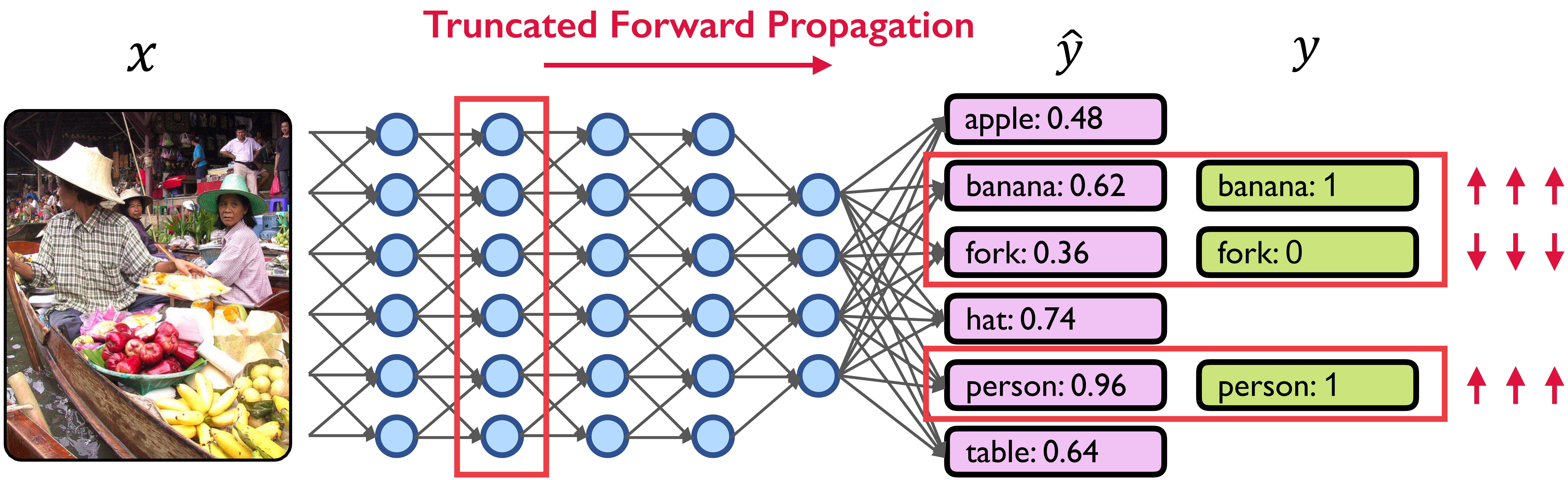
Pivot variable: a_l

Pivot variable update: $a_l = a_l - \lambda \, dL/da_l$

$L(\hat{y}_K, y_K)$

Step 3: Update a pivoting intermediate representation so that the partial loss is minimized.

Feedback-prop: Convolutional Neural Network Inference with Partial Evidence. CVPR 2018

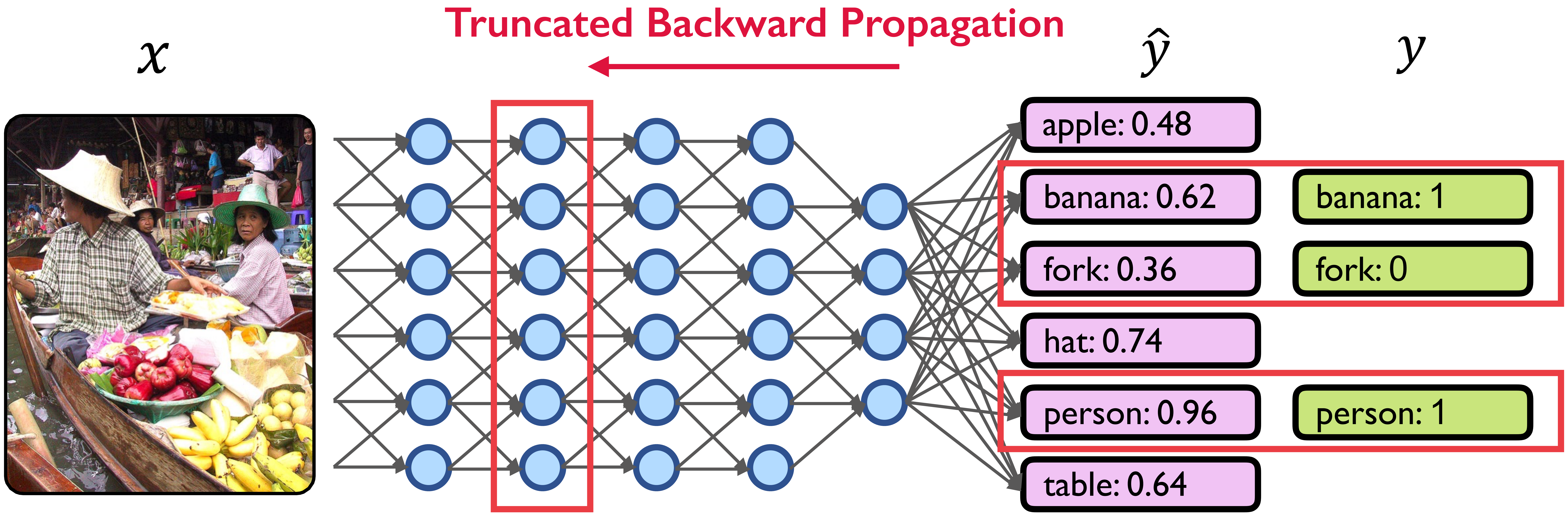


Updated pivot variable: a_l

$L(\hat{y}_K, y_K)$

Step 4: Forward-propagate with updated pivoting variable and recompute partial loss.

Feedback-prop: Convolutional Neural Network Inference with Partial Evidence. CVPR 2018



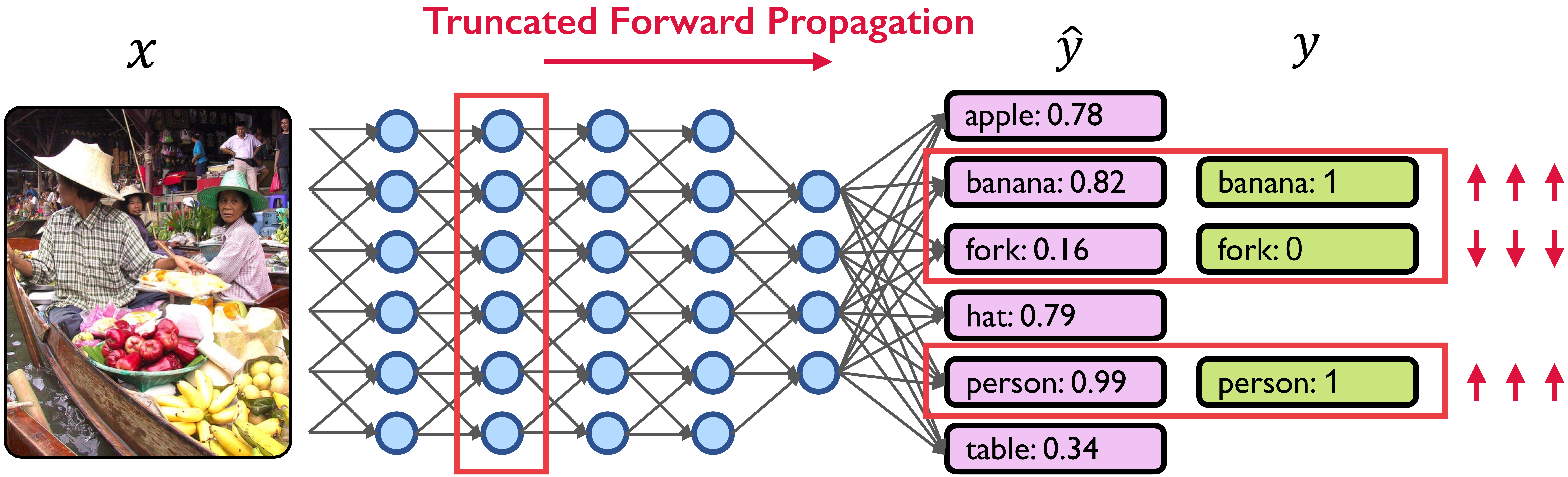
Pivot variable: a_l

Pivot variable update: $a_l = a_l - \lambda\, dL/da_l$

$L(\hat{y}_K, y_K)$

Step 3: Update a pivoting intermediate representation so that the partial loss is minimized.

Feedback-prop: Convolutional Neural Network Inference with Partial Evidence. CVPR 2018

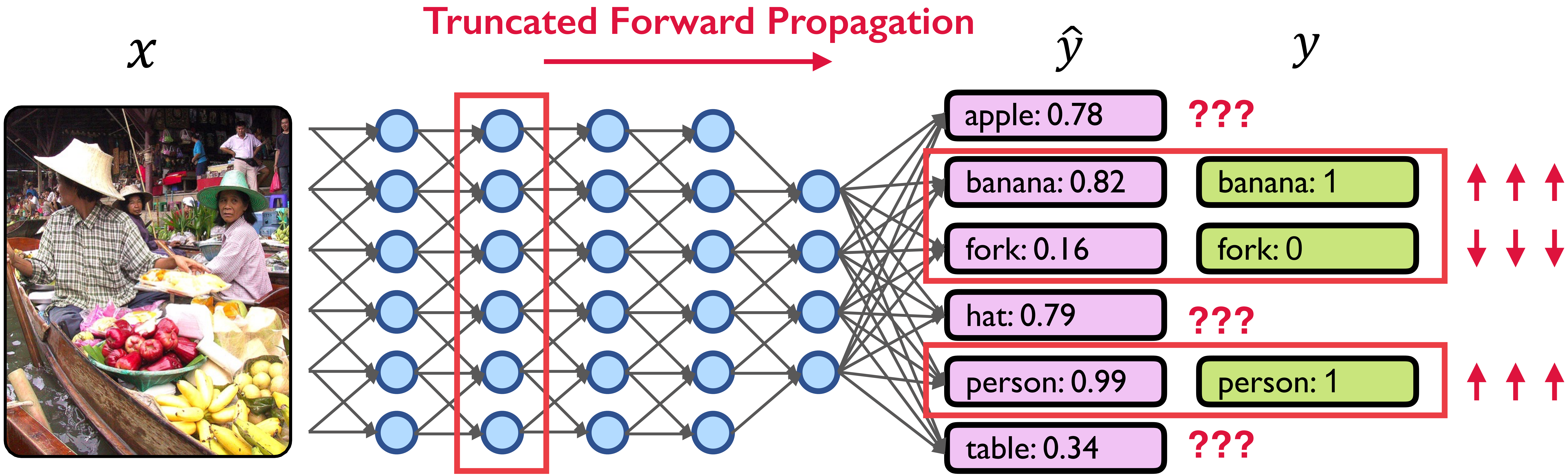


Updated pivot variable: a_l

$L(\hat{y}_K, y_K)$

Step 4: Forward-propagate with updated pivoting variable and recompute partial loss. Repeat until stopping criteria

