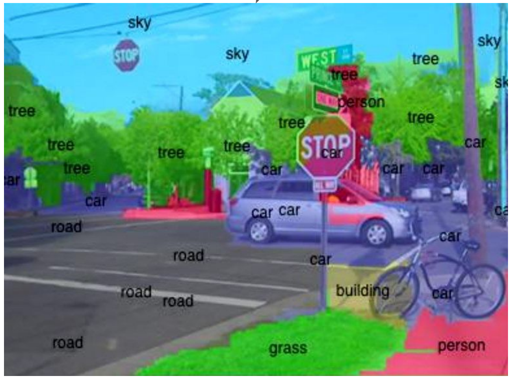
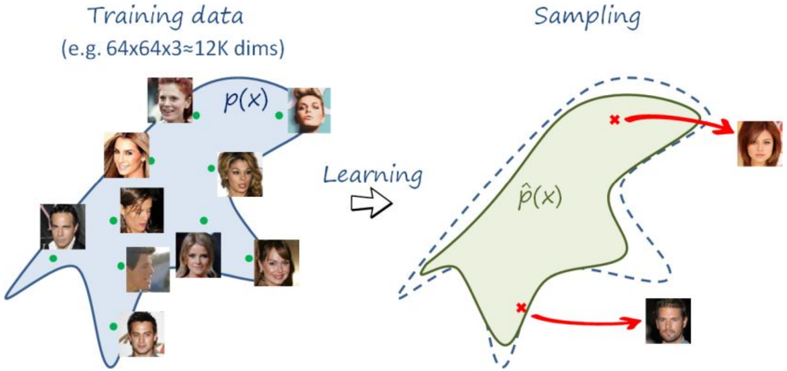
Stanford Artificial Intelligence Lab.

Machine Learning for Sustainability



Computer Vision



Generative Models

Layout

- Large Scale Pre-training Using Multi-modal Data.

- Learning When and Where to Zoom Using Deep Reinforcement Learning

- Poverty Mapping using Multi-modal data and Machine Learning.

Learning to Interpret Satellite Images using Wikipedia Articles

IJCAI - 2019

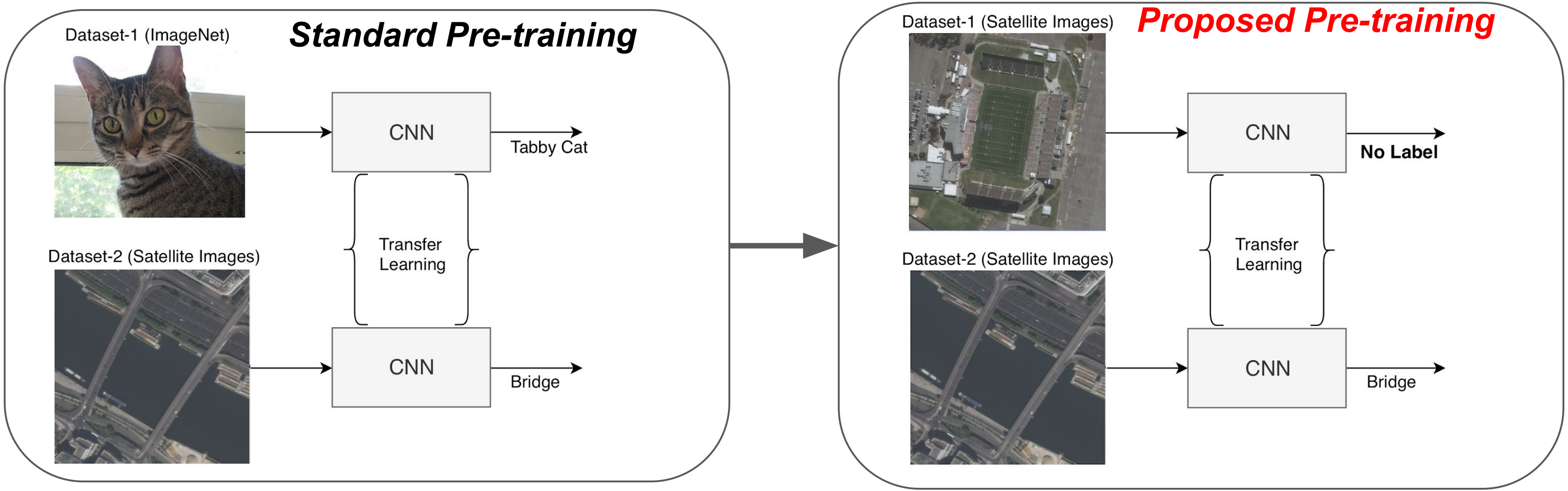
*Burak Uzkent, *Evan Sheehan, *Chenlin Meng, **David Lobell, **Marshall Burke, and *Stefano Ermon

*Department of Computer Science, Stanford University

*Department of Earth Science, Stanford University

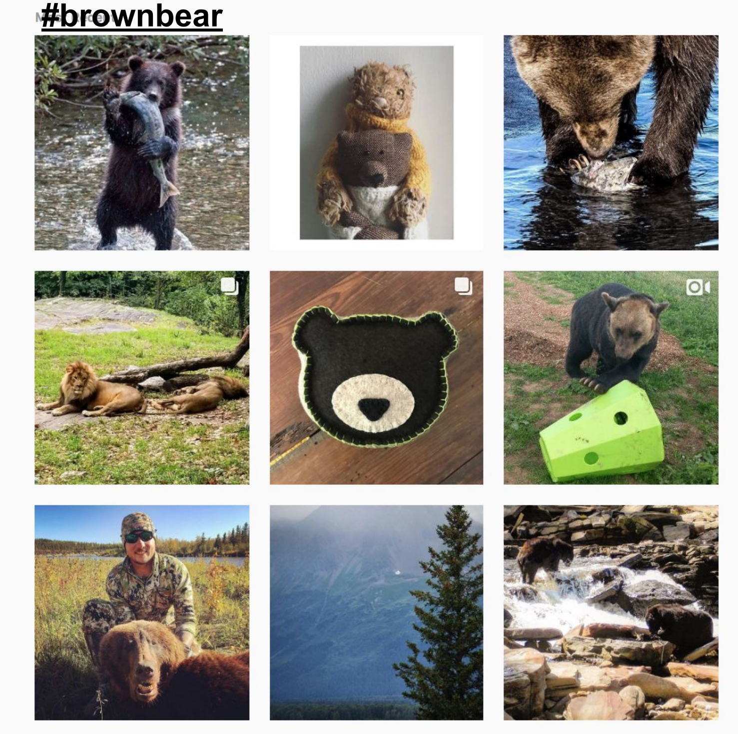
Introduction

- Almost all of the state-of-the-art deep learning models rely on the following framework.
  - *Pre-train on **ImageNet Dataset**.*
  - *Fine-tune on the **Target Dataset**.*



Related Work - Learning from Instagram Images with Hashtags

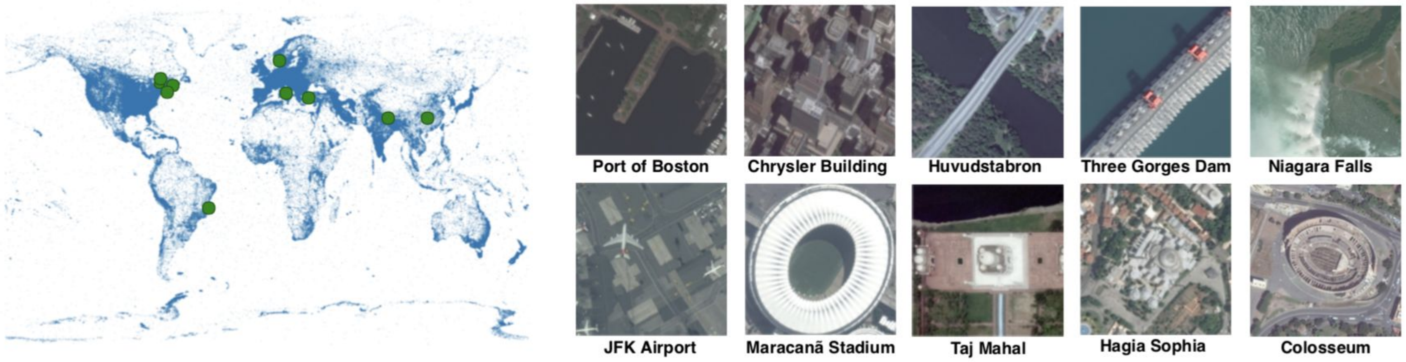
- Mahajan et al. builds an image recognition dataset consisting of 3 billion images from Instagram.
- They label the images using the hashtags given by the users.
- Two sets of labels are used:
    - ***ImageNet labels*** *(1k)*
    - ***WordNet synsets*** *(17k)*
- Pre-training improves recognition accuracy on **ImageNet by *%5*.**



Mahajan, Dhruv, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. "Exploring the limits of weakly supervised pretraining." In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 181-196. 2018.

Learning from Satellite Images using Wikipedia Articles

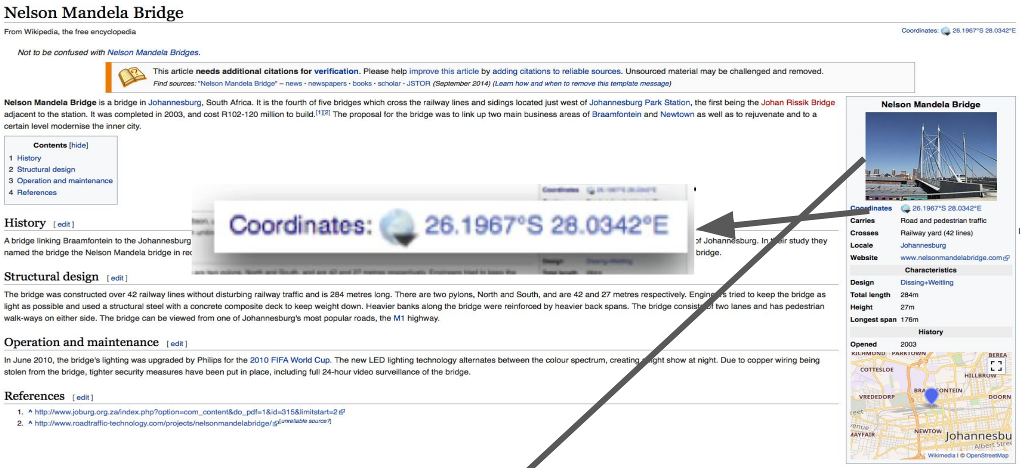
- In its most recent dump, Wikipedia contains *~5 million articles* (English) and *~1 million articles* are geo-referenced.



