Transfer Learning Techniques for Building Damage Assessment

Yandi Wu

ERDC
ENGINEER RESEARCH & DEVELOPMENT CENTER

About my project

Mentor: Charlotte Ellison

Project/Team: Battle Damage Assessment (BDA)

Objectives: Building damage assessment via semantic segmentation of satellite images

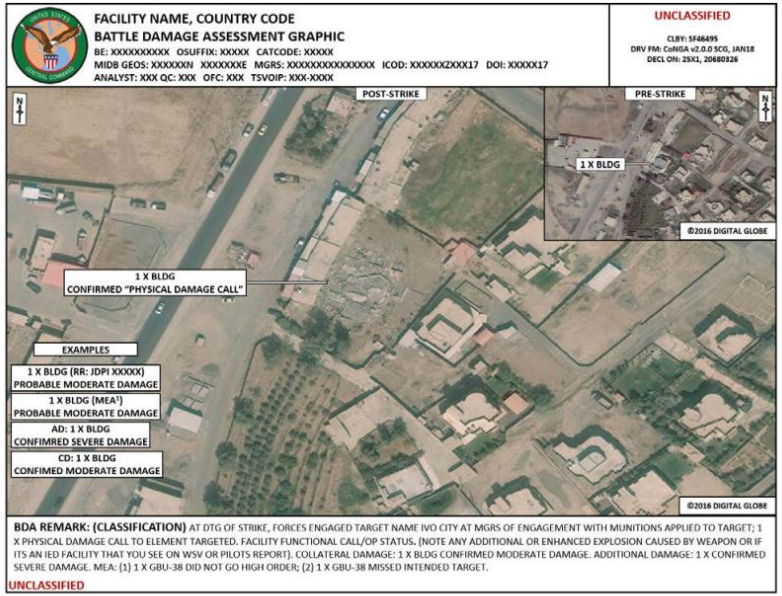


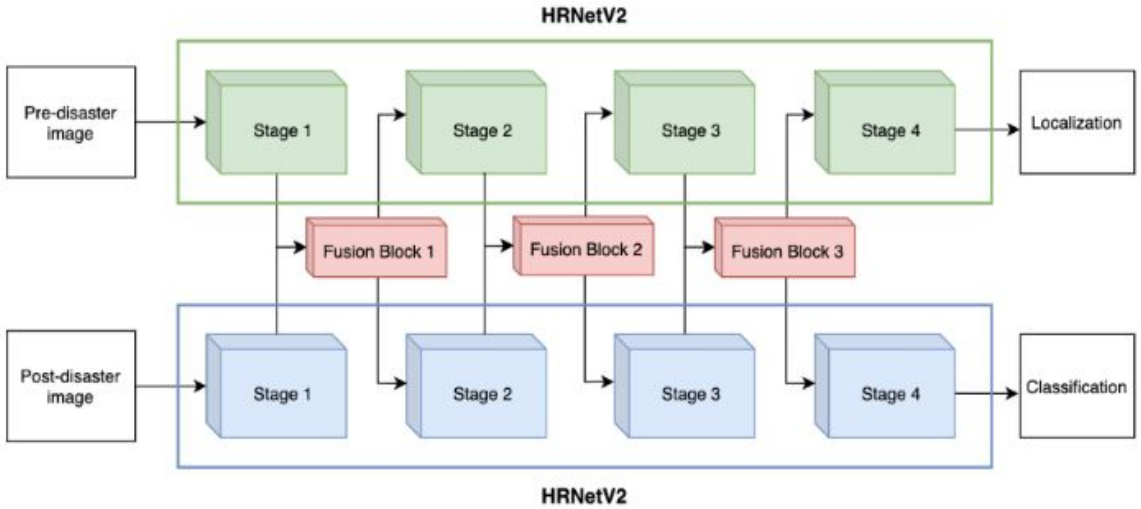
Figure 8. BDA-G Standard

Source: https://www.jcs.mil/Portals/36/Documents/Doctrine/training/jts/cjcsi_3162_02.pdf?ver=2019-03-13-092459-350

Main Tool: Dual HRNet (Koo et al 2019)

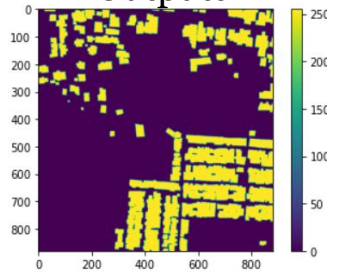
- GPU-powered convolutional neural network (CNN), 5th place winner of xView2 challenge
- Two Inputs: post/ pre disaster images
- Two Outputs: building localization, damage classification
- Benefits: clean code base, only one model (HRNet) in its backbone rather than composite of several models (important because laptop only has one GPU!)