Feedback-prop: Convolutional Neural Network Inference with Partial Evidence. CVPR 2018



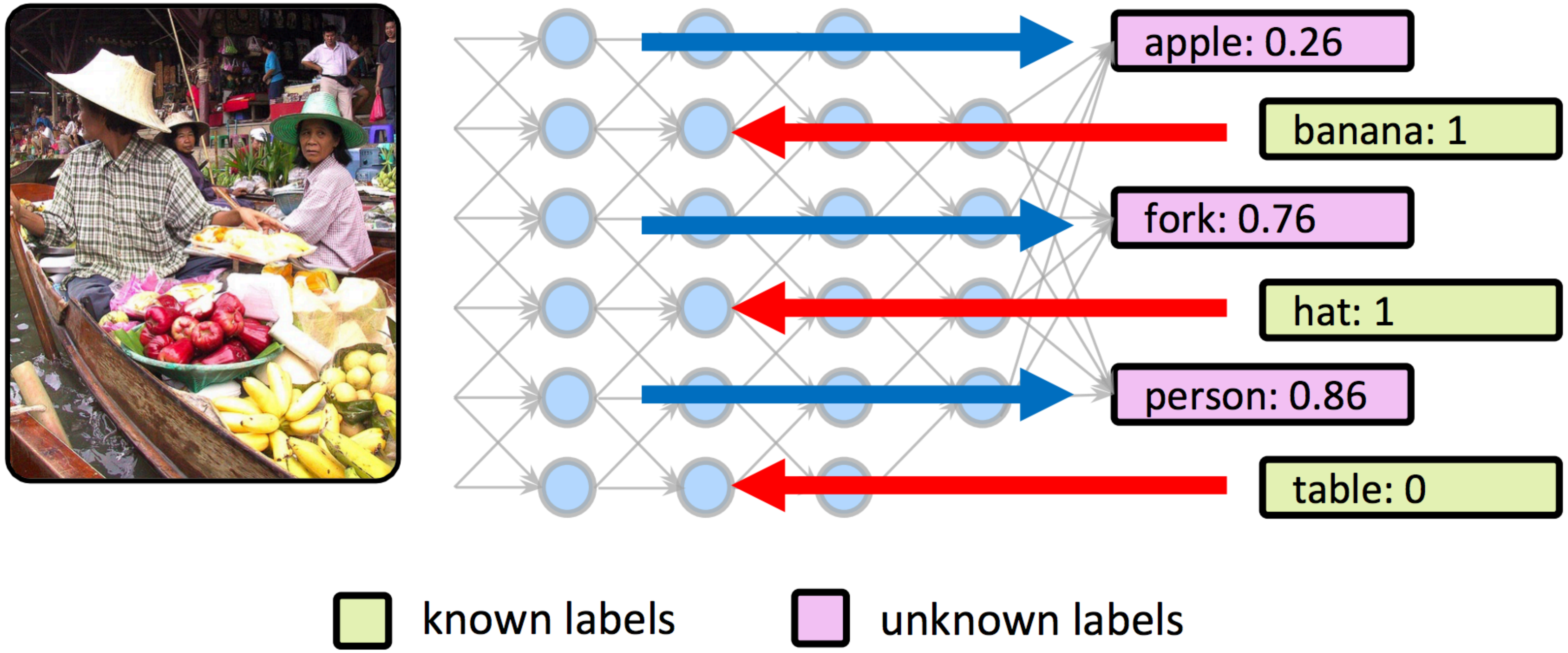
Updated pivot variable: a_l

$L(\hat{y}_K, y_K)$

It is clear the effect of the pivoting variable on the known labels

But what is the effect on the unknown labels? Do they improve?

Feedback-prop: Convolutional Neural Network Inference with Partial Evidence. **CVPR 2018**

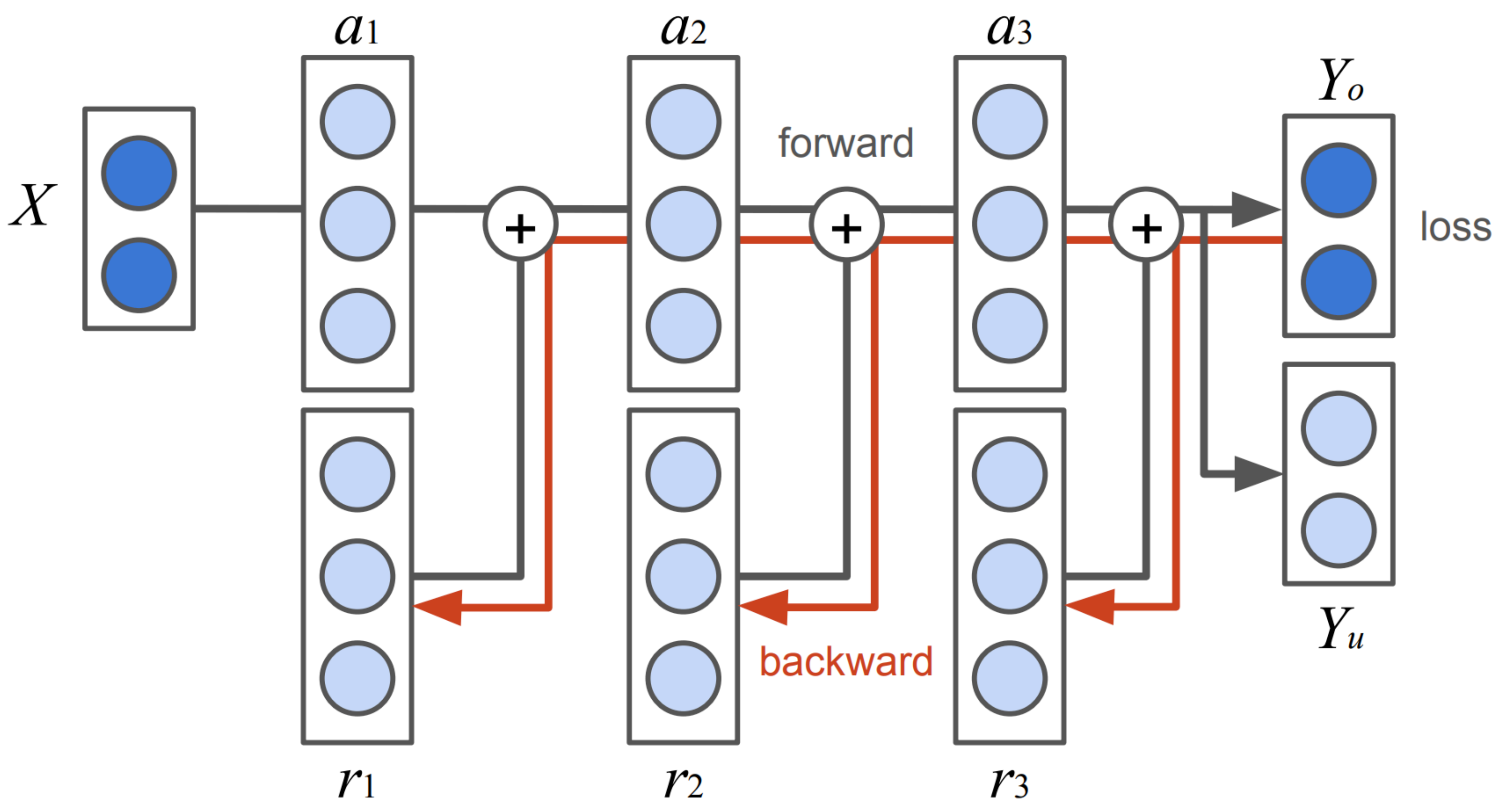


Answer: Their accuracy improves! This just works!

Feedback-prop: Convolutional Neural Network Inference with Partial Evidence. **CVPR 2018**

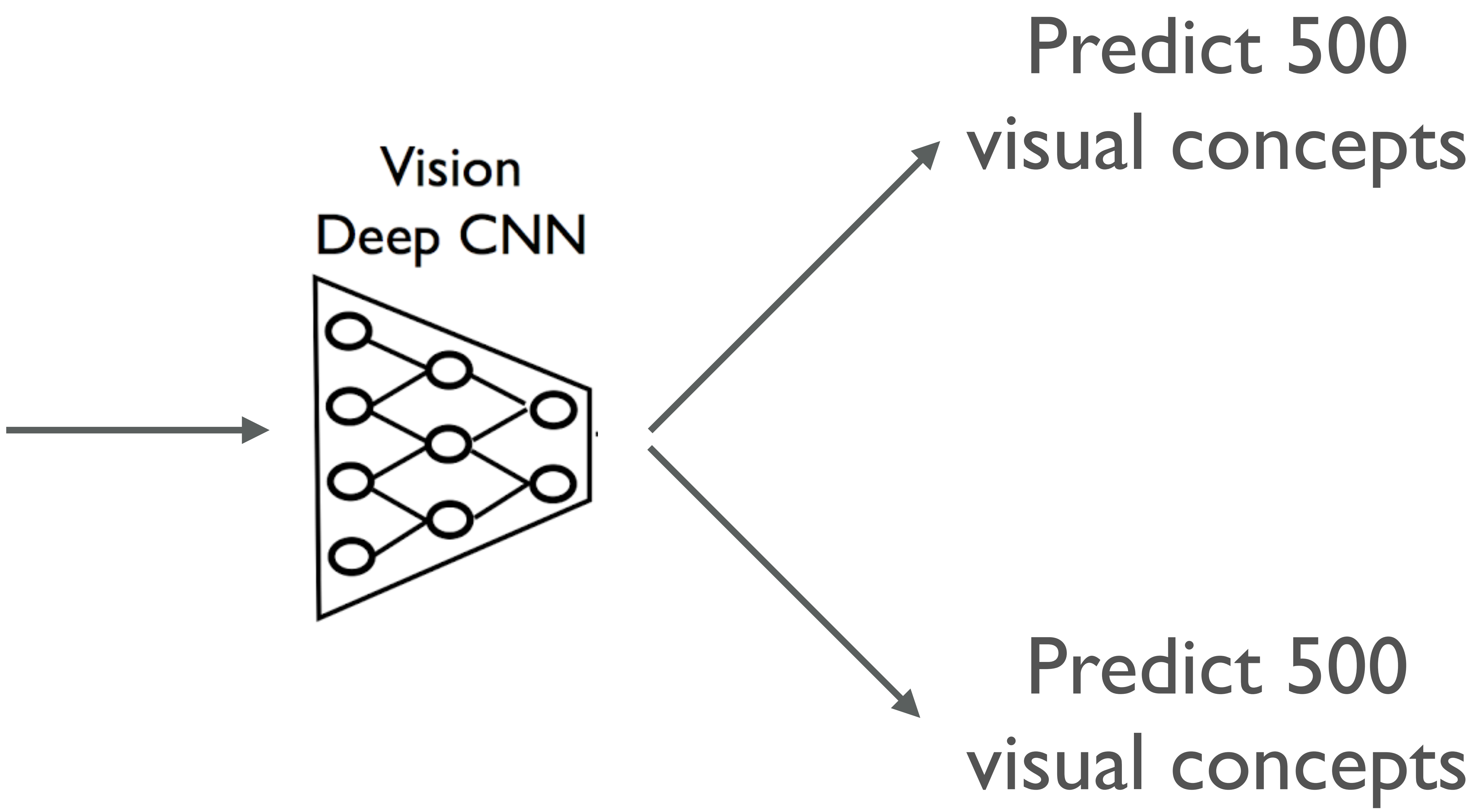
We also propose more technical contributions with the same underlying idea:

How to avoid committing to a specific intermediate pivot representation



Refer to the paper for more details. Available on arxiv for now.
https://arxiv.org/abs/1710.08049

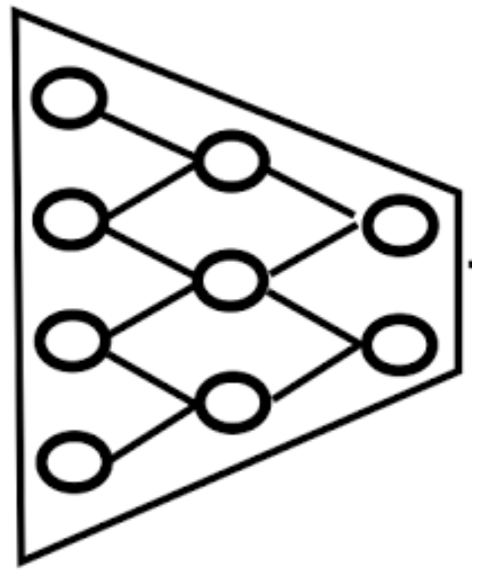
Task 1: Multi-label Prediction between sets of non-overlapping concepts



Train model to predict two non-overlapping sets of visual concepts.

Task 1: Multi-label Prediction between sets of non-overlapping concepts

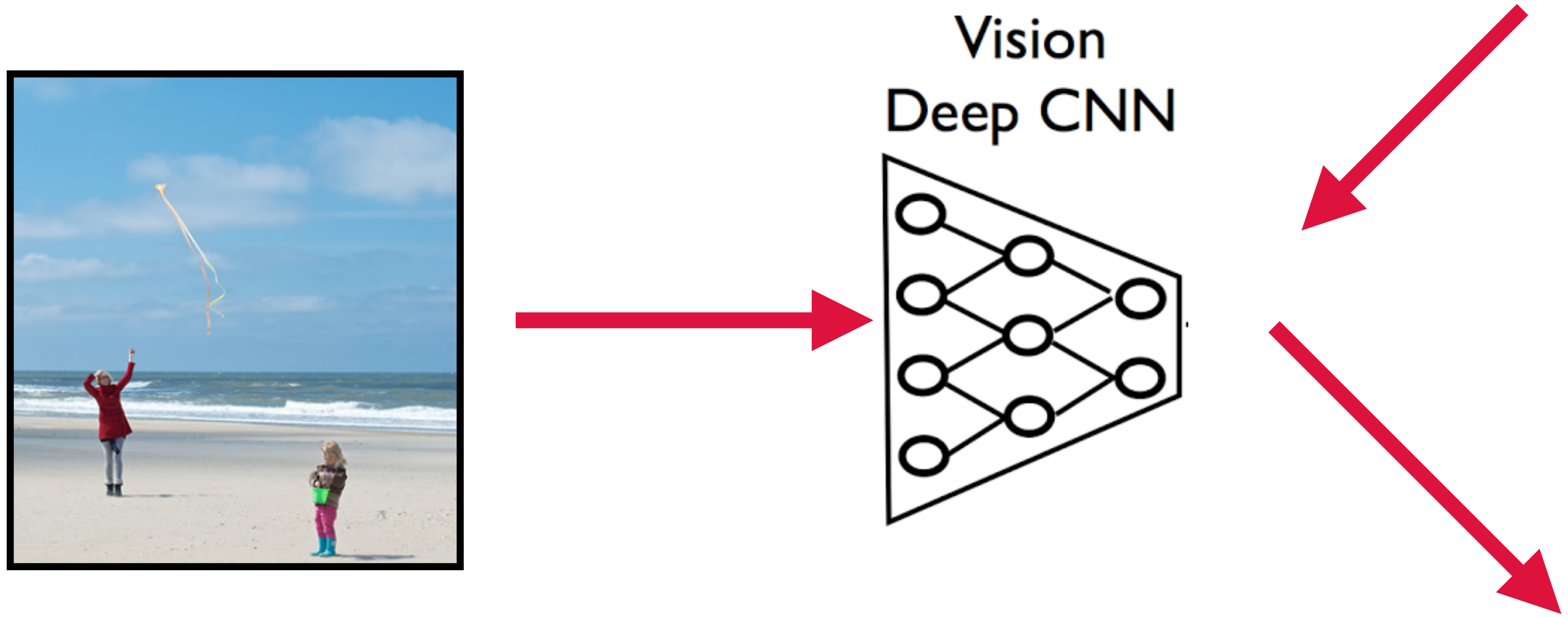
Vision
Deep CNN



Predict 500
visual concepts

Evaluate on one of the sets: **meanAP: ~27%**

Task 1: Multi-label Prediction between sets of non-overlapping concepts



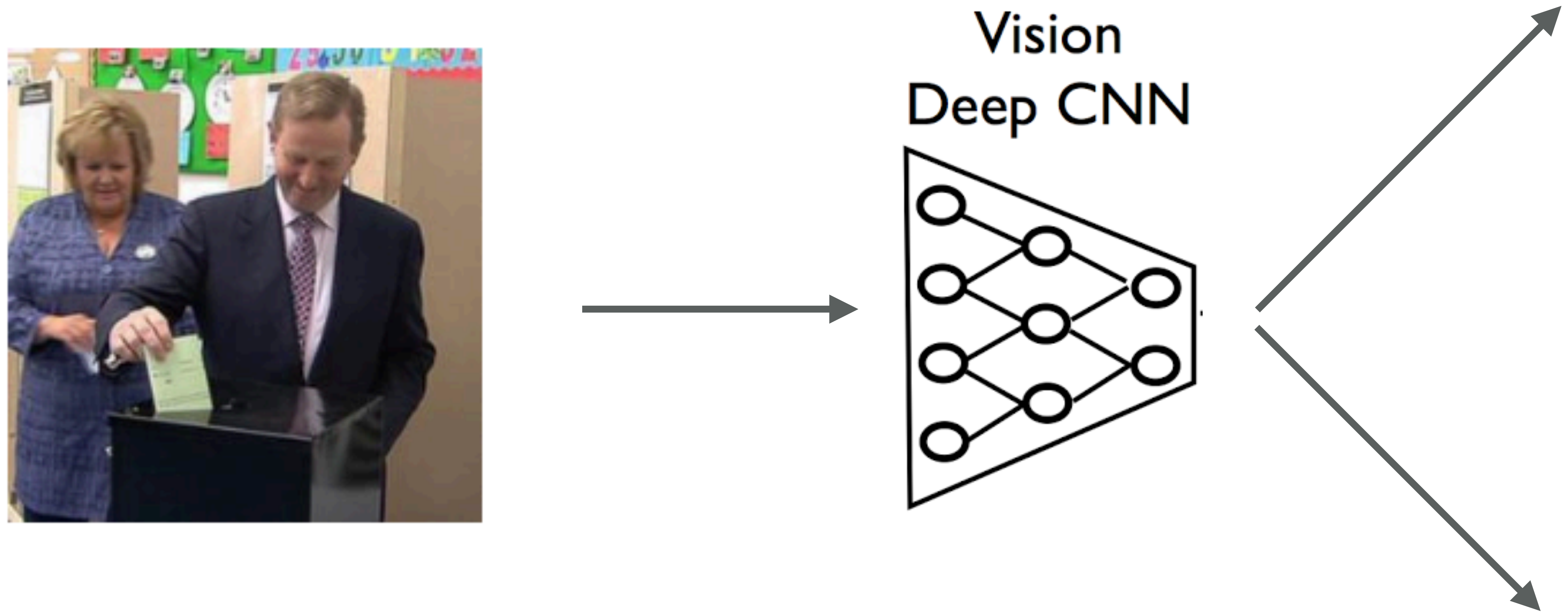
Predict 500 visual concepts

Predict 500 visual concepts

Evaluate on one of the sets: **meanAP: ~27%** **meanAP: ~29.5%**

* Averaged across many possible set selections.

Task II: Images + News Text + News Captions



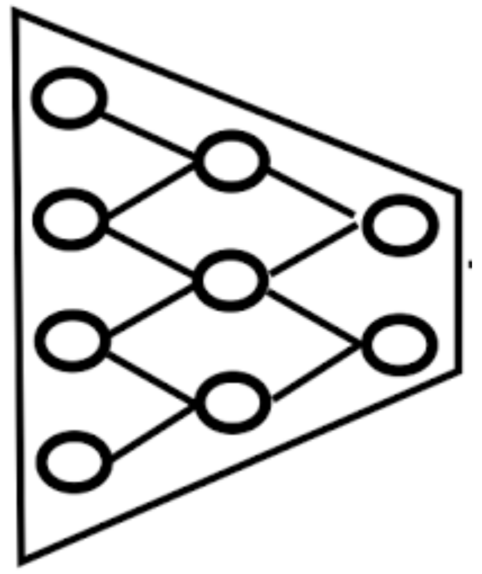
Words from news text

Words from news captions

Train model to predict words from news captions + words from news articles with non-overlapping vocabularies.