Scatter plot of the distribution of geo-tagged Wikipedia articles together with corresponding high resolution images.

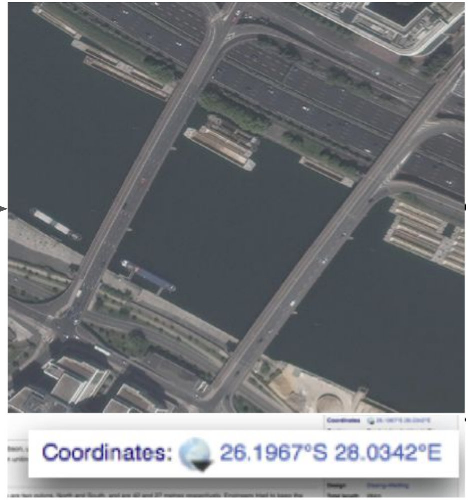
Pairing Articles to Satellite Images - WikiSatNet

$$\mathcal{D} = \{(c_1, x_1, y_1), (c_2, x_2, y_2), \cdots, (c_N, x_N, y_N)\}$$


Nelson Mandela Bridge
From Wikipedia, the free encyclopedia
Coordinates: 26.1967°S 28.0342°E
Not to be confused with Nelson Mandela Bridges.
This article needs additional citations for verification. Please help improve this article by adding citations to reliable sources. Unsourced material may be challenged and removed.
Find sources: "Nelson Mandela Bridge" – news · newspapers · books · scholar · JSTOR (September 2014) (Learn how and when to remove this template message)
Nelson Mandela Bridge is a bridge in Johannesburg, South Africa. It is the fourth of five bridges which cross the railway lines and sidings located just west of Johannesburg Park Station, the first being the Johan Rissik Bridge adjacent to the station. It was completed in 2003, and cost R102-120 million to build.[1][2] The proposal for the bridge was to link up two main business areas of Braamfontein and Newtown as well as to rejuvenate and to a certain level modernise the inner city.
Contents [hide]
1 History
2 Structural design
3 Operation and maintenance
4 References
History [edit]
A bridge linking Braamfontein to the Johannesburg ... of Johannesburg. In their study they named the bridge the Nelson Mandela bridge in re... bridge.
Structural design [edit]
The bridge was constructed over 42 railway lines without disturbing railway traffic and is 284 metres long. There are two pylons, North and South, and are 42 and 27 metres respectively. Engineers tried to keep the bridge as light as possible and used a structural steel with a concrete composite deck to keep weight down. Heavier banks along the bridge were reinforced by heavier back spans. The bridge consists of two lanes and has pedestrian walk-ways on either side. The bridge can be viewed from one of Johannesburg's most popular roads, the M1 highway.
Operation and maintenance [edit]
In June 2010, the bridge's lighting was upgraded by Philips for the 2010 FIFA World Cup. The new LED lighting technology alternates between the colour spectrum, creating a light show at night. Due to copper wiring being stolen from the bridge, tighter security measures have been put in place, including full 24-hour video surveillance of the bridge.
References [edit]
1. ^ http://www.joburg.org.za/index.php?option=com_content&do_pdf=1&id=315&limitstart=2
2. ^ http://www.roadtraffic-technology.com/projects/nelsonmandelabridge/ [unreliable source?]
Coordinates: 26.1967°S 28.0342°E

Nelson Mandela Bridge	
Coordinates	26.1967°S 28.0342°E
Carries	Road and pedestrian traffic
Crosses	Railway yard (42 lines)
Locale	Johannesburg
Website	www.nelsonmandelabridge.com
Characteristics	
Design	Dissing+Weitling
Total length	284m
Height	27m
Longest span	176m
History	
Opened	2003

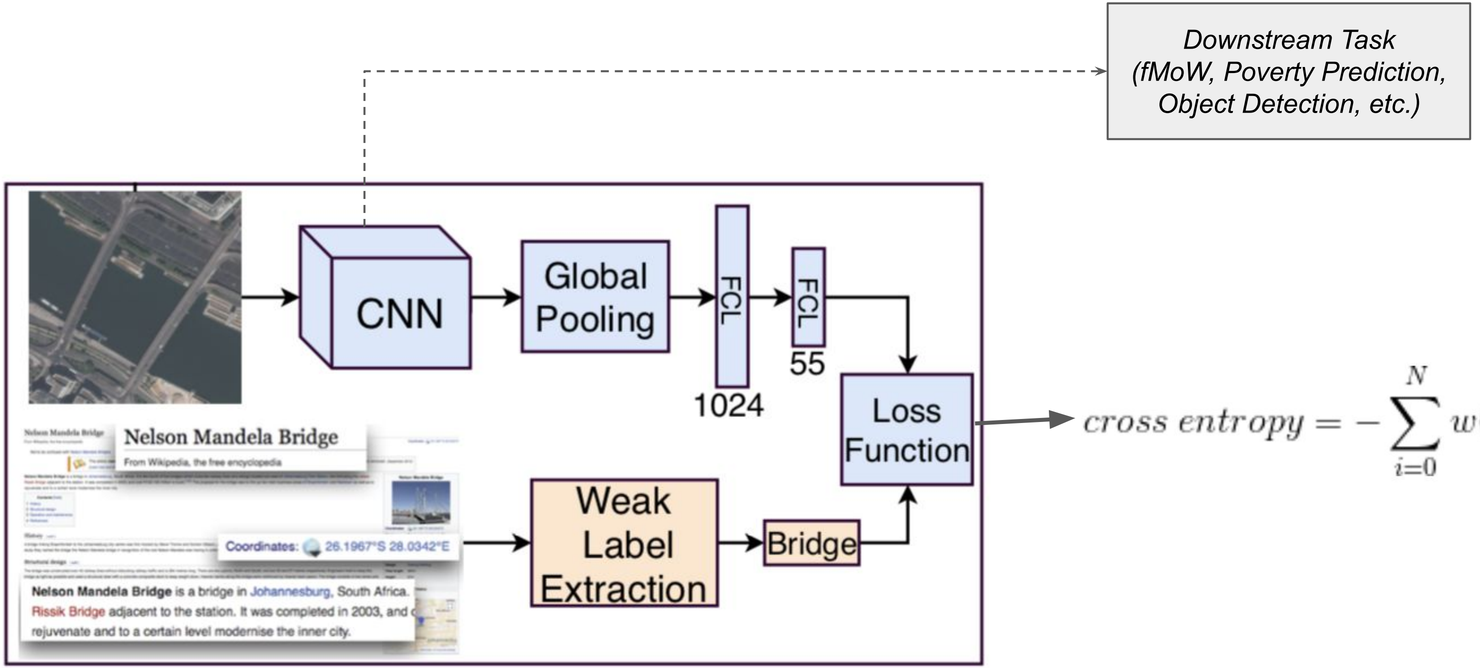
Pair to an overhead image



***Images embedded into Wikipedia Articles can also be used to learn deep visual representations. (Gomez et al. 2017)**

Gomez, L., Patel, Y., Rusiñol, M., Karatzas, D. and Jawahar, C.V., 2017. Self-supervised learning of visual features through embedding images into text topic spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4230-4239).

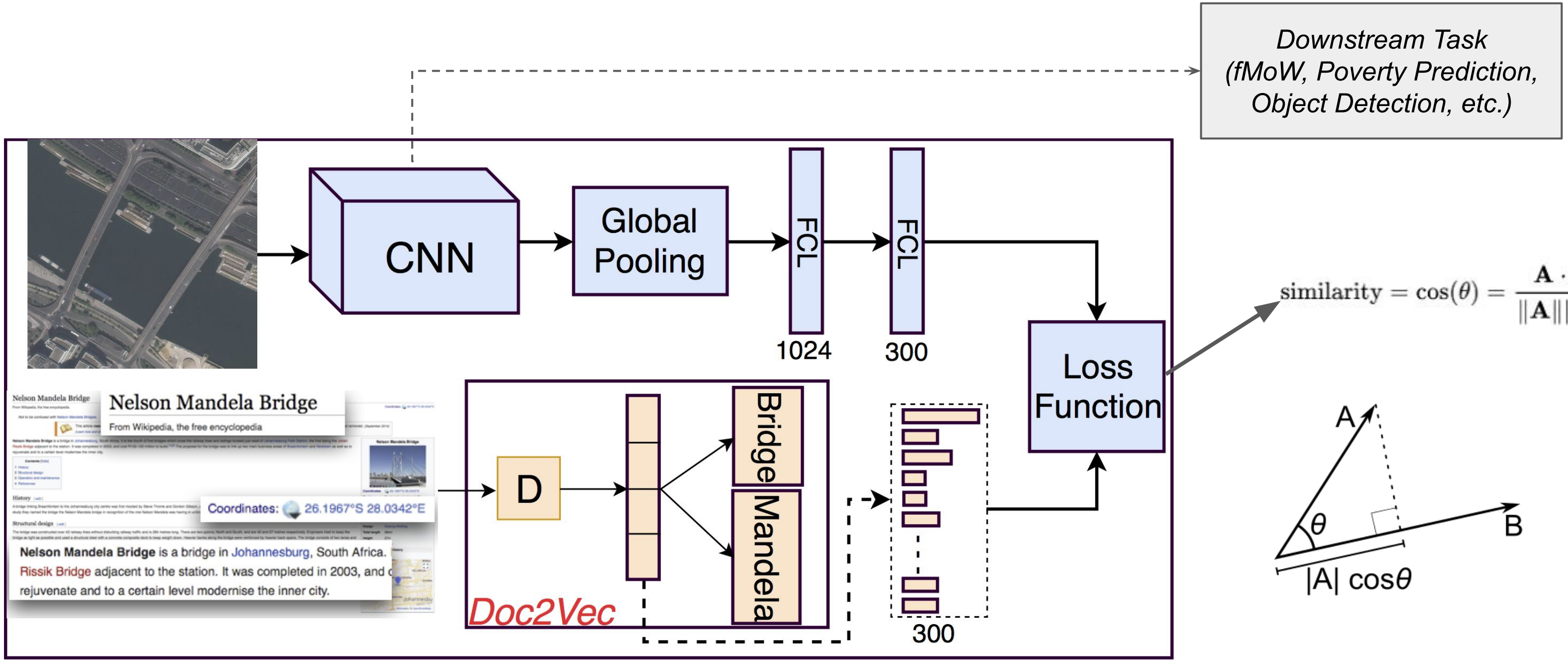
Representation Learning with Weak Labels