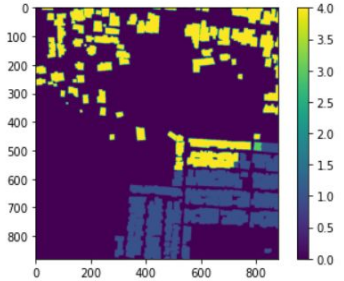
Inputs



Source: https://github.com/SIAnalytics/dual-hrnet

Outputs





Dataset: xBD

- 45,361 sq km of annotated data
- 17,654 post/pre disaster images used for training, 2208 validation, 2206 testing
- 850,736 buildings
- 6 disaster types, 4 damage categories
- Used for xView2 challenge in 2020

Score	Label	Visual Description of the Structure
0	No damage	Undisturbed. No sign of water, structural damage, shingle damage, or burn marks.
1	Minor damage	Building partially burnt, water surrounding the structure, volcanic flow nearby, roof elements missing, or visible cracks.
2	Major damage	Partial wall or roof collapse, encroaching volcanic flow, or the structure is surrounded by water or mud.
3	Destroyed	Structure is scorched, completely collapsed, partially or completely covered with water or mud, or no longer present.

Top: Hurricane Florence
Bottom: Palu tsunami

Source: https://www.ibm.com/cloud/blog/the-xview2-ai-challenge

More realistic training dataset

- Aim: simulate battle setting
- Smaller datasets: 300 training, 70 validation
- Train on densely populated areas (urban areas often targeted)
- Focus on disasters with fire damage (e.g. wildfires and bushfires)
- Choose images with high amounts of damage

Mexico City earthquake: urban setting

Santa Rosa wildfire: fire damage

Socal Fire: fire damage

Test data: 20 most damaged samples, 3 sets of disasters

Fires (top, original test set): pinery bushfire and Woolsey wildfire (Australia), Portugal wildfire

Tornadoes (middle): Joplin (MO), Tuscaloosa (AL), Moore (OK) tornadoes

Water (bottom): Palu tsunami (Indonesia), Nepal flooding, Midwest flooding (not shown)

Generalization Gap

- First row: disasters not in training data; second row: disasters show up in the training data
- Numbers in second row are (generally) higher; need to apply transfer learning techniques to decrease difference
- Columns with most room for improvement are highlighted in yellow

Test Set	Total F1	Damage F1	Loc F1	No Damage F1	Minor Damage F1	Major Damage F1	Destroyed F1
Portugal wildfire, Pinery bushfire, Woolsey Fire (ood)	0.4461	0.3719	0.6192	0.7043	**0.1398**	0.01572	0.6276
Mexico earthquake, Santa Rosa wildfire, SoCal fire (iid)	**0.5384**	**0.4266**	**0.7993**	**0.7938**	0.009785	**0.02173**	**0.8812**

Transfer Learning

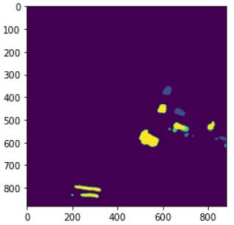
OOD data: Pinery bushfire

Trained on Mexico earthquake, Santa Rosa and SoCal fires (IID data)

Pretrained Baseline Model (Dual HRNet)

Fine-tuning: multiadaBN

New model

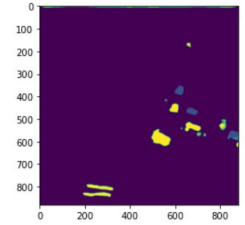


Trained on Mexico earthquake, Santa Rosa and SoCal fires (IID Data)

Pretrained Baseline Model (Dual HRNet)

Fine-tuning: SWA

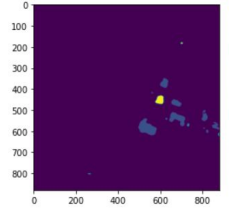
New model



OOD data: Pinery bushfire

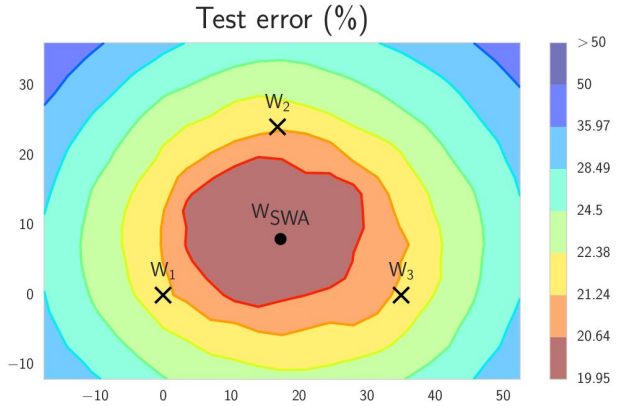
Fine-tuning: multiadaBN

New model



Transfer Learning Techniques

- **Stochastic Weight Averaging (SWA)** (Izmailov et al 2020)**:** Store weights for 40 epochs, then average weights (see picture for illustration). Publicly available code, easy to implement, little computational overhead
- **Multidomain Adaptive Batch Normalization (multiadaBN)** (Schneider et al 2020): Use test statistics for batch normalization layer instead of training statistics; code also publicly available: https://domainadaptation.org/batchnorm/
- **Benson/Ecker 2020**: implemented SWA, multiadaBN with some improvements in overall F1 score (but improvements not clear cut- see table)