Task II: Images + News Text + News Captions

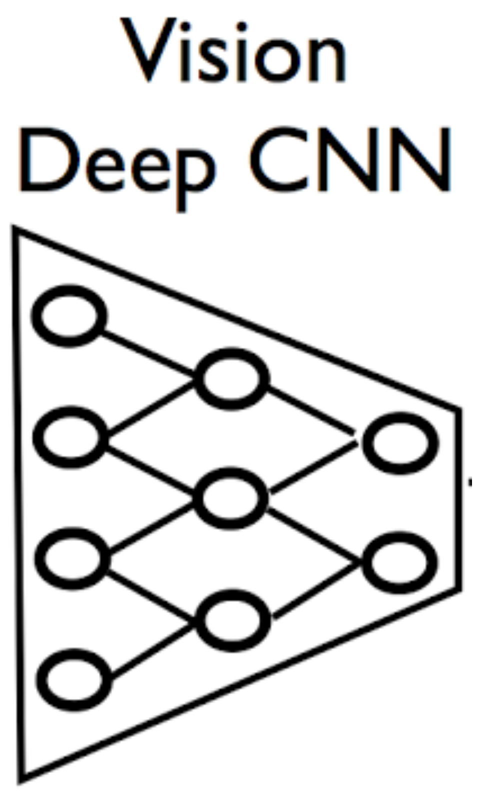
Vision
Deep CNN



Words from news captions

Evaluate on caption words: **meanAP: ~19.92%**

Task II: Images + News Text + News Captions

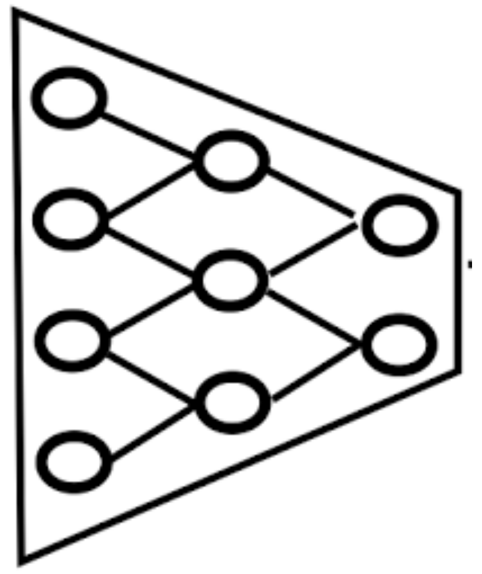


Words from news text

Words from news captions

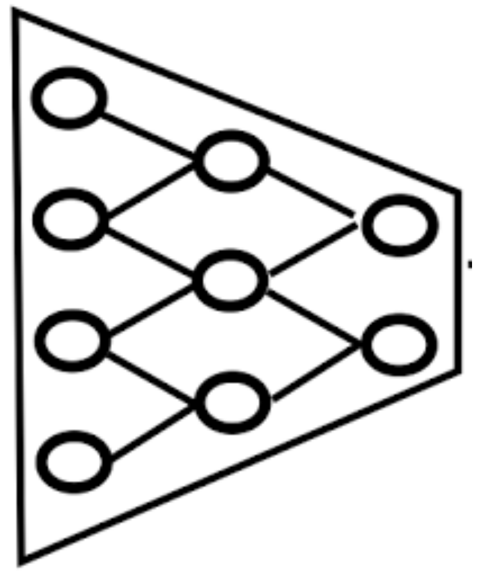
Evaluate on caption words: **meanAP: 19.92%** **meanAP: 22.57%**

Vision
Deep CNN



official:0.790290
home:0.310297
child:0.180287
people:0.139492
woman:0.088490
house:0.076746
camp:0.064999
use:0.063372

Vision
Deep CNN



action, start, fund, price, move, technology, syria, thousand, name, risk, offer, hope, saw, food, face, education, girl, act, crime, course, violence, crisis, book, age, return, france, organisation, space, access, try, hundred, provide, …

camp:0.925969
refugee:0.908903
home:0.293703
child:0.255574
woman:0.147657
people:0.104480
syria:0.088542
official:0.061292

no feedback-prop predictions:

claim:0.891679	school:0.060947	try:0.319411	official:0.790290	ceremony:0.506596	people:0.494557
try:0.592581	people:0.054434	show:0.186112	home:0.310297	thousand:0.159579	light:0.325617
attack:0.278426	light:0.050388	scene:0.158961	child:0.180287	pay:0.132895	launch:0.279506
city:0.155168	part:0.045863	news:0.110425	people:0.139492	game:0.104834	sir:0.270729
hundred:0.133139	force:0.043337	people:0.092683	woman:0.088490	deal:0.080287	point:0.243272
woman:0.120313	area:0.042076	attack:0.059946	house:0.076746	people:0.071572	leave:0.150900
police:0.119733	include:0.042012	pay:0.050996	**camp:0.064999**	open:0.048961	centre:0.133657

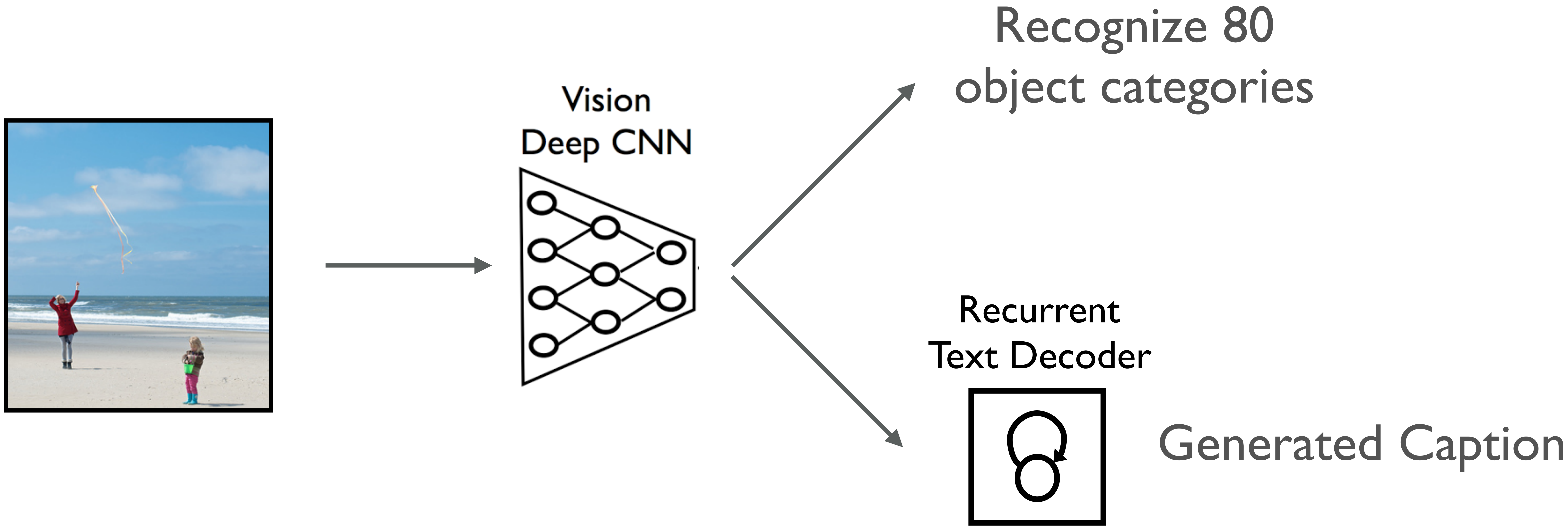
news text labels:

people, government, tell, police, country, state, group, report, find, place, school, public, news, attack, force, want, official, mean, support, death, security, put, use,	country, work, part, party, minister, report, number, school, leader, news, meet, house, force, court, power, want, official, end, council, support, election, death,	people, government, tell, police, country, part, family, child, party, group, report, company, president, need, leader, public, news, business, house, help, force, court, case,	action, start, fund, price, move, technology, syria, thousand, name, risk, offer, hope, saw, food, face, education, girl, act, crime, course, violence, crisis, book,	union, today, secretary, offer, speak, key, executive, education, parent, development, stop, radio, energy, visit, mile, everyone, space, stage, club, opportunity, trust,	prime, start, statement, mark, station, act, person, age, return, ireland, morning, provide, island, couple, poll, candidate, referendum, amount, ask, voter, protect,

with feedback-prop predictions:

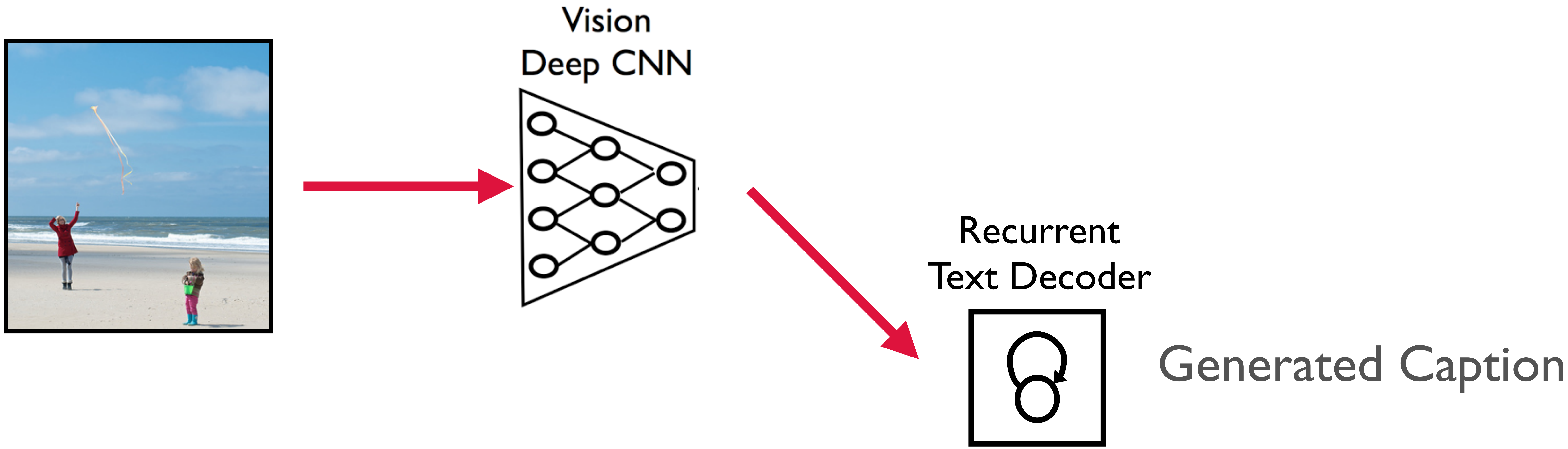
claim:0.913860	**clash:0.948569**	try:0.385340	**camp:0.925969**	**school:0.858543**	**vote:0.488819**
attack:0.910921	protester:0.774579	protest:0.260692	**refugee:0.908903**	game:0.284368	campaign:0.447369
bomb:0.267836	pro:0.520027	medium:0.130189	home:0.293703	play:0.234772	people:0.388327
try:0.240699	security:0.405497	china:0.119549	child:0.255574	thousand:0.112460	centre:0.309245
body:0.159527	force:0.176731	**court:0.100340**	woman:0.147657	parent:0.085781	ireland:0.271122
woman:0.123605	**police:0.159598**	show:0.086785	people:0.104480	people:0.076458	leave:0.263814
relative:0.121821	anti:0.122141	**police:0.069903**	syria:0.088542	start:0.061948	point:0.179191

Task III: Image Captioning + Object Categorization



Train model to predict captions and objects in the image.

