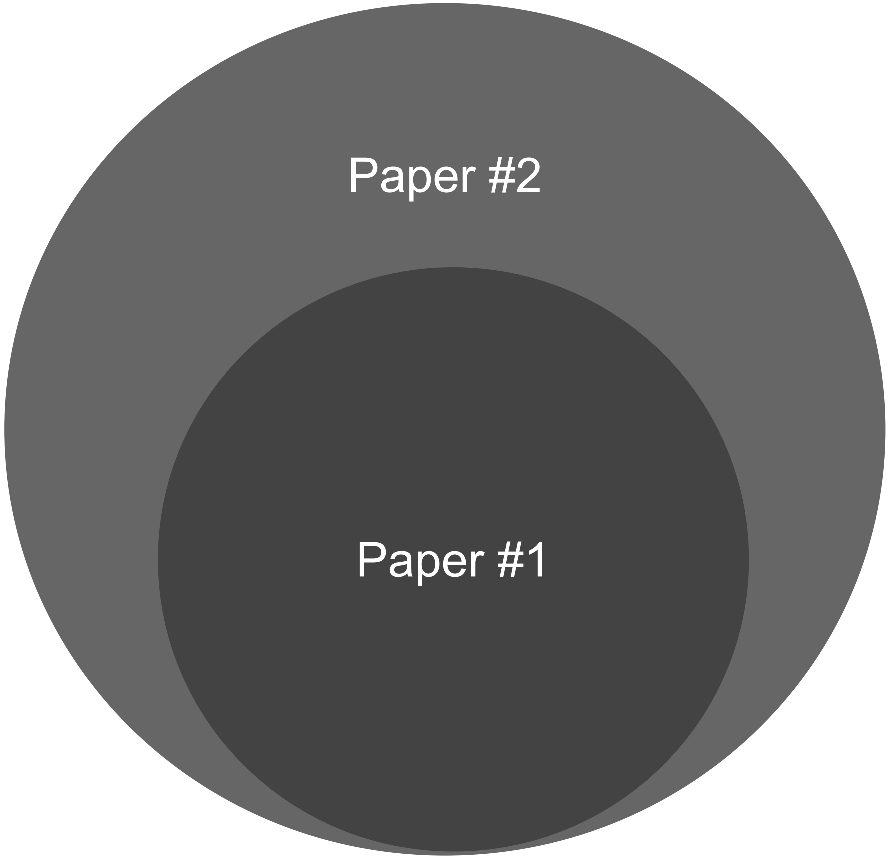
MMU
MULTIMEDIA UNIVERSITY

REVIEW TO:

RAPID: Rating Pictorial Aesthetics using Deep Learning (Paper #1)
Rating Image Aesthetics Using Deep Learning (Paper #2)

Authors: Xin Lu, Zhe Lin, Hailin Jin, Jianchao Yang, James Z. Wang

By: Kairanbay Magzhan



~90%
same

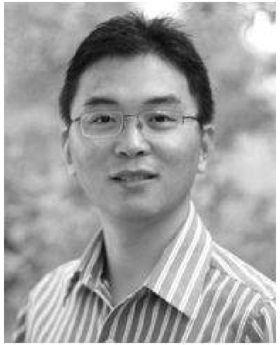
Authors

Xin Lu

PhD student at the College of Information Science and Technology, The Pennsylvania State University, USA

Zhe Lin

Senior Research Scientist at Adobe Research, USA. PhD from University of Maryland, USA



Hailin Jin

Principal Scientist at Adobe Systems Inc., Postdoctoral Researcher from University of California, USA

Jianchao Yang

Researcher scientist at Adobe Technology Laboratory, USA. PhD from University of Illinois at Urbana-Champaign, USA

James Z. Wang

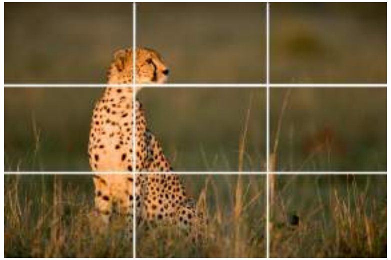
PhD from Stanford University, USA. Professor and the Chair of Faculty Council at the College of Information Science and Technology, The Pennsylvania State University, USA

Handcrafted features

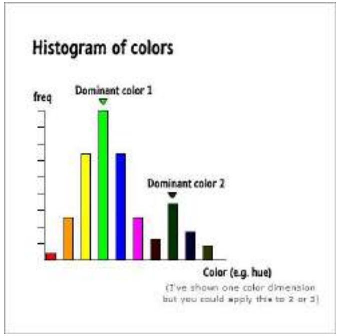
Low level

High level

Edge distribution



Rule of third



Color histogram

Golden ratio

Handcrafted features

- Some aesthetic-relevant attributes may be unexplored and thus poorly defined
- Vagueness of certain photographic or psychologic rules. Difficulty in implementing them computationally, these handcrafted features are often merely approximations of such rules.