$$cross\ entropy = -\sum_{i=0}^{N} w(i) log(p(i))))$$

***Requires human intervention and heuristics.**

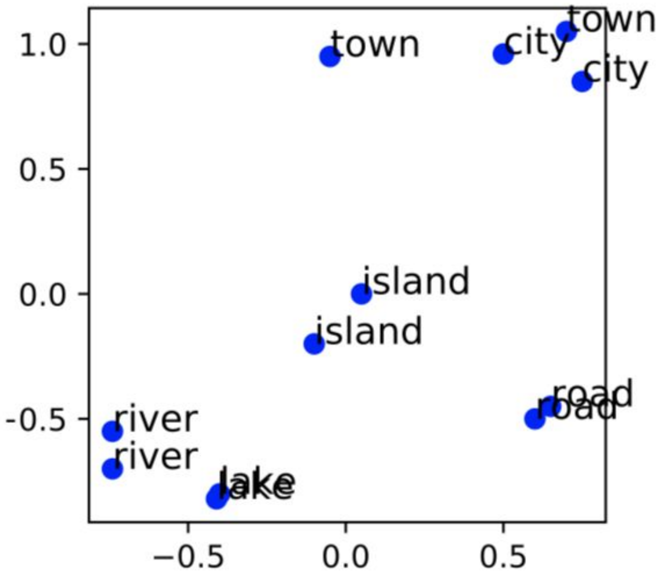
Representation Learning with Image2Text Matching



$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

***A more automatic approach.**

Analyzing Doc2Vec Model

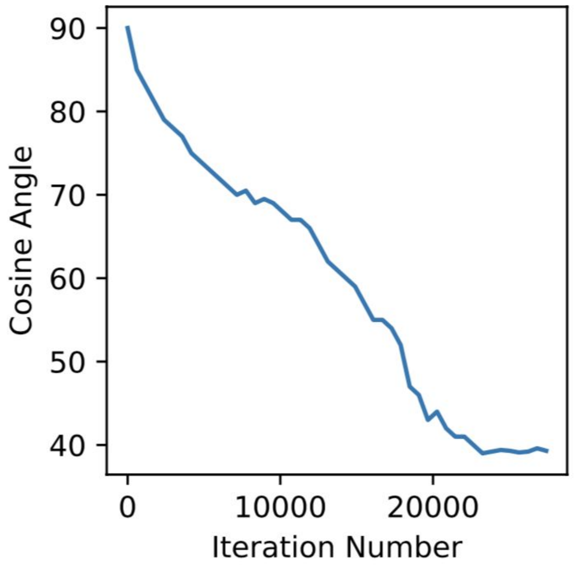


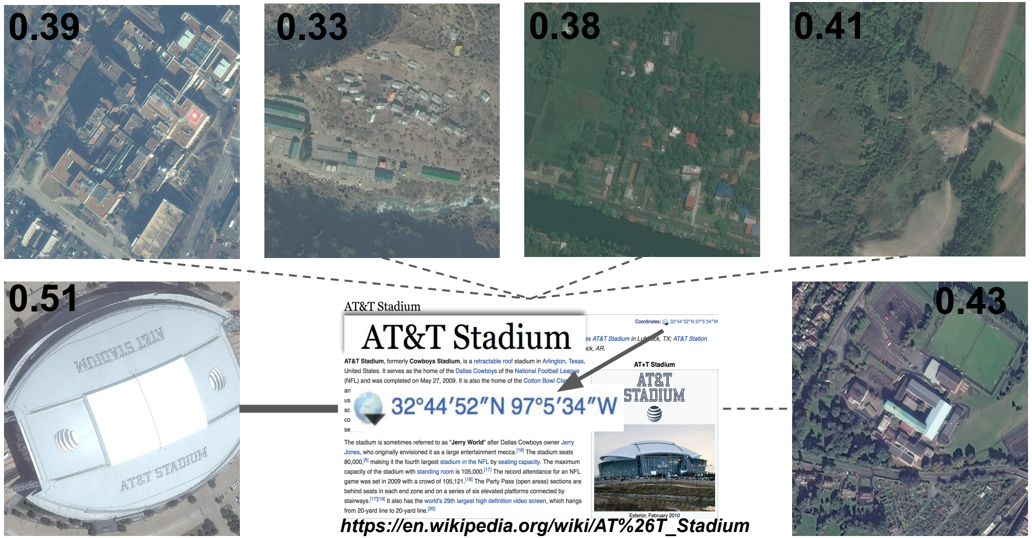
City - Middletown, Connecticut
City - Milton, Georgia
Lake - Timothy Lake
Lake - Tinquilco Lake
Town - Mingona Township, Kansas
Town - Moon Township, Pennsylvania
Road - Morehampton Road, Dublin
Road - Motorway M10 Pakistan
River - Motru River
River - Mousam River
Island - Aupaluktok Island
Island - Avatanak Island

***Articles with similar content are projected to the similar latent space.**

Image2Text Matching Pre-training Experiments

- We use DenseNet with 121 layers to parameterize the CNN.

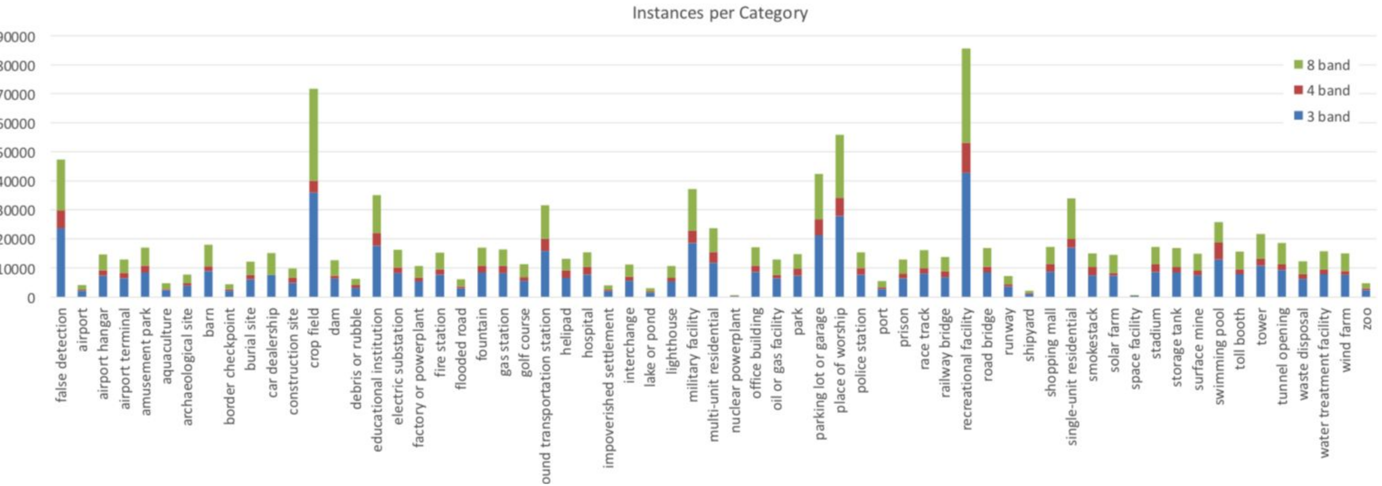




*Trained model matches the Wikipedia Article of AT&T Stadium to its corresponding overhe
image with higher similarity than it does to other images.

Target Task- functional Map of the World (fMoW)

- We use the recently released functional map of the world (fMoW) dataset consisting of high resolution satellite images.
- It includes 350k, 50k, 50k samples across 62 classes from the training, validation, and test sets.



Christie, Gordon, Neil Fendley, James Wilson, and Ryan Mukherjee. "Functional map of the world." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6172-6180. 2018.