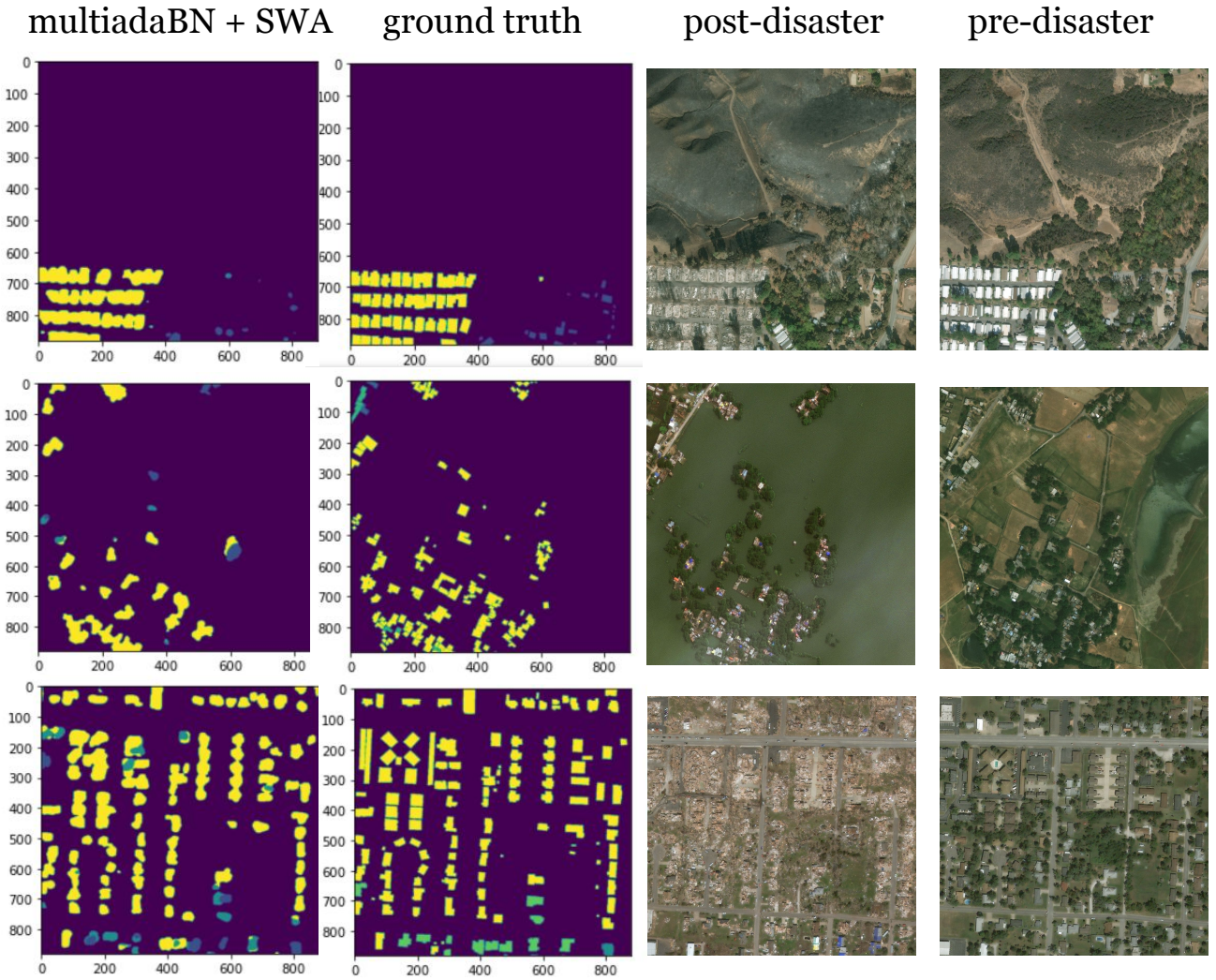


Left: test error loss surface. Source: https://github.com/timgaripov/swa

Model	Gupta & Shah (2020) OOD Score↑	Gap↓	Gain↑	OOD-xBD Score↑	Gap↓	Gain↑
Two-stream ResNet50	0.44	0.30	–	0.60 ±0.01	0.14	–
+SWA	0.50	0.23	0.06	0.62 ±0.01	0.13	0.02
+multi-domain AdaBN	0.52	0.17	0.07	0.62 ±0.05	0.08	0.02
+multi-domain AdaBN +SWA	**0.59**	0.15	0.15	**0.65** ±0.03	0.07	0.05
Dual-HRNet	0.61	0.14	–	0.67 ±0.02	0.06	–
+SWA	0.62	0.13	0.00	0.66 ±0.03	0.05	-0.01
+multi-domain AdaBN	0.67	0.04	0.06	**0.69** ±0.03	0.04	0.02
+multi-domain AdaBN +SWA	**0.68**	0.05	0.07	0.68 ±0.03	0.06	0.01

Results

- Good results for datasets with mostly destroyed/no damage pixels
- Shown: Woolsey fire (top), Nepal flood (middle), Joplin tornado (bottom)



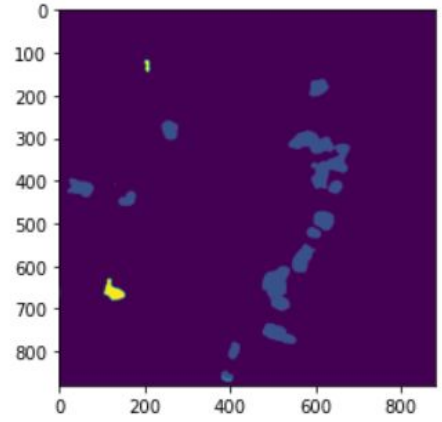
Results

1. Multi adaBN consistently decreases minor damage F1 scores and scores a major damage F1 score of 0, regardless of testing or training set
2. Efficacy of multiadaBN and SWA depends on the disaster type: no clear effect on fire/water disaster F1 scores, appear to improve tornado F1 scores
3. Efficacy of multi adaBN/SWA depends on training set
4. Multi adaBN has more noticeable effect on model than SWA