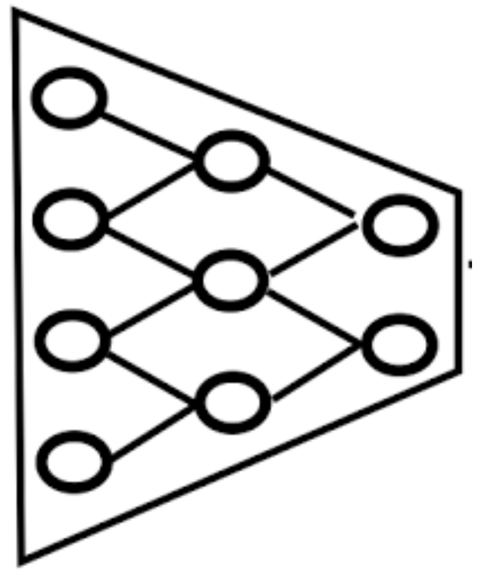
Task III: Image Captioning + Object Categorization



Evaluation on captions: **CIDEr: 94.6**

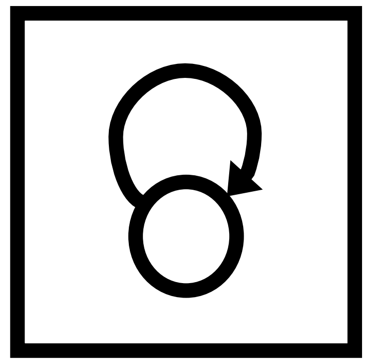
Task III: Image Captioning + Object Categorization

Vision
Deep CNN



Recognize 80
object categories

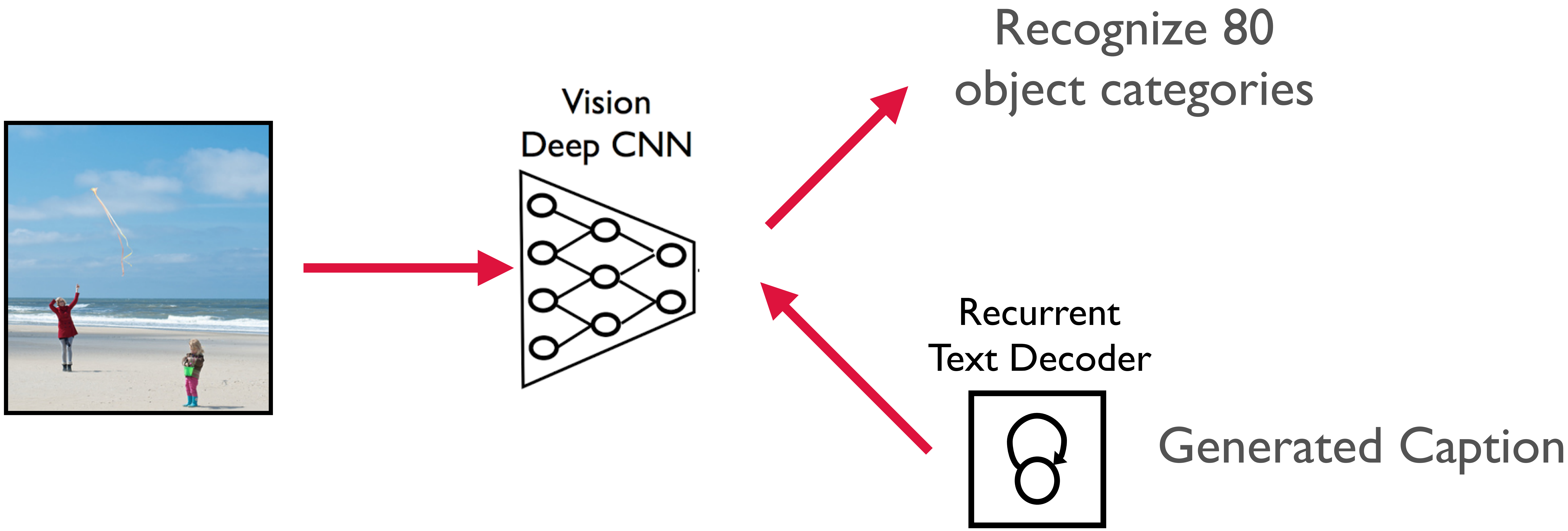
Recurrent
Text Decoder



Generated Caption

CIDEr: 99.2

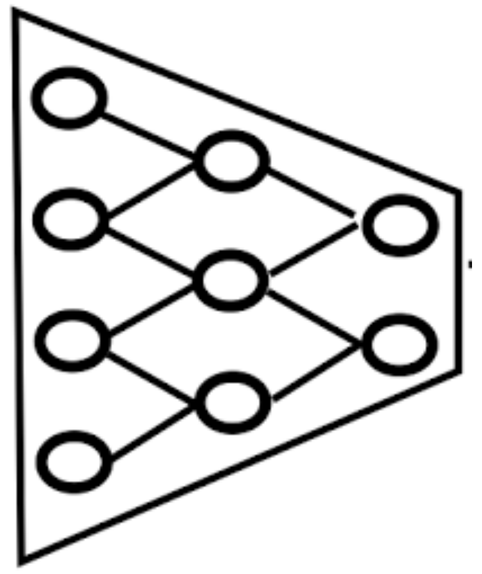
Task III: Image Captioning + Object Categorization



We did not try this but should also work!

Task IV: Scene Attributes + Scene Coarse Labels: SUN Dataset

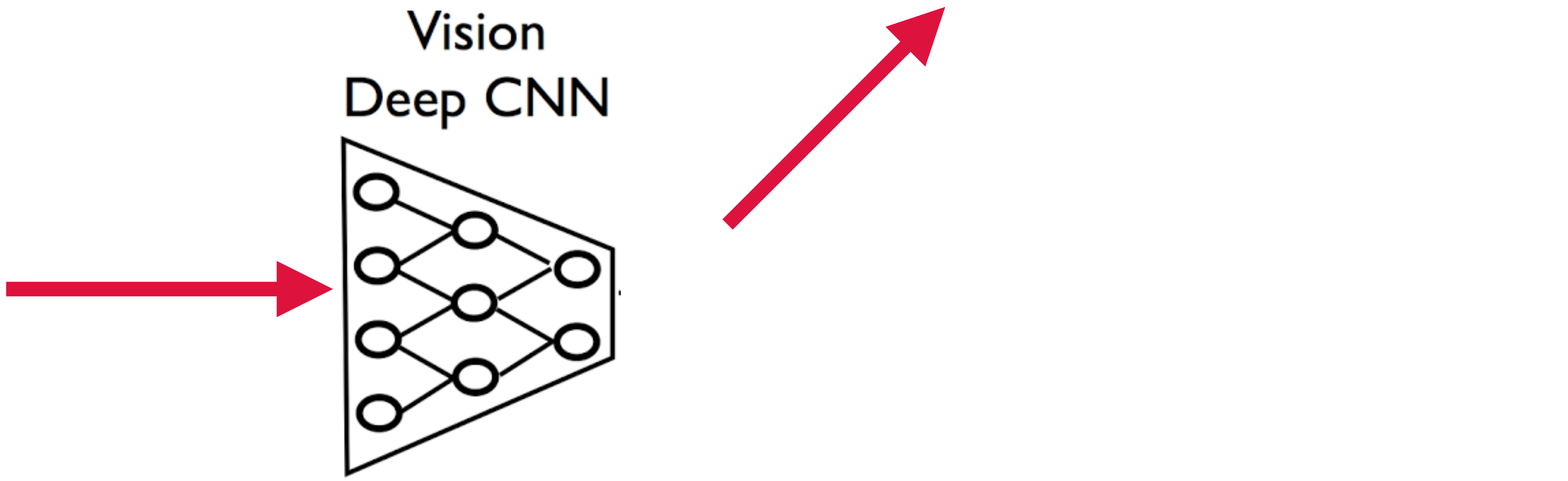
Vision
Deep CNN



Scene Attributes

Coarse Scene Labels

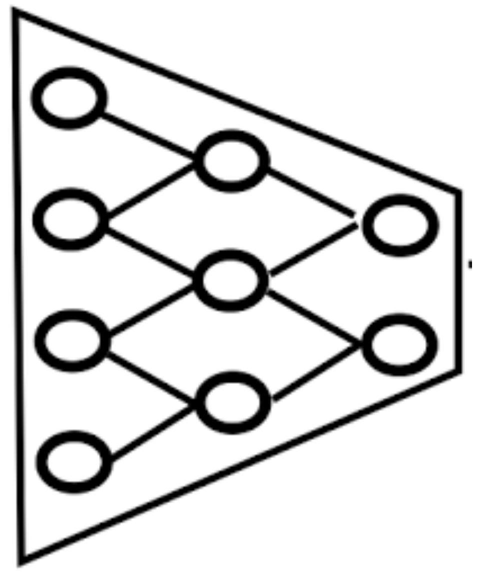
Task IV: Scene Attributes + Scene Coarse Labels: SUN Dataset