Examples

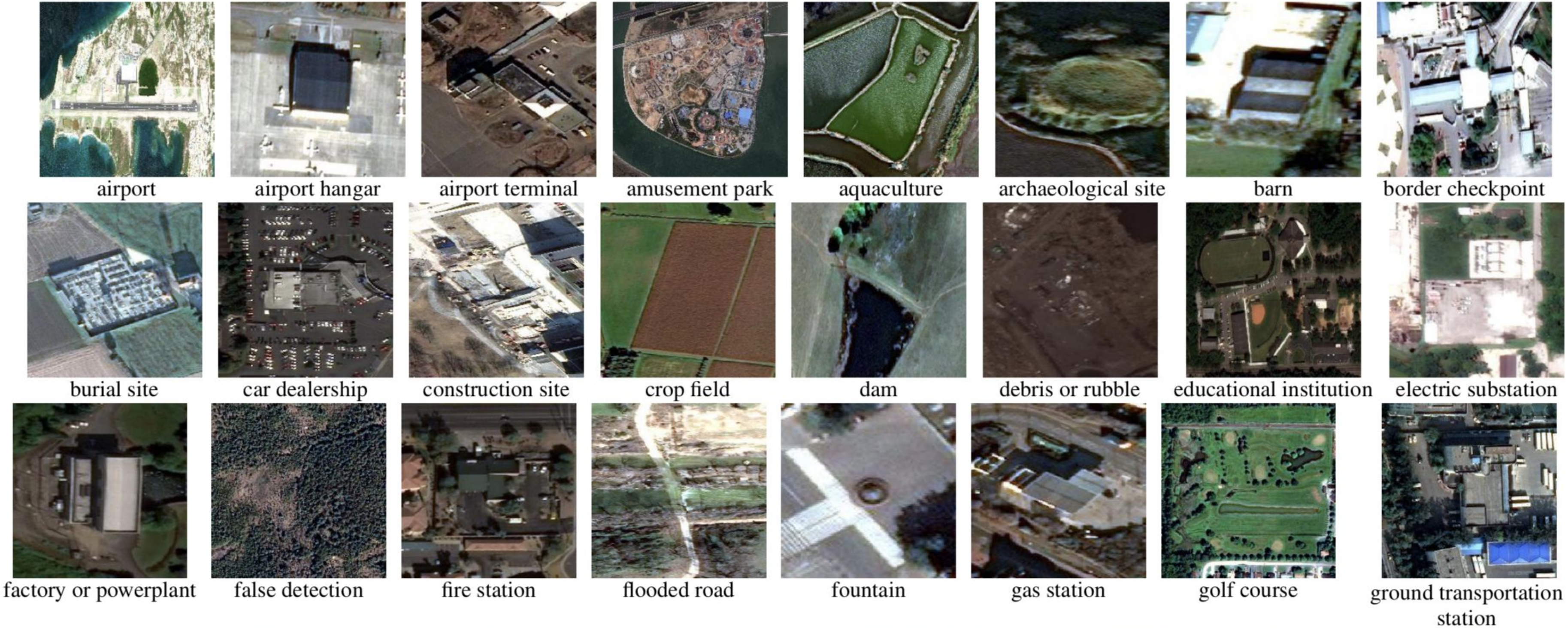
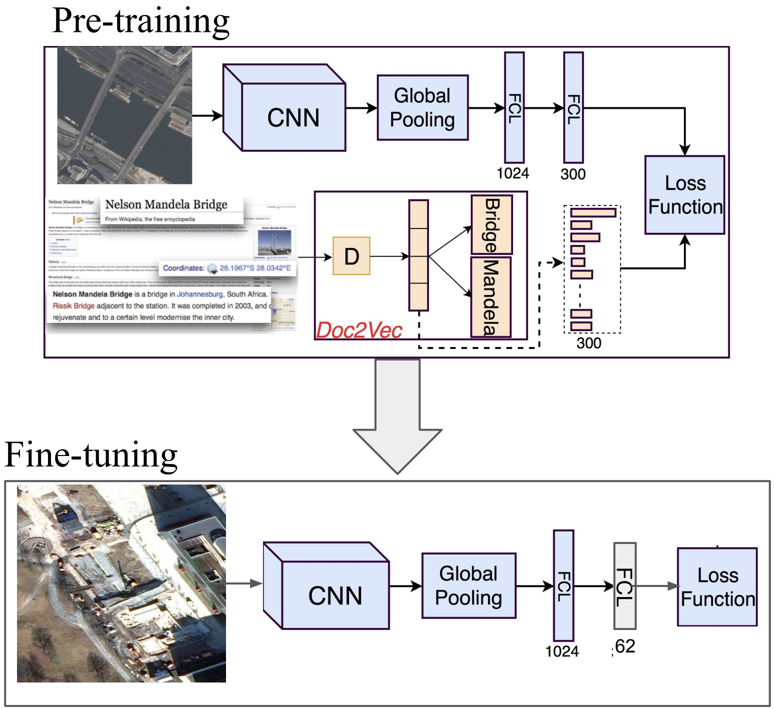
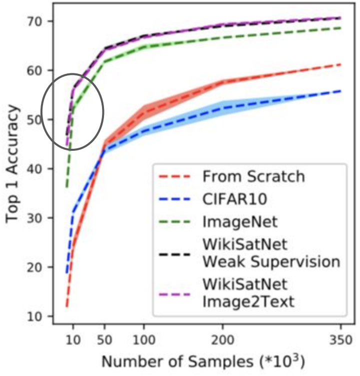
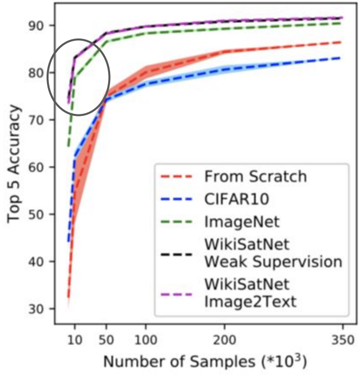

airport
airport hangar
airport terminal
amusement park
aquaculture
archaeological site
barn
border checkpoint
burial site
car dealership
construction site
crop field
dam
debris or rubble
educational institution
electric substation
factory or powerplant
false detection
fire station
flooded road
fountain
gas station
golf course
ground transportation station

Image Classification on fMoW





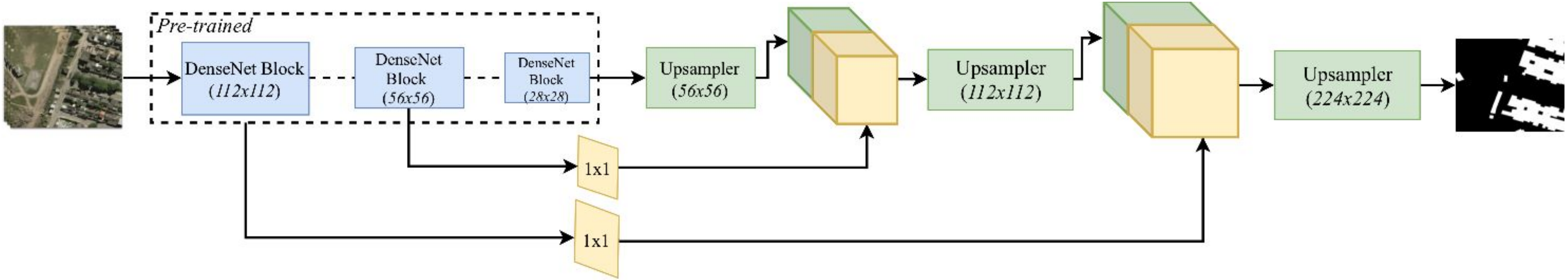
Gap decreases w.r.t sample complexity



Gap decreases w.r.t sample complexity

***Pre-training on a dataset with similar data distribution to the target dataset is very helpful when there is low sample complexity in the target dataset.**

Building Segmentation on SpaceNet



Model	From Scratch	ImageNet	WikiSatNet *Image2Text*
200 Samples	42.11 (%)	50.75 (%)	**51.70** (%)
500 Samples	48.98 (%)	54.63 (%)	**55.41** (%)
5000 Samples	57.21 (%)	59.63 (%)	**59.74** (%)

Mean IoU scores on SpaceNet test set

***Pre-training works best when we consider the same level tasks (image recognition - image recognition, semantic segmentation - semantic segmentation). (He et. al CVPR 2019)**

Learning Where and When to Zoom using Deep Reinforcement Learning

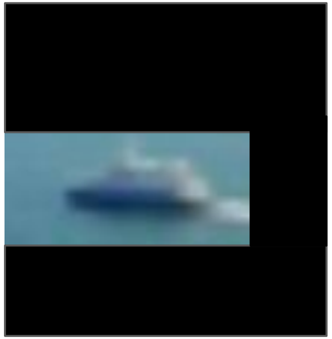
CVPR - 2020 (Under Review)

Burak Uzkent and Stefano Ermon

Department of Computer Science, Stanford University

Motivation

- Understanding the salient parts of an image is an important research field in computer vision.
- Previous approaches train a model and check the activation maps in test time to visualize the salient parts.
- In our study, we pose it as a Reinforcement Learning task and train an RL agent to learn *patch dropping policies*.



*Do we need the full image to be able to classify this image as ship?

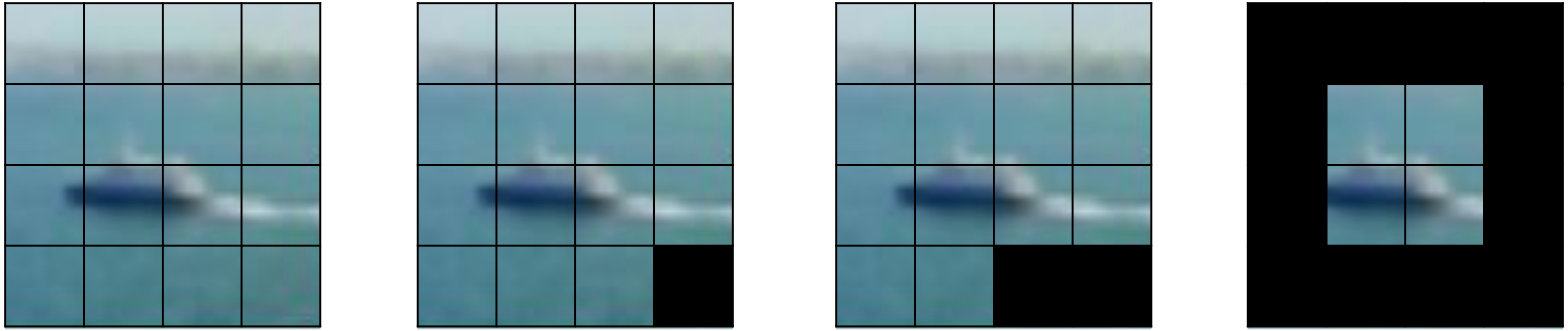
*Can we just process small part of this image and identify that it is ship?

<u>*If we process less number of pixels, we can build more efficient models.</u>

PatchDrop - An Adaptive Patch Sampling Framework

Do we need all the patches in an image to infer correct decisions?