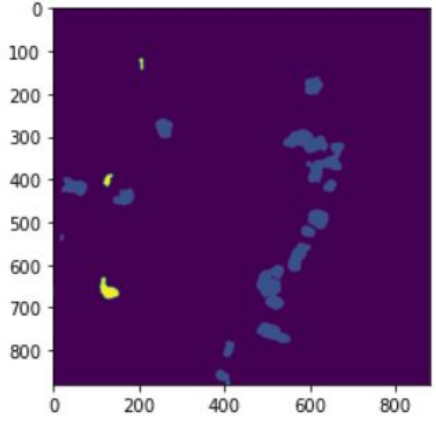
Multiadabn struggles to find major/minor damage

(But is it really a bad thing?)

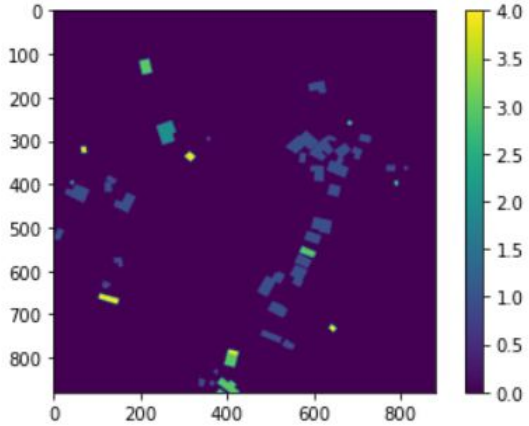
Mult adaBN + SWA



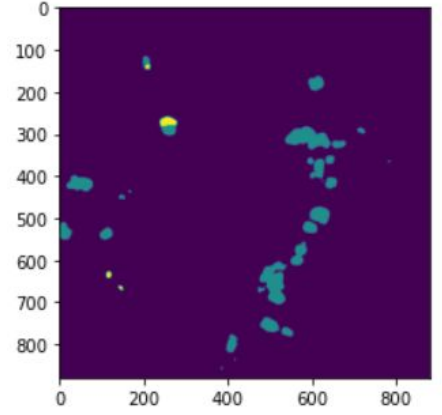
multi adaBN



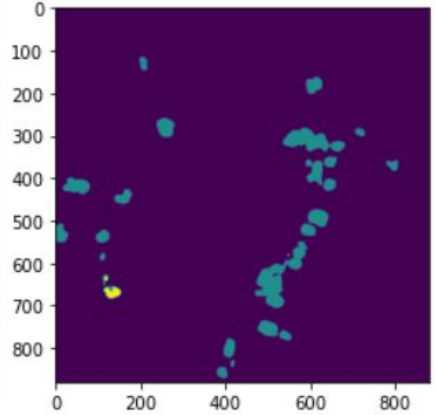
Ground truth



Pre disaster



SWA



Baseline

Destroyed

Major damage

Minor damage

No damage

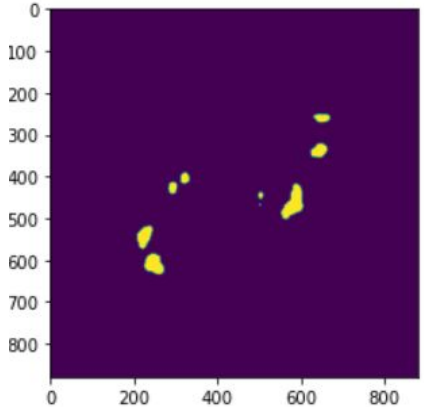
Background

Post disaster

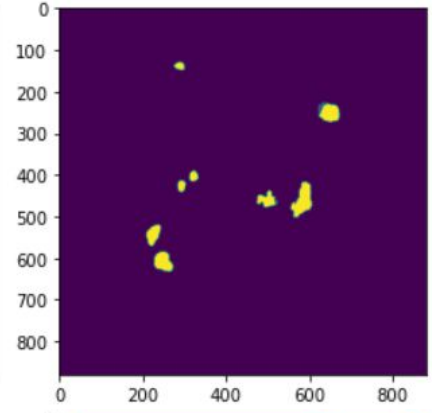
Multiadabn struggles to find major/minor damage

(But is it really a bad thing?)