Evaluate on scene attributes: **meanAP: 52.83%**

Task IV: Scene Attributes + Scene Coarse Labels: SUN Dataset

Vision
Deep CNN



Scene Attributes

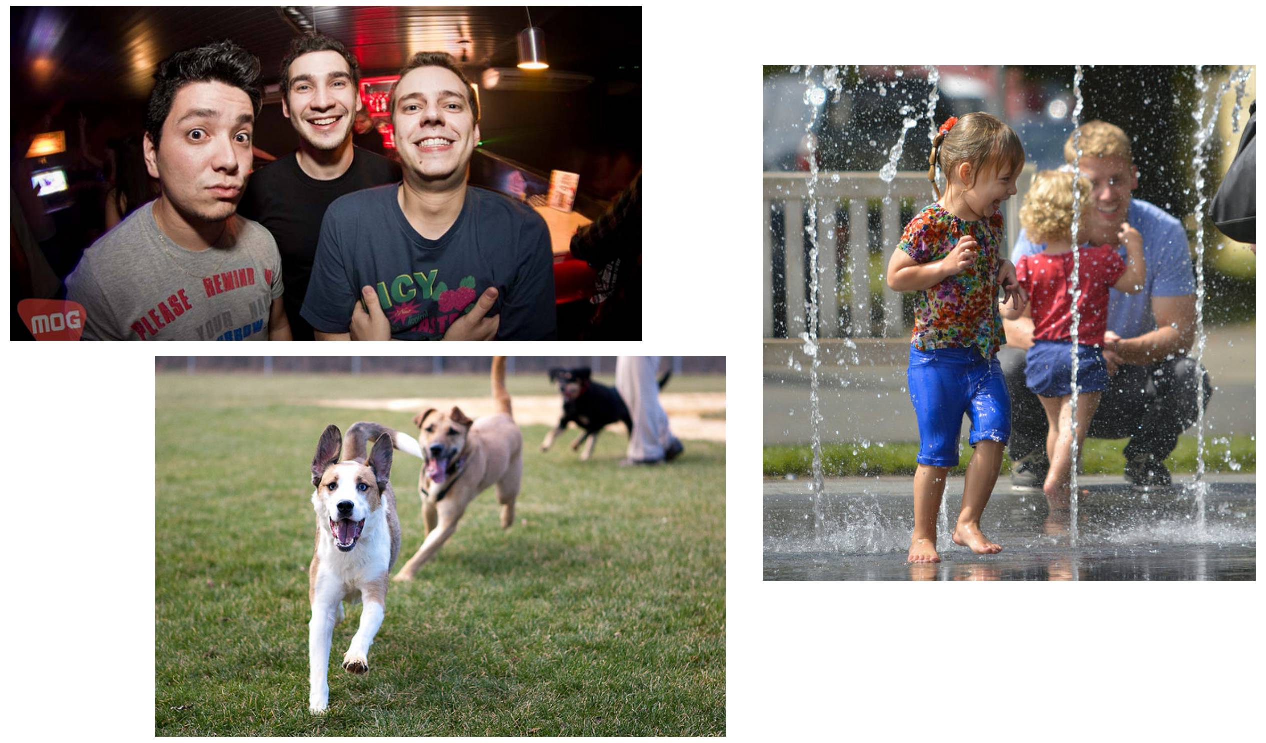
Coarse Scene Labels

Evaluate on scene attributes: **meanAP: 52.83%**

Hu et al 2016
meanAP: 58.45%

meanAP: 58.70%

Very Practical: Images don’t exist in a vacuum



“They seem to be having a lot of fun”

Images on social media have comments

"People pay respects to the victims”

“A man protests in the middle of the street”

“A lone Jewish settler challenges Israeli security forces”

Many other examples: geo-location, uploader information, context.

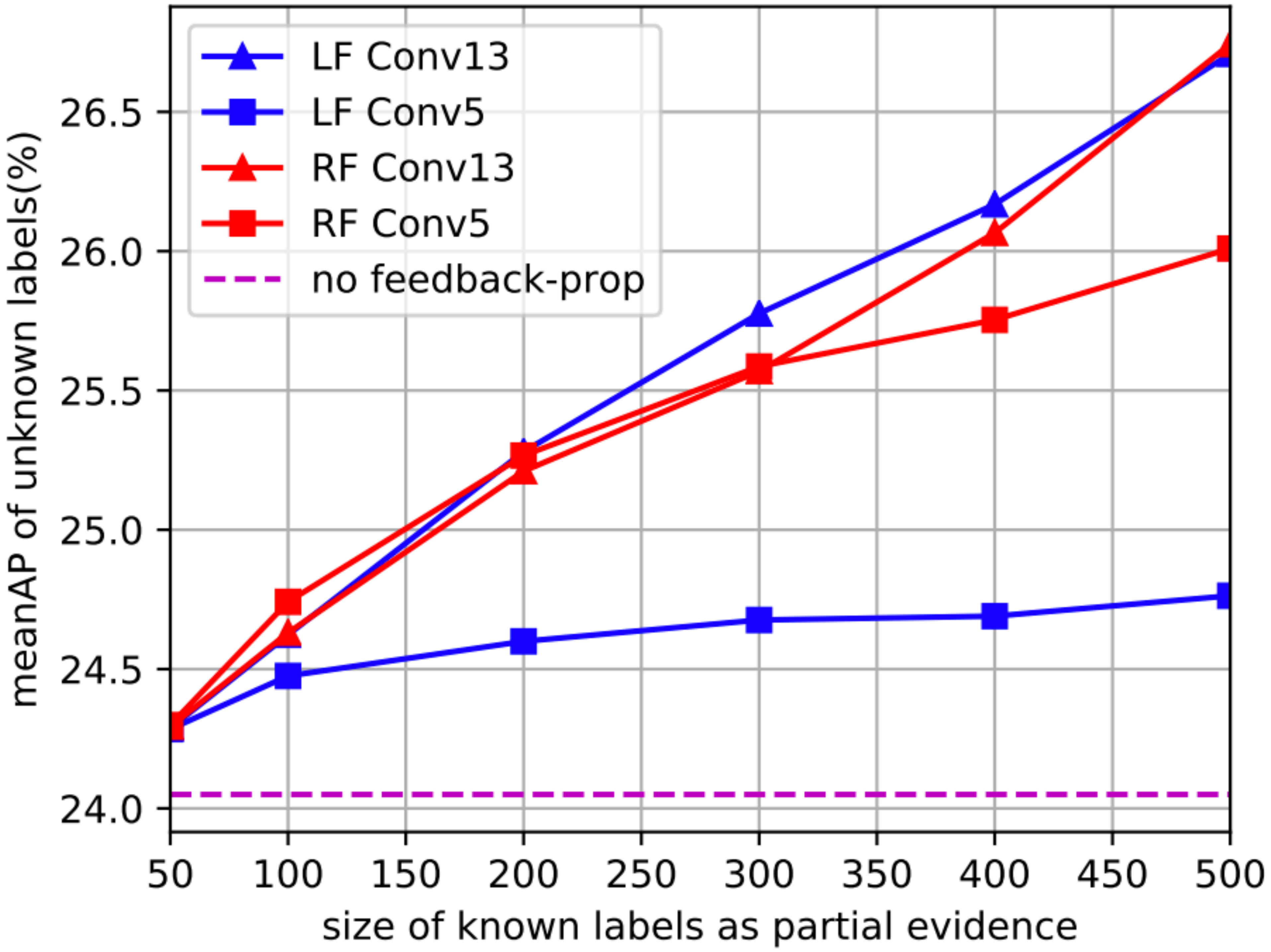
News images have captions and content.

Other Findings in our Paper

- Layer-wise analysis for Resnet-50 and VGG-16 for the best pivoting layers (where shared structure info is presumably maximal): **Happens in the middle layers! Not too close to input, not too close to outputs.**

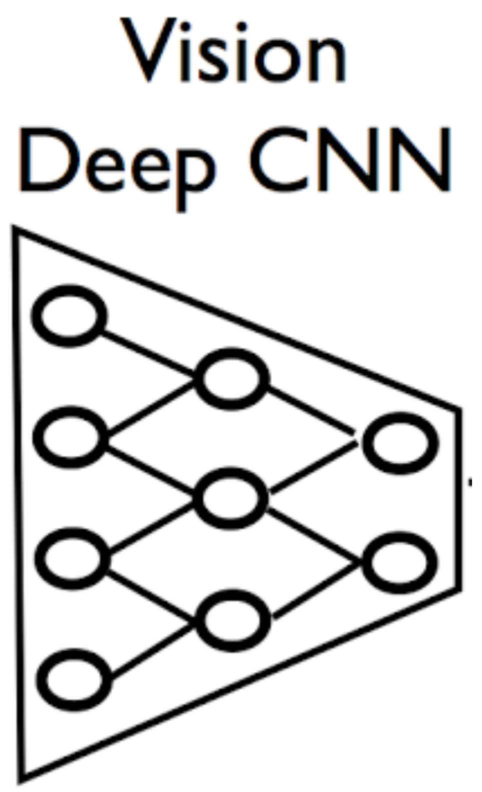
- **Extra information under this framework, even if noisy, or misleading, improves the predictions for the other tasks! and we did not even witness significant diminishing returns!!**