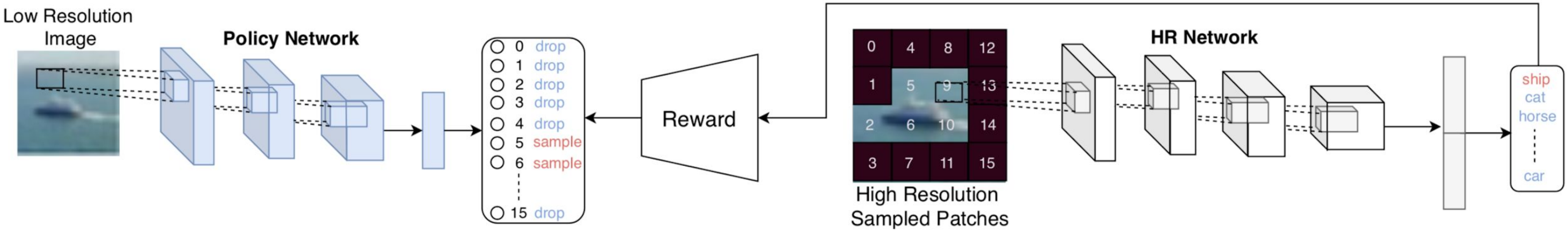
We train a ResNet32 on CIFAR10 and test it with random patch drop policy.



92.3% 91.1% 88.4% 46.3%

Can we design a conditional patch dropping strategy?

Proposed Framework



Policy Network

Classifier

Policies -> $\pi_1(\mathbf{a_1}|x_l;\theta_p) = p(\mathbf{a_1}|x_l;\theta_p)$ $\pi_2(\mathbf{a_2}|x_h^m;\theta_{cl}) = p(\mathbf{a_2}|x_h^m;\theta_{cl})$

Actions -> $\mathbf{a}_1 \in \{0,1\}^P$ $\mathbf{a}_2 \in \{0,1,\ldots N\}$

*Conditioning the Policy Network on low resolution images introduces minimal computational overhead.
*Additionally, in some domains, i.e. remote sensing, low resolution images are more affordable than high resolution images.

Modeling the Policy Network and Classifier

- The agent is trained using the predictions from the classification model.

Patch Sampling Policy->

$$\pi_1(\mathbf{a}_1|x_l, \theta_p) = \prod_{p=1}^{P} s_p^{\mathbf{a}_1^\mathbf{p}}(1 - s_p)^{(1-\mathbf{a}_1^\mathbf{p})}$$

Policy Network Predictions->

$$s_p = f_p(x_l; \theta_p) \qquad s_p \in [0, 1]$$

Classifier Predictions->

$$s_{cl} = f_c(x_h^m; \theta_{cl})$$

Cost Function->

$$\boxed{\max_{\theta_p} J(\theta_p, \theta_{cl}) = \mathbb{E}_p[R(\mathbf{a}_1, \mathbf{a}_2, y)]}$$

NOT Differentiable!

Training the Policy Network and Reward Function

- We train the Policy Network using the Policy Gradient Algorithm.

Cost Function to Maximize ->

$$\nabla_{\theta_p} J = \mathbb{E}[R(\mathbf{a}_1, \mathbf{a}_2, y) \nabla_{\theta_p} \boxed{\log \pi_{\theta_p}(\mathbf{a}_1 | x_l)}]$$

Differentiable!

$$\nabla_{\theta_p} J = \mathbb{E}[A \sum_{p=1}^{P} \nabla_{\theta_p} \log(s_p \mathbf{a}_1^{\mathbf{p}} + (1 - s_p)(1 - \mathbf{a}_1^{\mathbf{p}}))]$$

Advantage Function ->

$$A(\mathbf{a}_1, \hat{\mathbf{a}}_1, \mathbf{a}_2, \hat{\mathbf{a}}_2) = R(\mathbf{a}_1, \mathbf{a}_2, y) - R(\hat{\mathbf{a}}_1, \hat{\mathbf{a}}_2, y)$$

Temperature Scaling for Exploration/Exploitation Trade-off ->

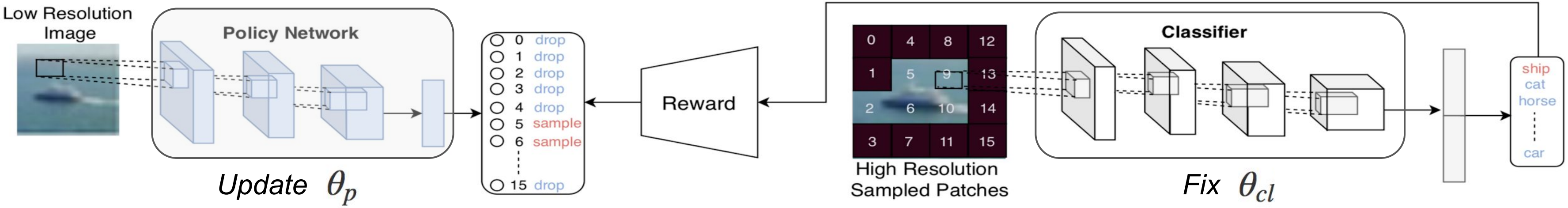
$$s_p = \alpha s_p + (1 - \alpha)(1 - s_p)$$

Reward Function ->

$$R(\mathbf{a}_1, \mathbf{a}_2, y) = \begin{cases} 1 - \left(\frac{|\mathbf{a}_1|_1}{P}\right)^2 & \text{if } y = \hat{y}(\mathbf{a}_2) \\ -\sigma & \text{Otherwise.} \end{cases}$$

Pre-training the Policy Network

- First, we train the classifier using original images without any masking.
- Next, we fix the classifier's weights and train the policy network.

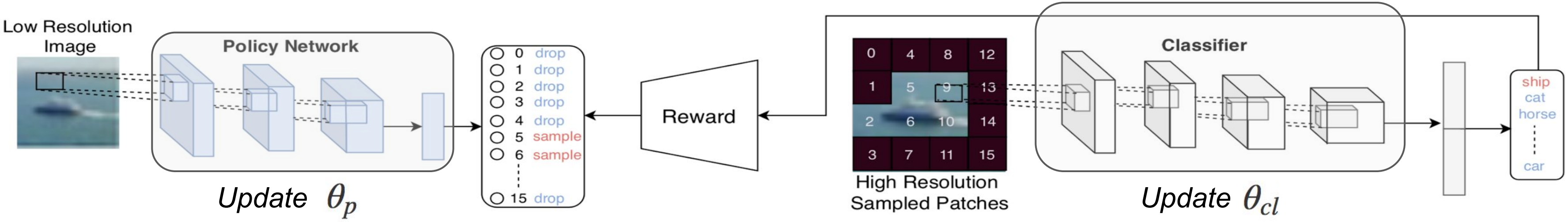


Pre-training Stage

- The policy network learns to understand *informative* patches however the overall accuracy is *reduced* since the classifier is not trained on *masked images*.

Jointly Fine-tuning the Policy Network and Classifier

- To boost the accuracy of the classifier, we finetune it jointly with the policy network.
- The classifier updates itself to adapt to the learned masked images and policy network updates the learned policies.



Joint Fine-tuning Stage

- At the end, in this step, we learn to drop more patches while increasing the accuracy w.r.t to the pre-training stage.

Experiments on CIFAR10/CIFAR100/ImageNet

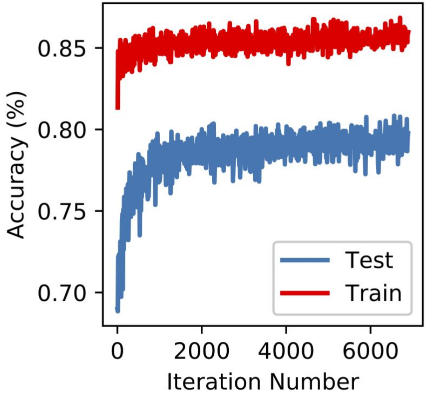
- For CIFAR10 and CIFAR100, we use 45k, 5k, and 10k training, validation and test samples.
- For ImageNet, we use 1.2 million, 50k, and 150k training, validation and test images.

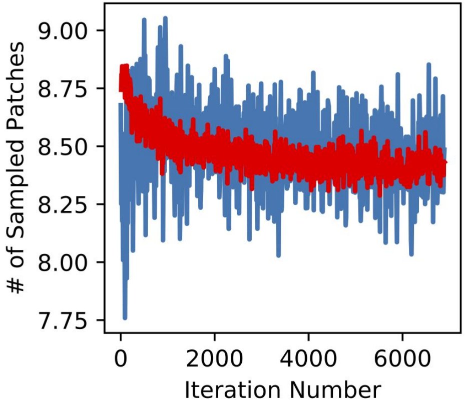
	CIFAR10				CIFAR100				ImageNet			
	Acc. (%) (Pre-training)	Acc. (%) (Ft-1)	Acc. (%) (Ft-2)	S	Acc. (%) (Pre-training)	Acc. (%) (Ft-1)	Acc. (%) (Ft-2)	S	Acc. (%) (Pre-training)	Acc. (%) (Ft-1)	Acc. (%) (Ft-2)	S
Fixed-H	71.2	88.8	89.2	9,9,9	48.5	65.8	68.0	9,10,10	59.8	68.6	71.9	10,9,7
Fixed-V	64.7	88.4	89.1	9,9,9	46.2	65.5	68.5	9,10,10	59.4	68.4	72.1	10,9,7
Stochastic	40.6	88.1	88.7	9,9,9	27.6	63.2	65.4	9,10,10	57.6	67.2	70.4	10,9,7
Activations Maps	56.6	88.9	89.5	9,9,9	40.4	64.0	67.6	9,10,10	59.4	67.2	70.3	10,9,7
SRGAN	78.8	78.8	78.8	0,0,0	69.1	56.1	56.1	0,0,0	69.1	69.1	69.1	0, 0, 0
STN	56.9	88.2	89.1	9,9,9	41.1	64.3	67.2	9,10,10	58.6	71.1	72.3	10, 9, 7
PatchDrop	**80.6**	**91.9**	**91.5**	**8.5,7.9,6.9**	**57.3**	**69.3**	**70.4**	**9,10,9.8**	**63.7**	**74.9**	**76.3**	**10.1, 8.5, 6.9**
No Patch Sampling	75.8	75.8	75.8	0,0,0	55.1	55.1	55.1	0,0,0	67.4	67.4	67.4	0,0,0
w/o Patch Dropping	92.3	92.3	92.3	16,16,16	69.3	69.3	69.3	16,16,16	76.5	76.5	76.5	16,16,16

The proposed framework drops about %40-%60 of the patches while maintaining the classification accuracy of the model using original HR images.