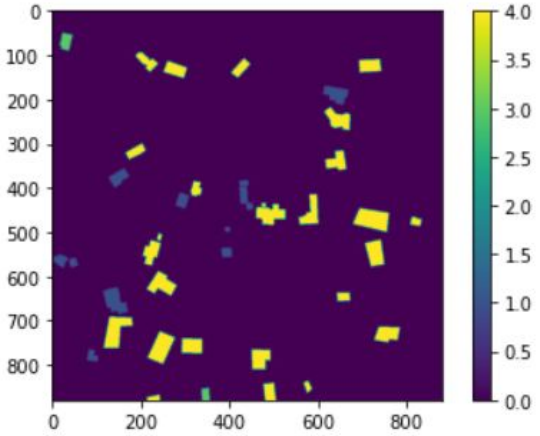
multiadaBN + SWA



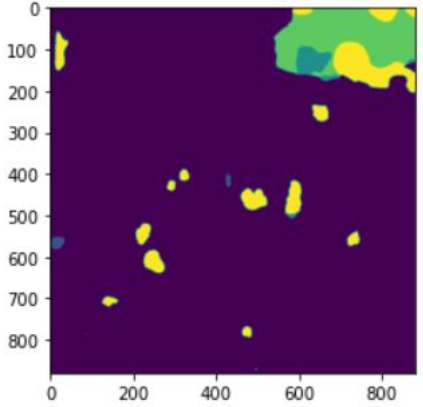
multiadaBN



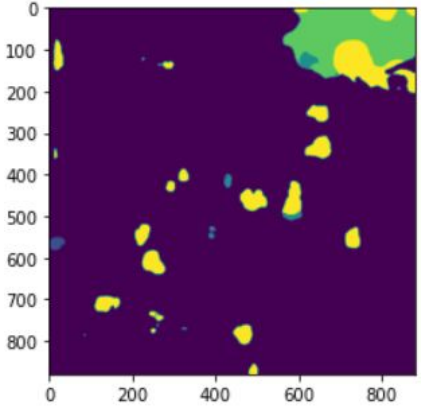
Ground truth



Pre disaster



SWA



Baseline

Destroyed

Major damage

Minor damage

No damage

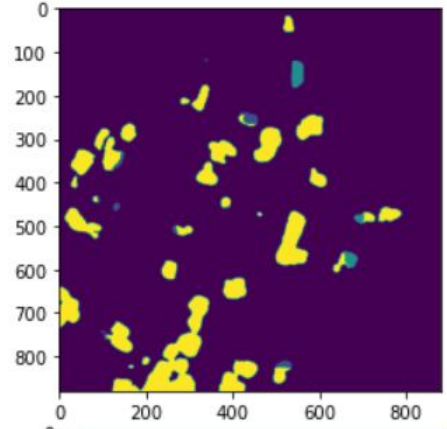
Background

Post disaster

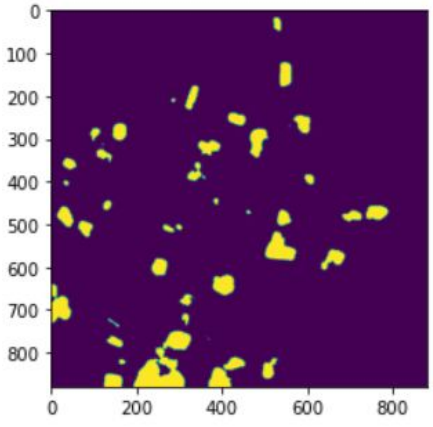
Multiadabn struggles to find major/minor damage

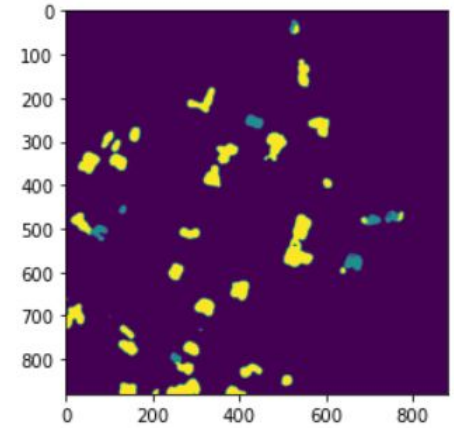
(It could be a bad thing.)

Multi adaBN + SWA

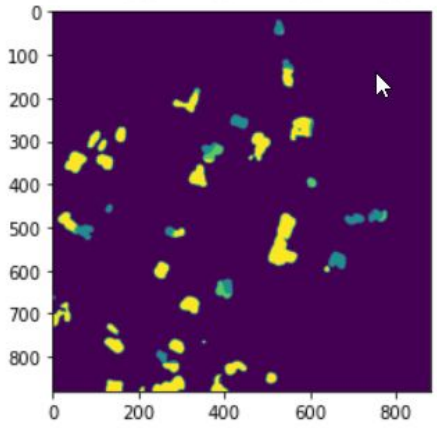


multiadaBN



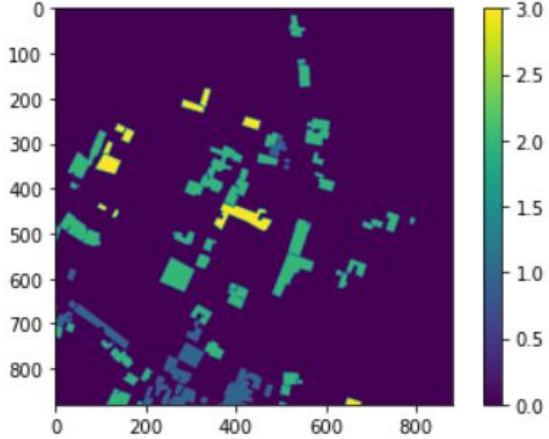


SWA



Baseline

Ground truth



- Destroyed
- Major damage
- Minor damage
- No damage
- Background

Pre disaster

Post disaster

Results: Fire test set (focus of project)