No diminishing returns? Just use more labels even if noisy



(b) Feedback-prop on ResNet18

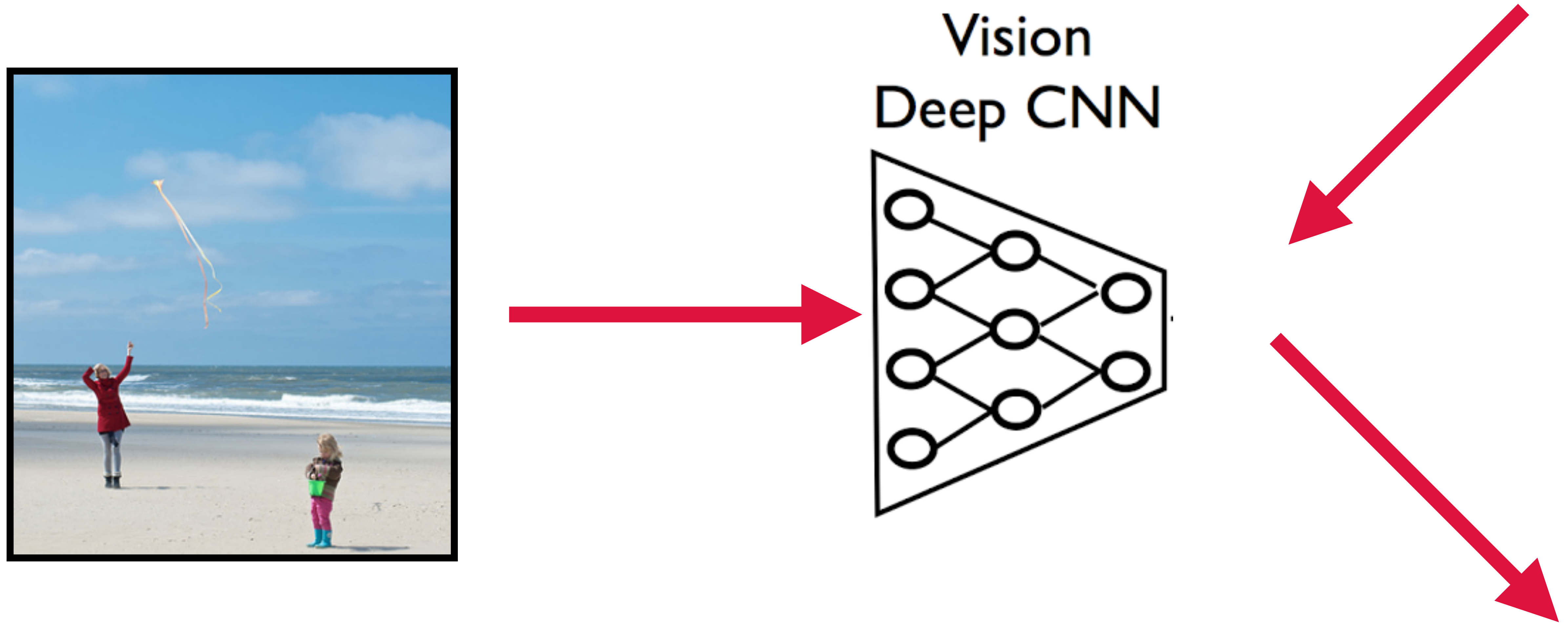
Future Directions? Holy Grail of Deep CNNs



Pretend Labels e.g.
Unsupervised Task

Semantic Task e.g.
Object classification

Future Directions? Holy Grail of Deep CNNs

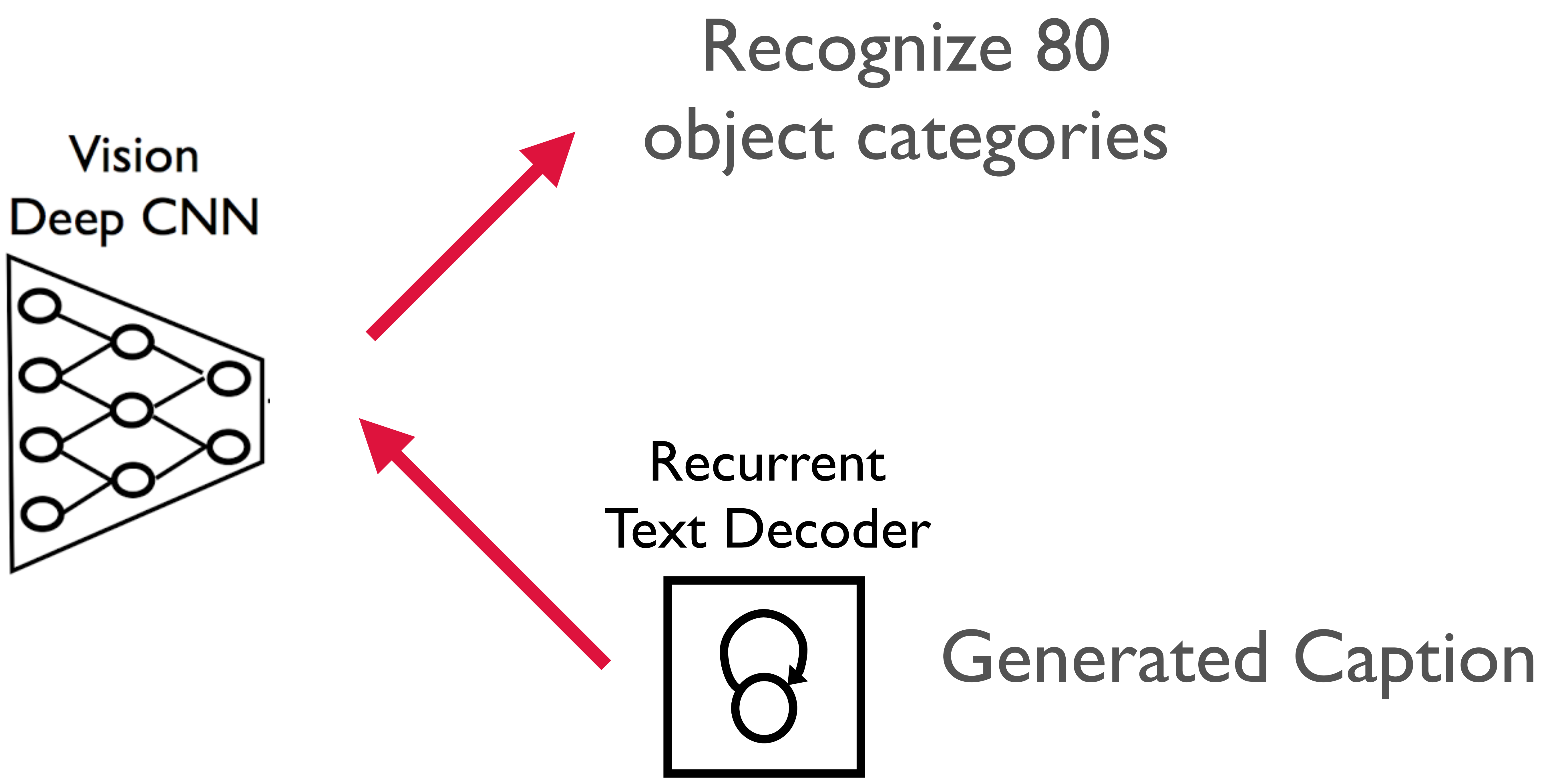


Pretend Labels e.g.
Unsupervised Task

Semantic Task

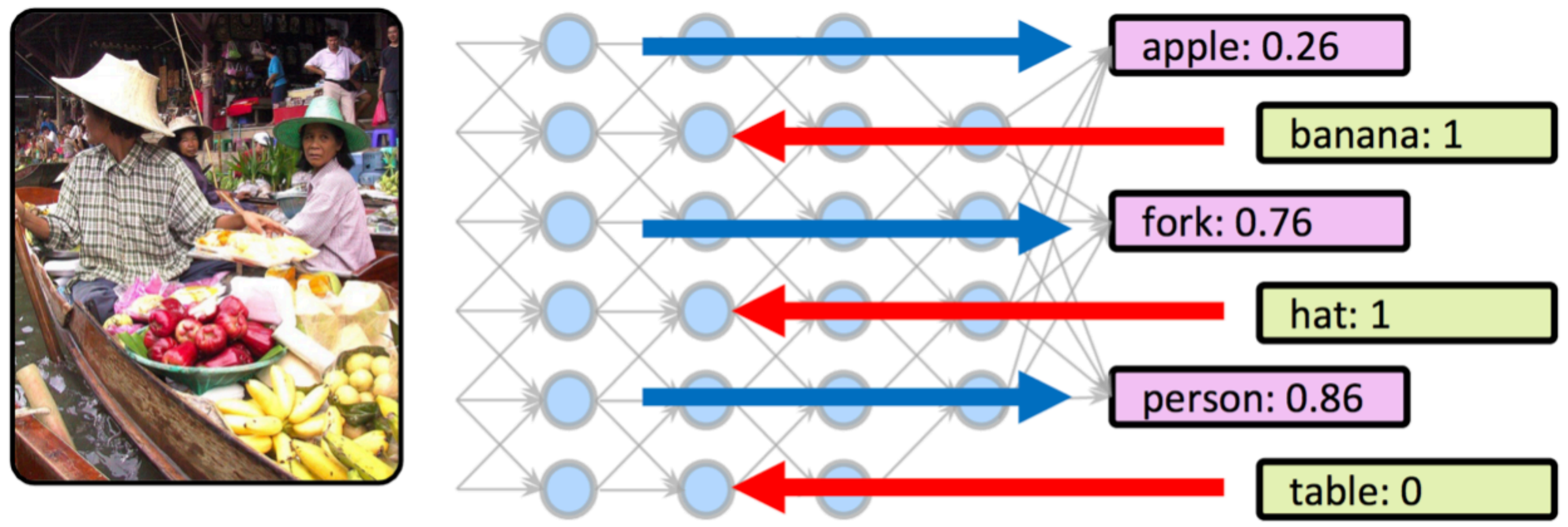
Future Directions? Learning Visual Common-sense Knowledge from Visual Sources for pure language tasks!

Can we discard the input image if only evidence after training is non-visual?