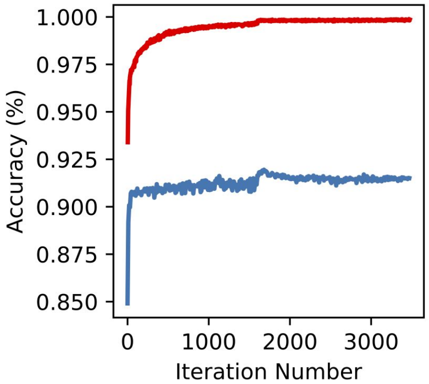
Impact of Joint Fine-tuning

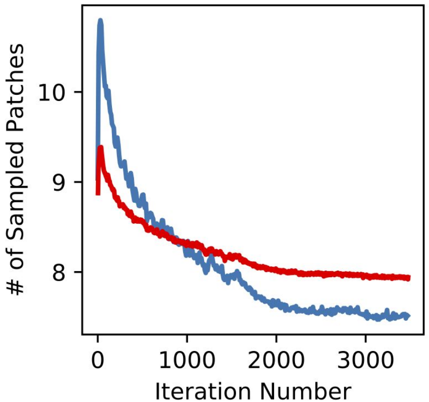
Pre-training



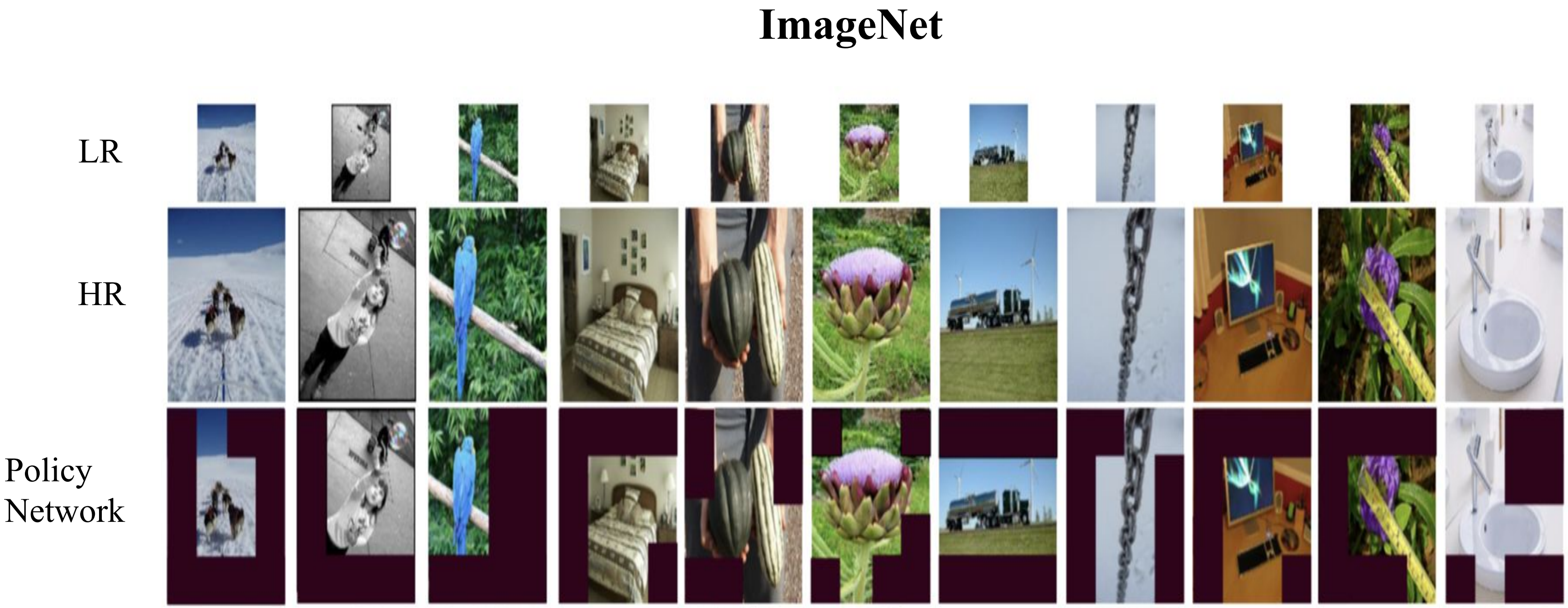


Joint Fine-tuning





Learned Patch Sampling Policies



Experiments on fMoW

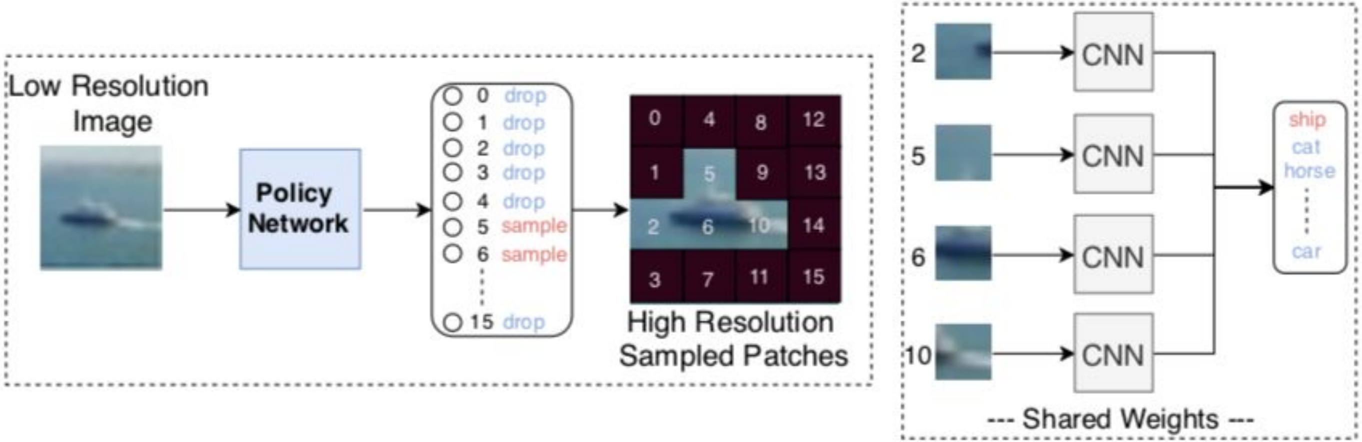
- For fMoW, we use 350k, 50k, and 50k training, validation and test samples.
- Original images are 224x224px whereas the images used by the policy network is 56x56px.

	Acc. (%) (Pre-training)	S	Acc. (%) (Ft-1)	S	Acc. (%) (Ft-2)	S
Fixed-H	47.7	7	63.3	6	65.5	6
Fixed-V	48.3	7	63.2	6	65.3	6
Stochastic	29.1	7±1.7	57.1	6±1.7	63.6	6±1.6
Activation Maps	37.1	7	61.1	6	64.6	6
SRGAN	63.3	0	63.3	0	63.3	0
STN	37.5	7	61.8	6	64.8	6
PatchDrop	**53.4**	**7**±2.7	**65.9**	**5.9**±2.4	**68.3**	**6.0**±2.4
No Patch Sampling	62.7	0	62.7	0	62.7	0
w/o Patch Dropping	67.3	16	67.3	16	67.3	16

Learned Patch Sampling Policies

Functional Map of the World

Conditional BagNets



Brendel, Wieland, and Matthias Bethge. "Approximating cnns with bag-of-local-features models works surprisingly well on imagenet." *arXiv preprint arXiv:1904.00760* (2019).

Conditional BagNets - Experiments on CIFAR10

	Acc. (%) (Pt)	S	Acc. (%) (Ft-1)	S	Run-time. (%) (ms)
BagNet (No Patch Drop) [1]	85.6	16	85.6	16	192
CNN (No Patch Drop)	92.3	16	92.3	16	77
Fixed-H	67.7	10	86.3	9	98
Fixed-V	68.3	10	86.2	9	98
Stochastic	49.1	10	83.1	9	98
STN [19]	67.5	10	86.8	9	112
BagNet (PatchDrop)	**77.4**	**9.5**	**92.7**	**8.5**	98

Brendel, Wieland, and Matthias Bethge. "Approximating cnns with bag-of-local-features models works surprisingly well on imagenet." *arXiv preprint arXiv:1904.00760* (2019).

Conditional Hard Positive Generation

	CIFAR10 (%) (ResNet32)	CIFAR100 (%) (ResNet32)	ImageNet (%) (ResNet50)	fMoW (%) (ResNet34)
No Augment.	92.3	69.3	76.5	67.3
CutOut [5]	93.5	70.4	76.5	67.6
PatchDrop	**93.9**	**71.0**	**78.1**	**69.6**

DeVries, Terrance, and Graham W. Taylor. "Improved regularization of convolutional neural networks with cutout." *arXiv preprint arXiv:1708.04552* (2017).