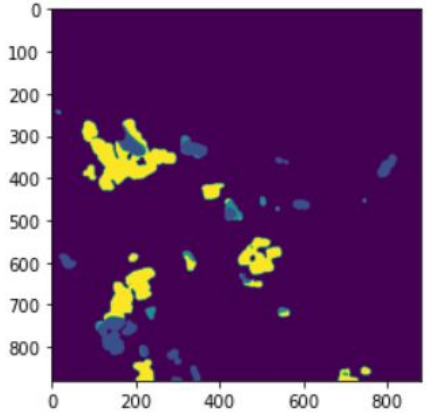
Model	Total F1	Damage F1	Loc F1	No Damage F1	Minor Damage F1	Major Damage F1	Destroyed F1
Baseline (fire)	0.4461	**0.3719**	0.6192	0.7043	0.1398	0.01572	**0.6276**
SWA (fire)	**0.4589**	0.3638	0.5890	0.6807	**0.1491**	**0.01591**	0.6094
Multi adaBN (fire)	0.4049	0.3167	0.6106	**0.7050**	0.03158	0	0.5300
Multi adaBN + SWA (fire)	0.4136	0.3253	**0.6196**	0.7036	0.0538	0	0.5458

Water data: Baseline seems to do the best...

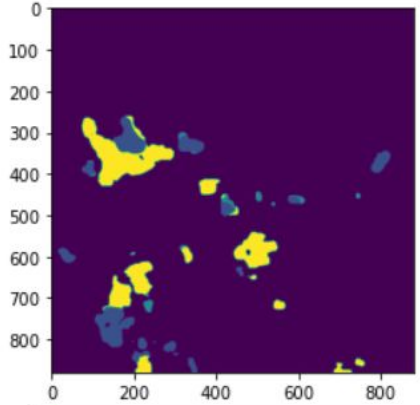
Model	Avg F1	Damage F1	Loc F1	No Damage F1	Minor Damage F1	Major Damage F1	Destroyed F1
Baseline (water)	**0.3288**	**0.2012**	0.6266	**0.5452**	**0.1297**	**0.06262**	0.06726
SWA (water)	0.3109	0.1876	0.5987	**0.5452**	0.09390	0.03066	0.0805
Multi adaBN (water)	0.3190	0.1799	0.6437	0.5222	0.03833	0	0.1591
Multi adaBN + SWA (water)	0.3274	0.1842	**0.6615**	0.5143	0.05441	0	**0.1682**