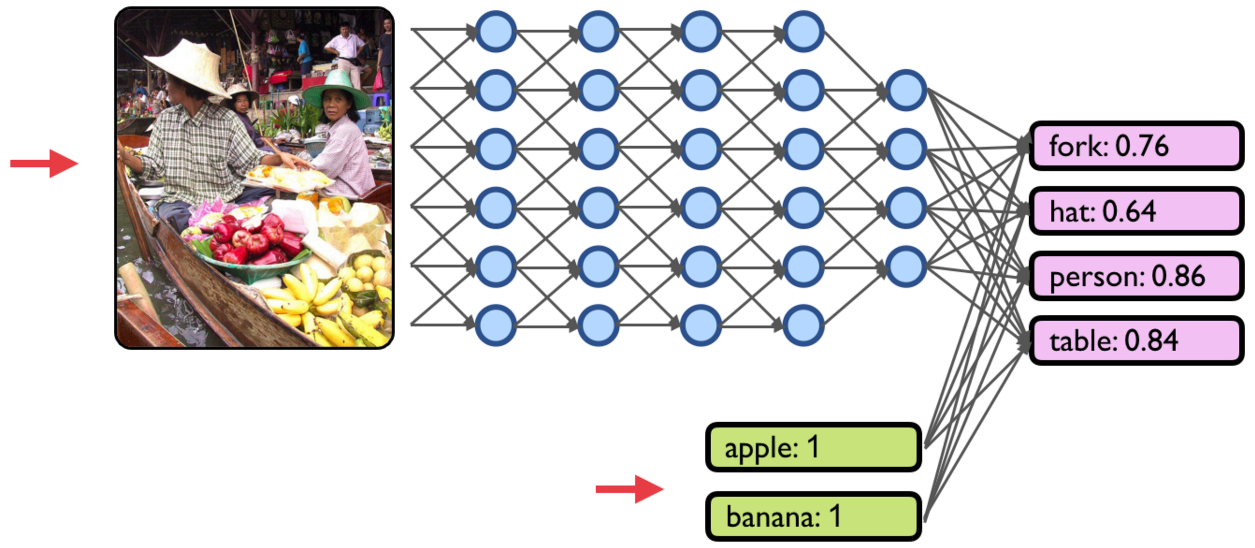
Still some way to go…

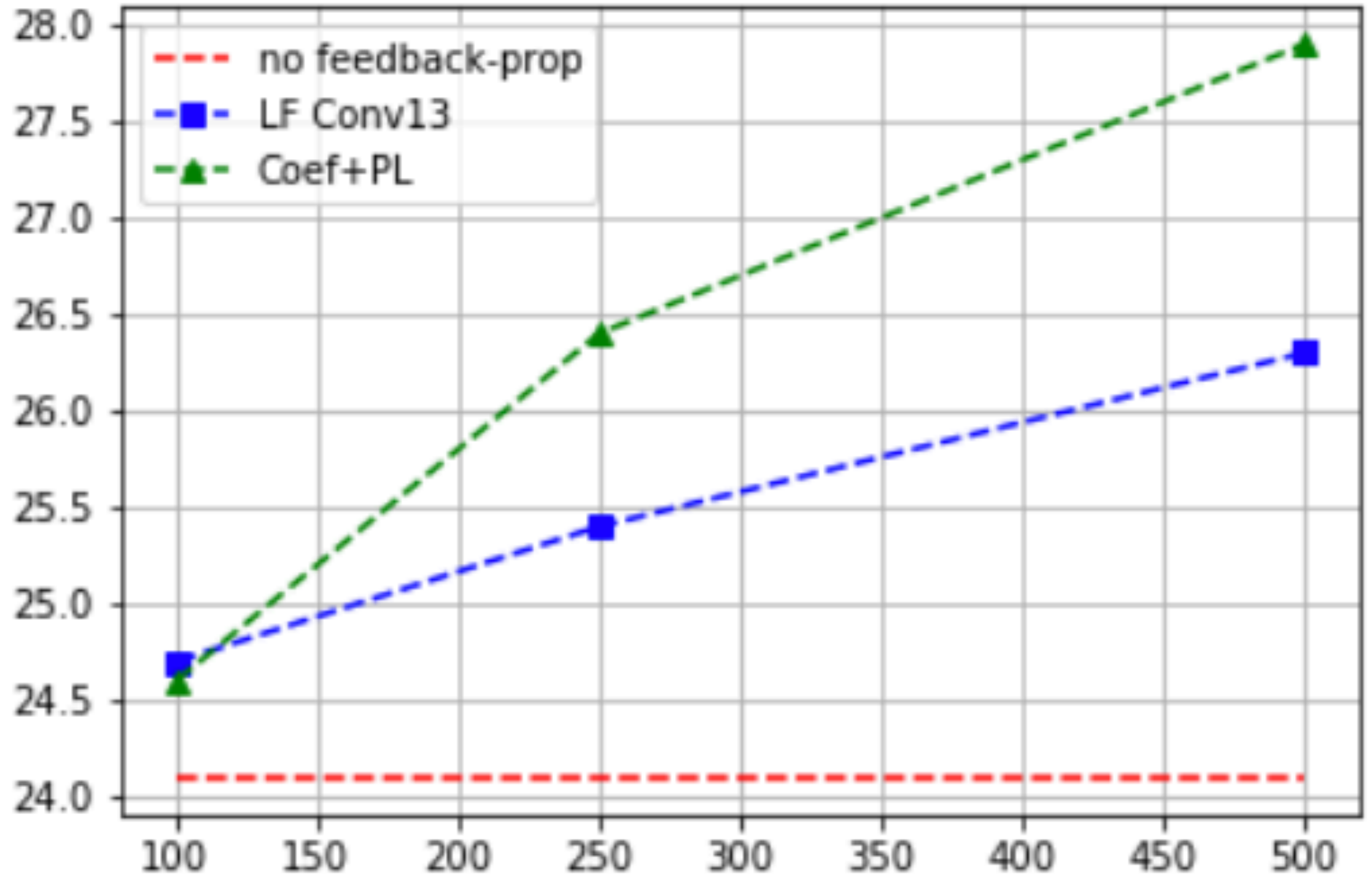
Feedback-prop



vs

Models trained explicitly under conditional labels





Other Future Directions

- **Unenrolled Feedback-propagation inference.**

  e.g simulate the feedback-process through a deeper network.

Other Future Directions

- **Unenrolled Feedback-propagation inference.**

  e.g simulate the feedback-process through a deeper network.

- **Incorporating Feedback-Loops during training.**

  e.g. feedback-loops break the DAG nature of DNNs but there could be workarounds.

Other Future Directions

- **Unenrolled Feedback-propagation inference.**

e.g simulate the feedback-process through a deeper network.

- **Incorporating Feedback-Loops during training.**

e.g. feedback-loops break the DAG nature of DNNs but there could be workarounds.

- **Feedback-loops are thought to be biologically plausible.**

Other Future Directions

- **Unenrolled Feedback-propagation inference.**

e.g simulate the feedback-process through a deeper network.

- **Incorporating Feedback-Loops during training.**

e.g. feedback-loops break the DAG nature of DNNs but there could be workarounds.

- **Feedback-loops are thought to be biologically plausible.**

Towards Biologically Plausible Deep Learning

Yoshua Bengio, Dong-Hyun Lee, Jorg Bornschein, Thomas Mesnard, Zhouhan Lin

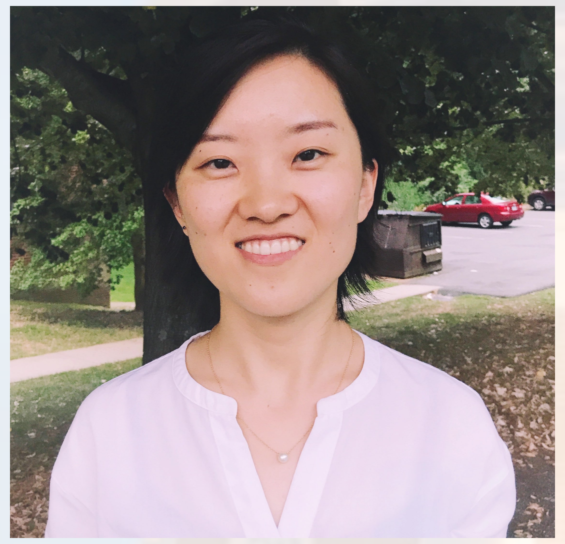
(Submitted on 14 Feb 2015 (v1), last revised 9 Aug 2016 (this version, v3))

Thanks

Kudos to students and collaborators!

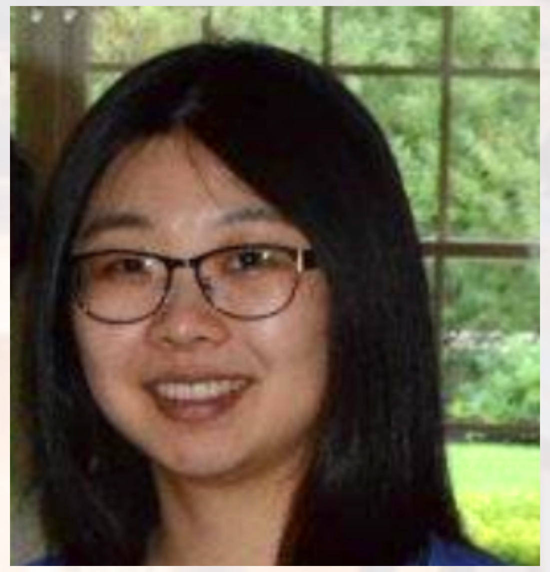
Tianlu Wang

Xuwang Yin

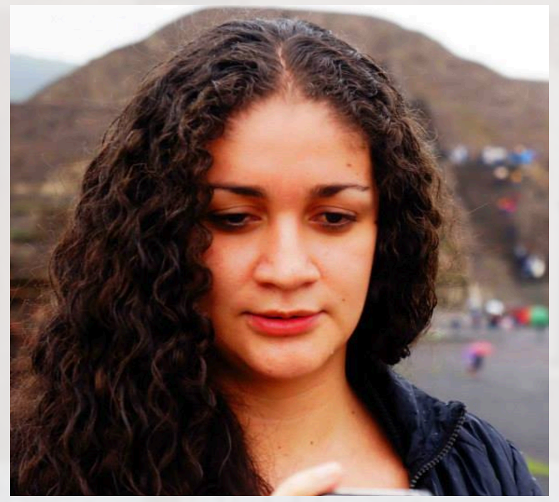


Jieyu Zhao

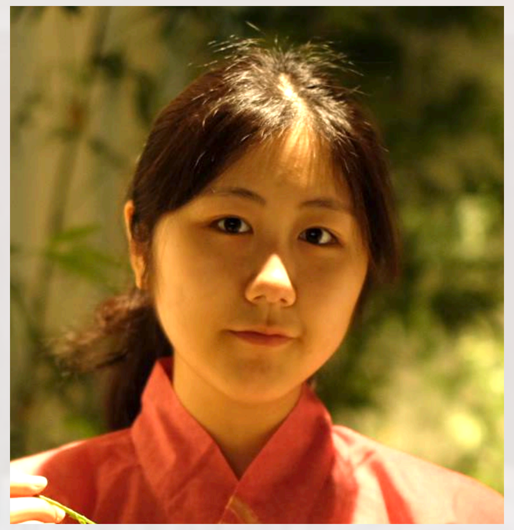
Mark Yatskar



Song Feng



Paola Cascante



Ziyan Yang

Fuwen Tan

Kai-Wei Chang

Kota Yamaguchi

Vicente Ordonez
Assistant Professor
Computer Science
University of Virginia
Twitter: @bluevincent
http://vicenteordonez.com

SAP®

Google

IBM Research