Predicting Economic Development using Geolocated Wikipedia Articles

KDD - 2019

*Evan Sheehan, *Chenli Meng, *Matthew Tan, *Burak Uzkent, *Neal Jean, **David Lobell, **Marshall Burke, and *Stefano Ermon

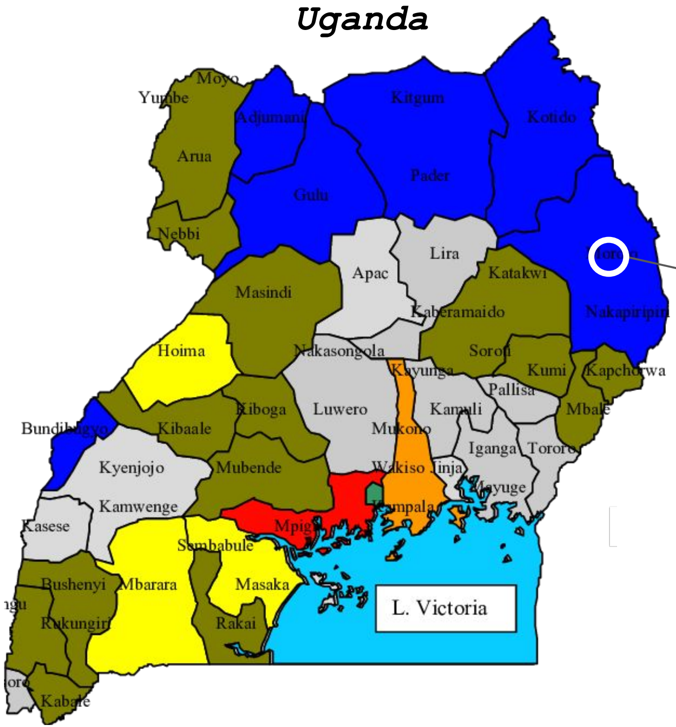
*Department of Computer Science, Stanford University

*Department of Earth Science, Stanford University

Motivation

- #1 UN Sustainable Development Goal:
  - Global Poverty Line : **$1.90** per person for one day.
- Understanding poverty can lead to:
  - Informed policy making
  - Targeted NGO and aid efforts.

Motivation



Representation
(Satellite, Nightlight Images, Wikipedia etc.) → **Build a Machine Learning Model** → **Poverty measures**

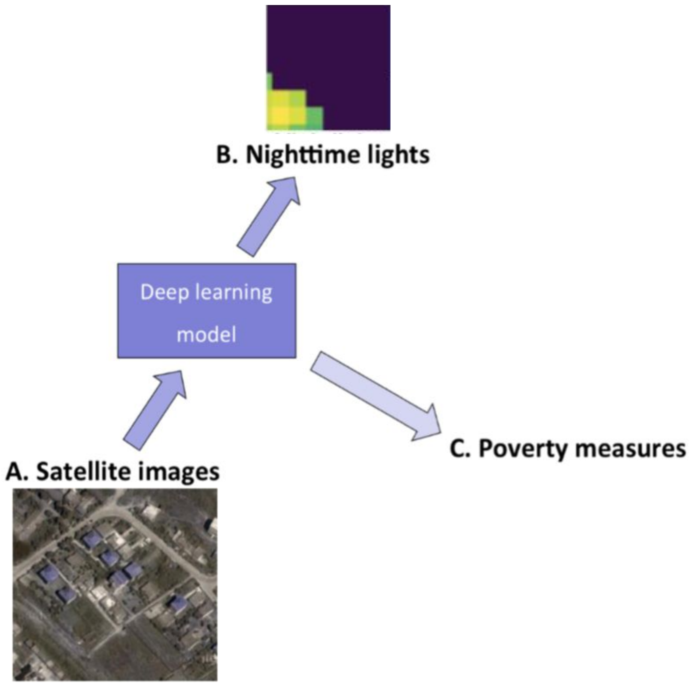

Satellite

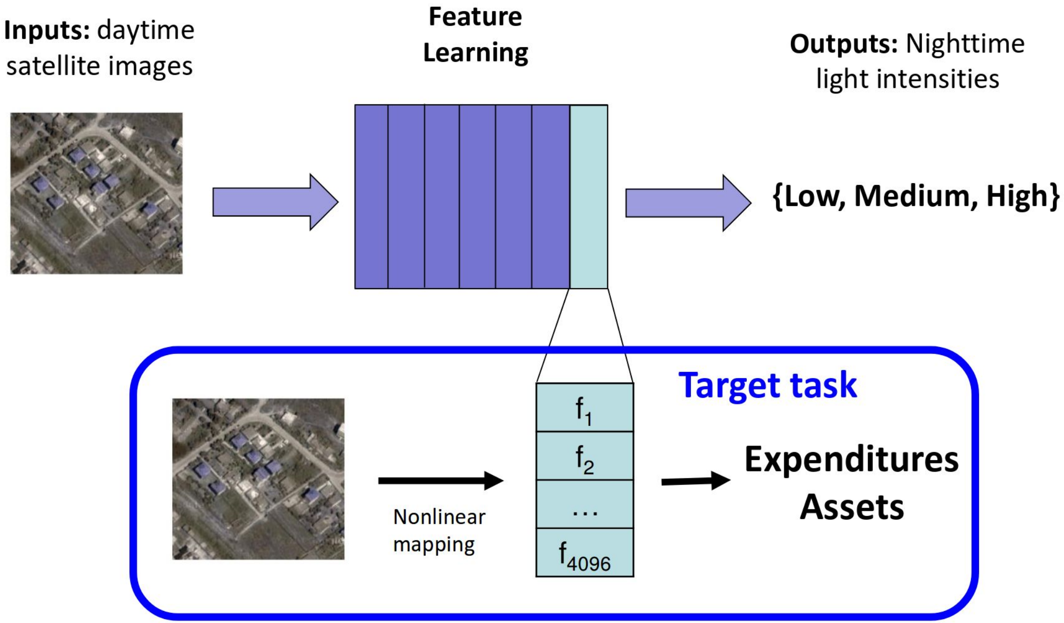

Nightlight

Wikipedia Article

Related Work

Jean et al. (Science 2016)





Geo-located Wikipedia Articles

- Poverty prediction has been previously tackled by nightlight images.
- We use geolocated Wikipedia articles to better predict poverty.

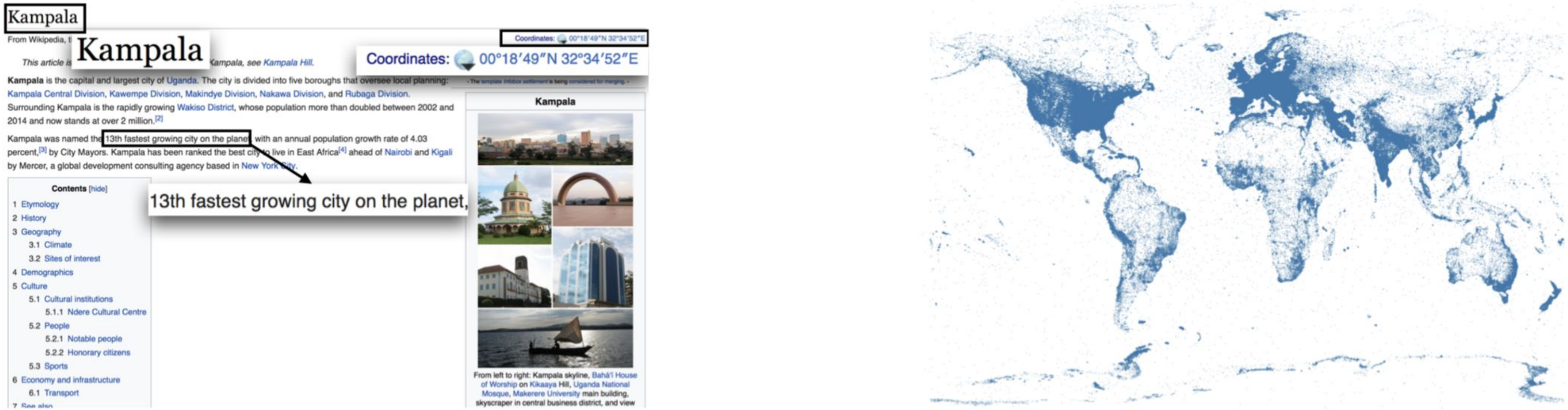
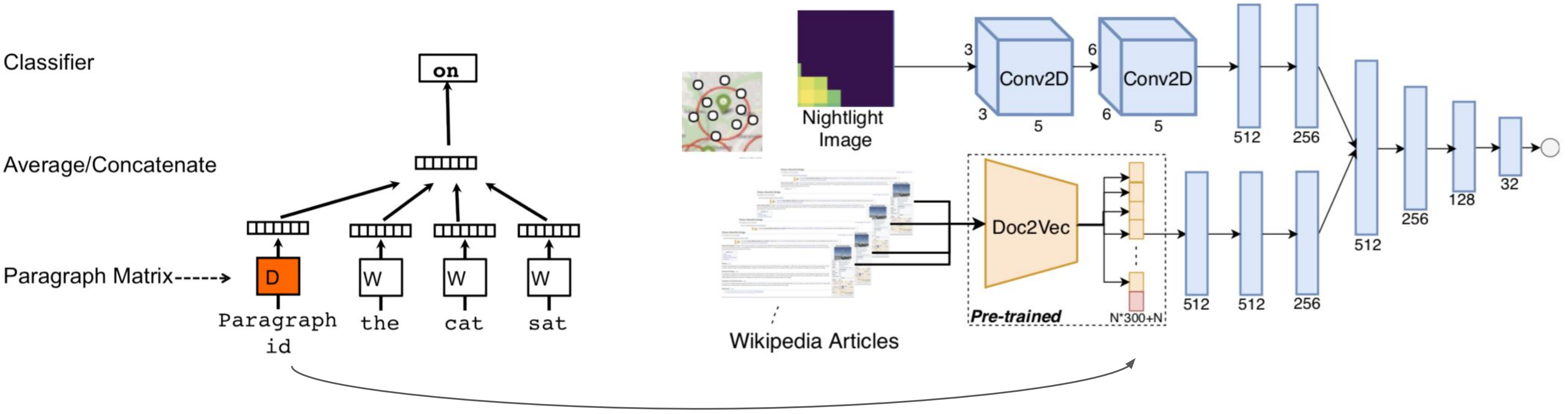


Figure 1: Left: Example of a geolocated Wikipedia article. Articles such as this contain a wealth of information relevant to economic development. Right: Global distribution of geolocated Wikipedia Articles. Note that there is no overlayed basemap, yet the shape of the continents arises naturally from the spatial distribution of articles.