Water disasters: baseline performed the best

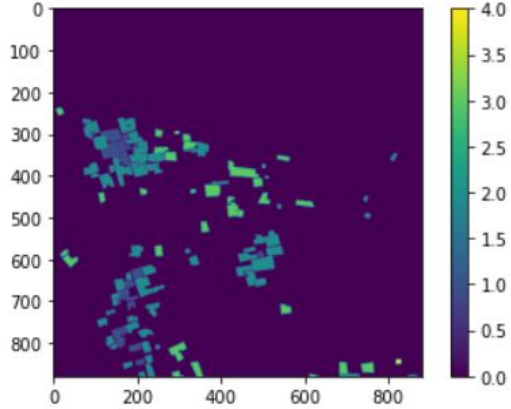
MultiadaBN + SWA



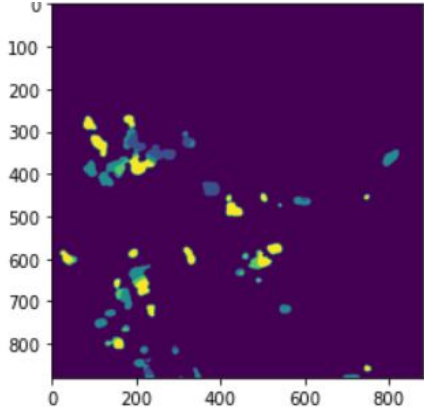
multiadaBN



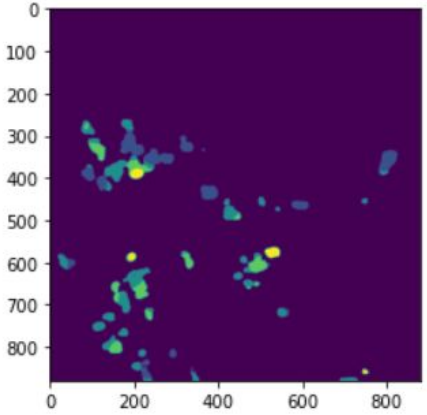
Ground truth



Pre disaster



SWA



baseline

- Destroyed
- Major damage
- Minor damage
- No damage
- Background

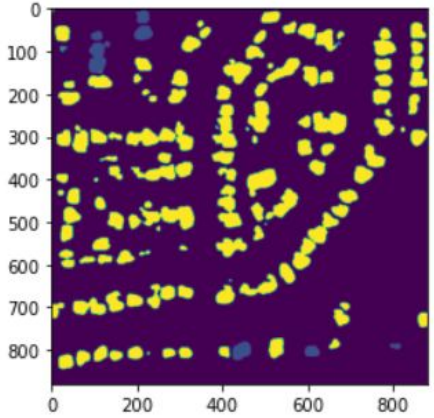
Post disaster

Tornadoes: SWA, Multi adaBN appear to help

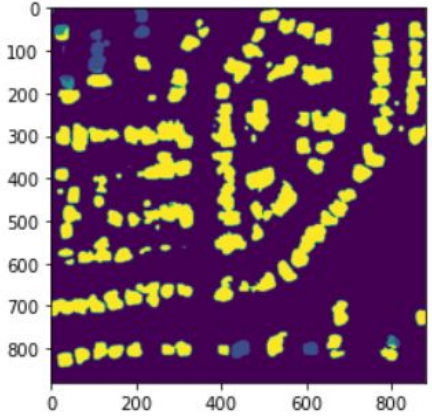
Model	Total F1	Damage F1	Loc F1	No damage F1	Minor Damage F1	Major Damage F1	Destroyed F1
Baseline (tornado)	0.4086	0.2646	0.7445	0.3477	0.01379	**0.1612**	0.5357
SWA (tornado)	0.4353	0.3073	0.7340	0.3566	0.08186	0.1277	0.6629
Multi adaBN (tornado)	0.4413	0.3211	0.7218	0.4179	0.1036	0	0.7630
Multi adaBN, SWA (tornado)	**0.4610**	**0.3368**	**0.7504**	**0.4515**	**0.1281**	0	**0.7683**

MultiadaBN, SWA help with tornado data

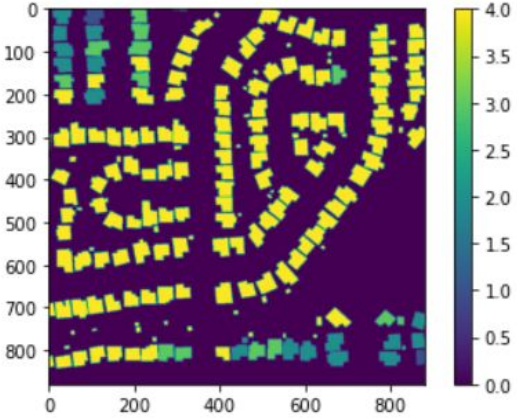
multiadaBN + SWA



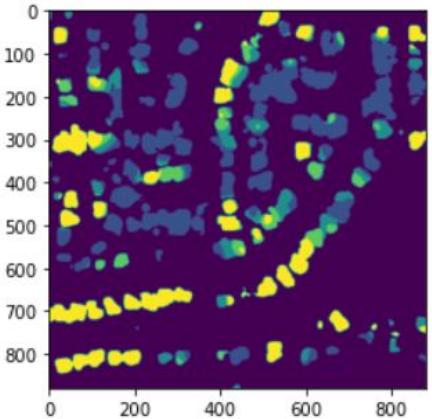
multi adaBN



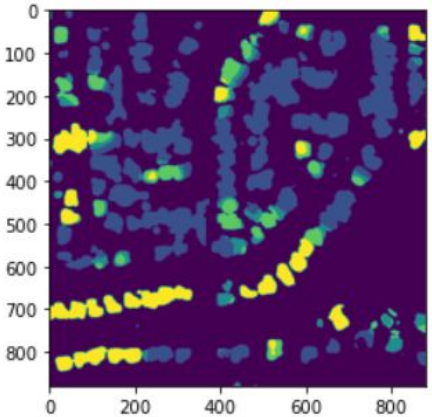
ground truth



Pre disaster



SWA



baseline

Destroyed

Major damage

Minor damage

No damage

Background

Post disaster

Comparing total F1 scores across disasters:

- Baseline + SWA perform best on fire data
- multiadaBN models perform best on tornado data
- Water data does the worst (higher amounts of minor/major damage, OOD problem)
- Fire and tornado results are comparable

Disaster	Baseline	SWA	MultiadaBN	MultiadaBN +SWA
Fire	**0.4461**	**0.4589**	0.4049	0.4136
Water	0.3288	0.3109	0.3190	0.3274
Tornado	0.4086	0.4353	**0.4413**	**0.4610**

New Problem: Data Imbalance

- Data imbalance problem: minor/major damage results are far lower than no damage/damage F1 scores since they are less common
- One approach: create a more balanced training dataset with more examples of major/minor damage
- Pixel count data for the two training datasets:

	No Damage	**Minor Damage**	**Major Damage**	**Destroyed**	**Total**
Most Damaged Dataset	5842894 (53.07%)	200768 (1.82%)	193245 (1.76%)	**4773680** (43.36%)	11010587
Balanced Dataset	**10053437** (67.4%)	**270168** (1.81%)	**217187** (1.46%)	4368461 (29.3%)	**14543366**

Results: Fire test set, balanced training set

- Multi adaBN and SWA appear to improve results
- Overall, model did not perform as well even though training sets had 108 images in common and more examples of minor/major damage (see numbers in parenthesis)

Model	Total F1	Damage F1	Loc F1	No damage	Minor Damage	Major Damage	Destroyed
Baseline	0.3565 (vs. 0.4461)	0.3562	0.3573	0.6043	**0.1581**	0	0.6624
SWA	0.3804 (vs. 0.4589)	0.3601	0.4276	0.6537	0.1316	0	0.6550
multi adaBN	0.3648 (vs. 0.4049)	0.3070	**0.4995**	0.5783	0.0014018	0	0.6483
Multi adaBN + SWA	**0.4131** (vs. 0.4136)	**0.3950**	0.4552	**0.7705**	0.09085	0	**0.7188**

Conclusion: Data is messy!

Original question: Are multiadaBN and SWA effective tools for transfer learning?

Answer: It depends.

- Yes for some disasters (e.g. tornado), no for others (e.g. water), inconclusive for others (e.g. fires)
- Yes for some training sets (e.g. balanced training set), no for others (e.g. most damaged training set)

Other observations:

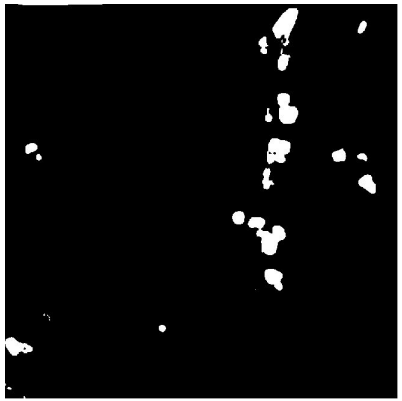
- Multi adaBN affects the models more than SWA does
- Multi adaBN is not effective for categories the model does not classify well (major, minor damage)- need to fix data imbalance problem

Challenges

- Debugging is a very time consuming process, and publicly available packages are full of bugs!
- Limited GPU availability: needed to crop test images, limit training size and batch size/number of workers
- Real world data is very messy and hard to draw conclusions from

Cropped training data

uncropped training data



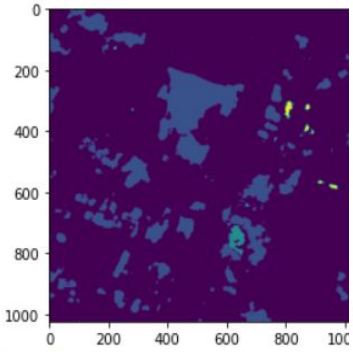
ground truth

Cropping the training data severely affected training accuracy.

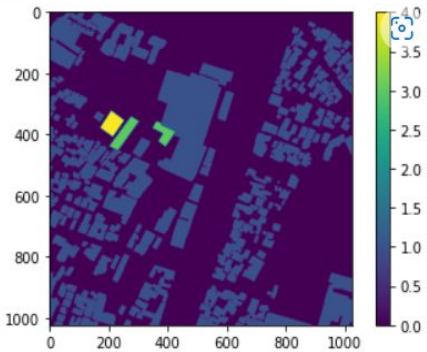
Other Accomplishments

- Compiled document summary of transfer learning techniques with focus on ones with publicly available code base
- Explored other backbones: ResNet50, UNet (via FastAI); these did not perform as well
- Successfully ran CNN with Max. Square Loss (loss function designed for transfer learning) on GTAV/Cityscapes datasets
- Visited GRL to attend annual picnic!

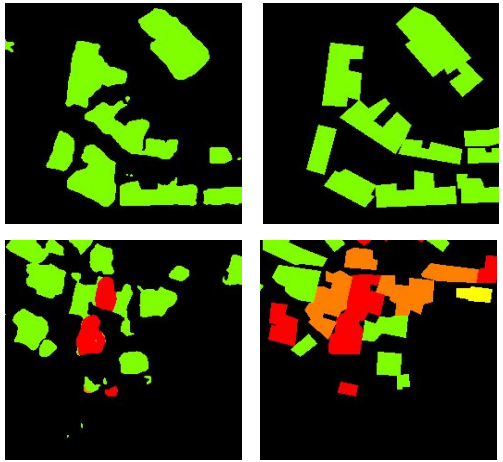
FastAI prediction

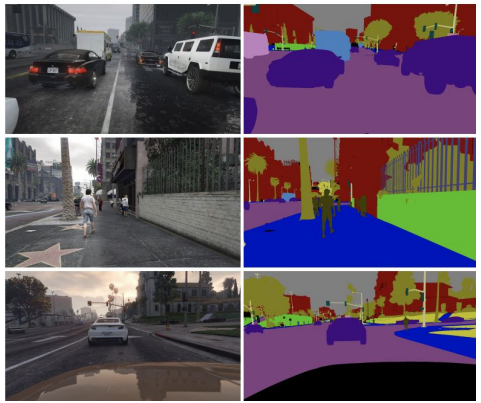


Ground Truth



ResNet50 prediction Ground Truth





Future Work for BDA

- Still unclear why multidomain adaBN, SWA are effective in some settings but not others- more experiments are necessary
- Figure out ways to mitigate data imbalance problem (common for smaller, and more realistic datasets)
- Try other loss functions, e.g. Maximum Square Loss (used on the GTAV/Cityscapes transfer learning problem): https://github.com/ZJULearning/MaxSquareLoss
- Generative Adversarial Networks (GANs): https://github.com/wasidennis/AdaptSegNet (model winning 3rd place in VisDA challenge)
- Write a survey of transfer learning techniques applied to semantic segmentation, compare efficacy of various methods

Acknowledgements

- **Charlotte Ellison**
- Nikki Wayant
- Ray Dos Santos
- Jennifer Smith
- Matt Reichenbach
- NSF