Proposed Method

- We train the Doc2Vec model on ~1.2 million geolocated articles w/o supervision.
- Our multi-modal model uses nightlight images and features from articles to predict poverty.



Proposed approach to perform poverty prediction on Africa.

Dataset

- There is 8k ground truth samples from African continent including countries Ghana, Malawi, Tanzania, Nigeria, Uganda.

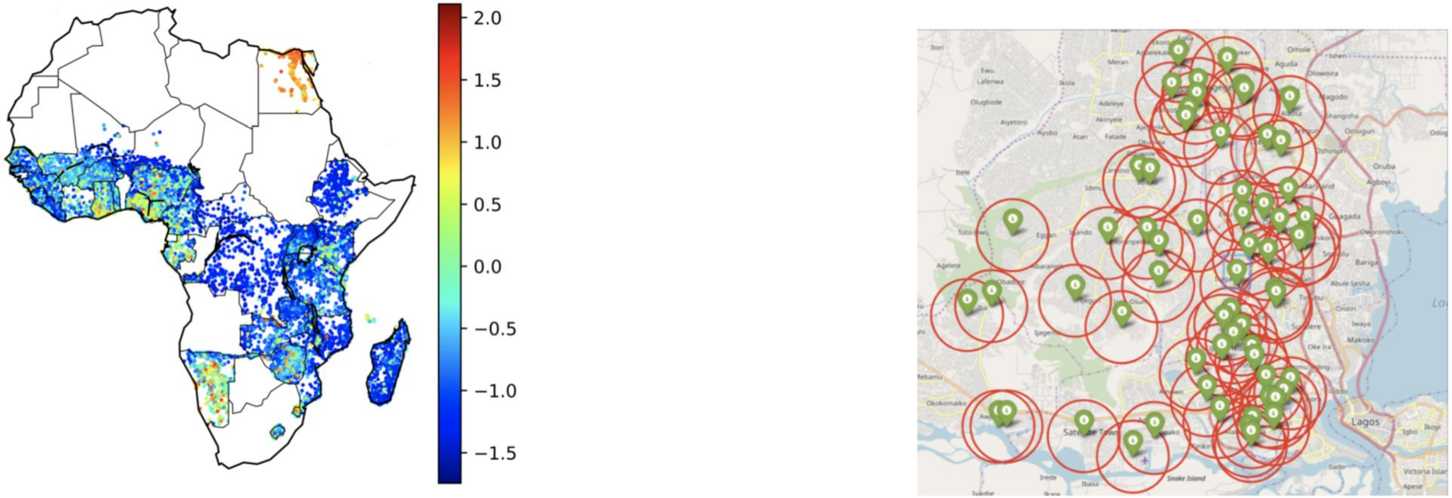


Figure 2: Left: Visualization of ground-truth Asset Wealth Index (AWI) data. Higher values (red) indicate wealthier communities. Right: Jitter in Lagos, Nigeria. Coordinates have up to a 2 km jitter radius in urban areas and 5 km in rural ones.

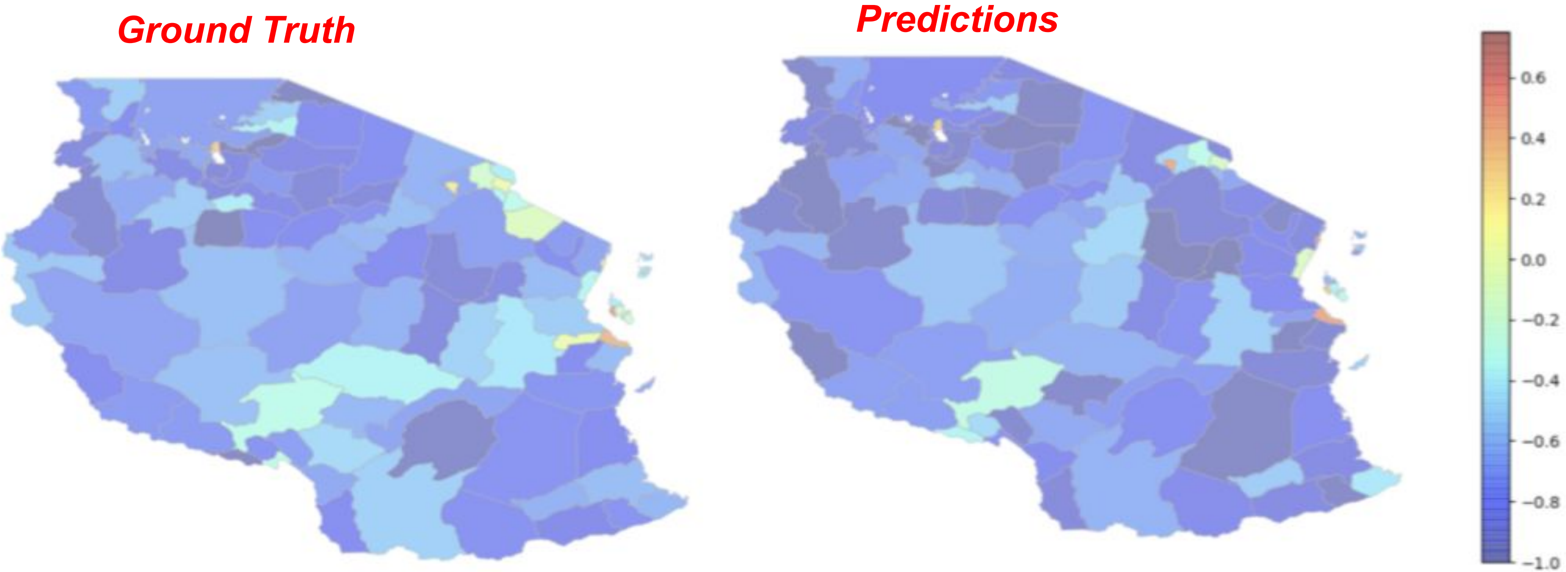
Experiments

- We follow two training strategies to perform experiments in African countries:
  - Train on one country and test on another country
  - Train on all the countries and test on all the countries.

	Trained on																	
	Ghana			Malawi			Nigeria			Tanzania			Uganda			All		
Tested on	NL	WE	MM	NL	WE	MM	NL	WE	MM	NL	WE	MM	NL	WE	MM	NL	WE	MM
Ghana	0.41	0.47	**0.76**	0.43	0.42	**0.61**	**0.64**	0.37	0.45	0.46	0.44	**0.51**	**0.65**	0.34	0.58	**0.61**	0.40	0.60
Malawi	0.30	0.40	**0.48**	0.24	0.49	**0.64**	0.34	0.35	**0.55**	0.37	0.42	**0.56**	0.34	0.25	**0.52**	0.40	0.38	**0.56**
Nigeria	0.44	0.32	**0.60**	0.31	0.37	**0.52**	0.30	0.52	**0.70**	0.46	0.37	**0.57**	0.48	0.24	**0.57**	0.48	0.35	**0.61**
Tanzania	0.50	0.52	**0.58**	0.46	0.52	**0.63**	0.52	0.48	**0.64**	0.60	0.64	**0.71**	0.52	0.49	**0.63**	0.54	0.50	**0.59**
Uganda	0.61	0.45	**0.70**	0.58	0.50	**0.74**	0.62	0.40	**0.70**	0.64	0.49	**0.75**	0.53	0.57	**0.76**	0.62	0.52	**0.71**
All	0.44	0.32	**0.46**	**0.55**	0.26	0.51	**0.51**	0.37	0.48	0.49	0.32	**0.65**	0.46	0.27	**0.48**	0.45	**0.77**	0.76
Average	0.45	0.41	**0.60**	0.43	0.43	**0.61**	0.49	0.42	**0.59**	0.50	0.45	**0.63**	0.50	0.36	**0.59**	0.52	0.49	**0.64**

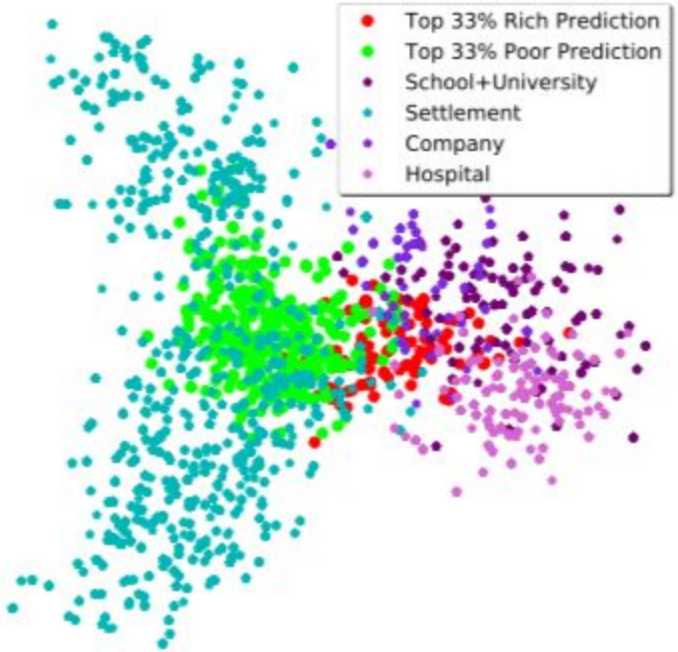
Table 1: Pearson's r^2 values for the Nightlight-Only (NL), Wikipedia Embedding (WE), and Multi-Modal (MM) models on in-country and out-of-country experiments. Columns and rows represent the countries the models were trained and tested on, respectively. The Multi-Modal model outperforms the other models on both in-country (shaded) and cross-country experiments.

Analyzing the Model



Visualization of predictions and ground truth on Tanzania. Lower score represent poor areas.

Analyzing the Predictions



Rich places are projected to latent space closely to School, University, Company and Hospital related articles. Poor places are embedded closely to the Settlement related articles.