Comparison

Generic features (SIFT, Fisher Vector etc.) > Handcrafted features (Color histogram, rule of third etc.)

Motivation

*"Image aesthetics depends on on a combination of **local cues** (e.g. sharpness and noise level) and **global cues** (e.g. the rule of third)"*

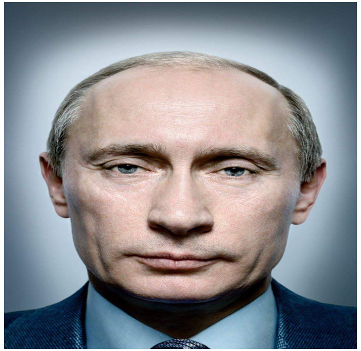
Square

cnn

landscape

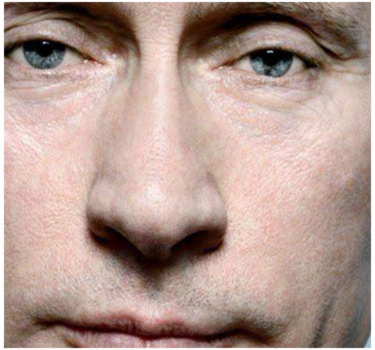
portrait

Resize (Warp)

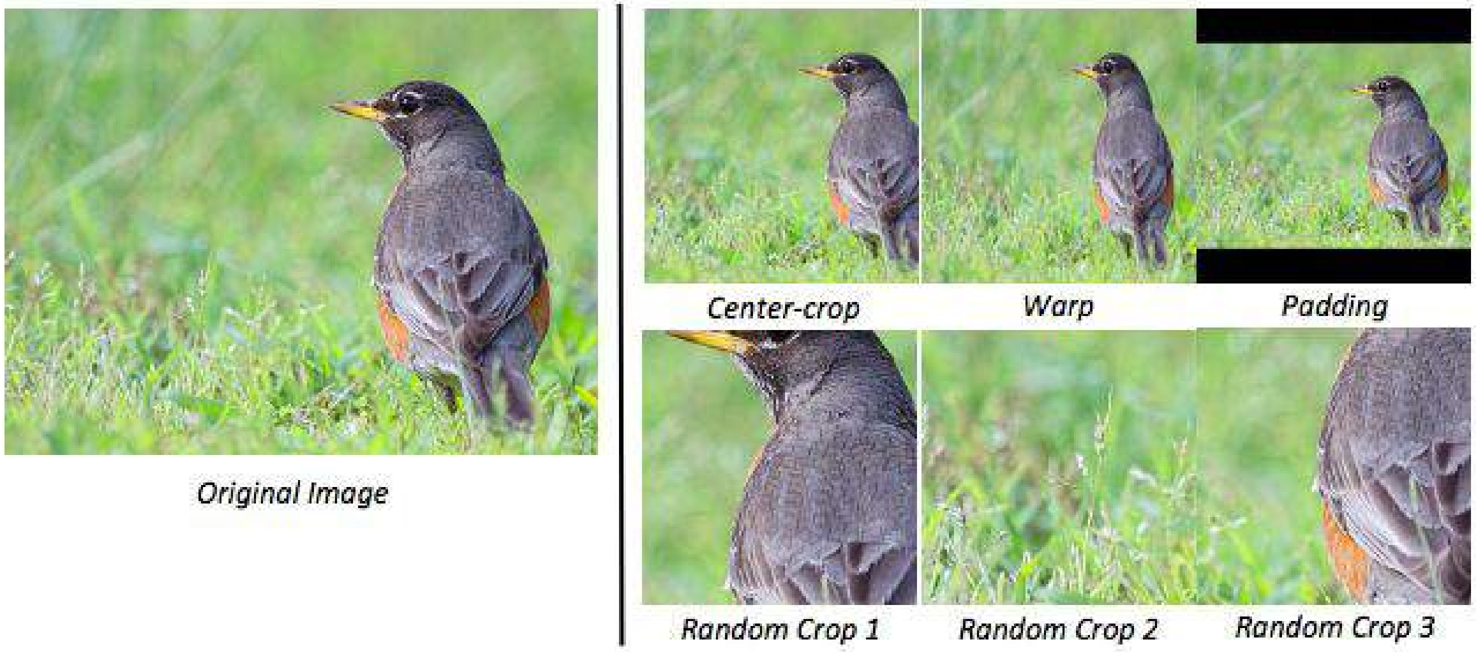


Normalized but lost important information

Random crop

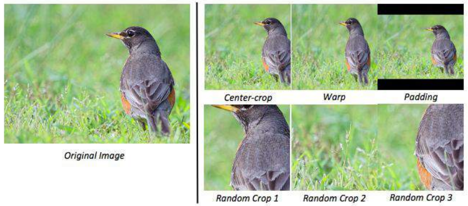


Normalized but lost important information

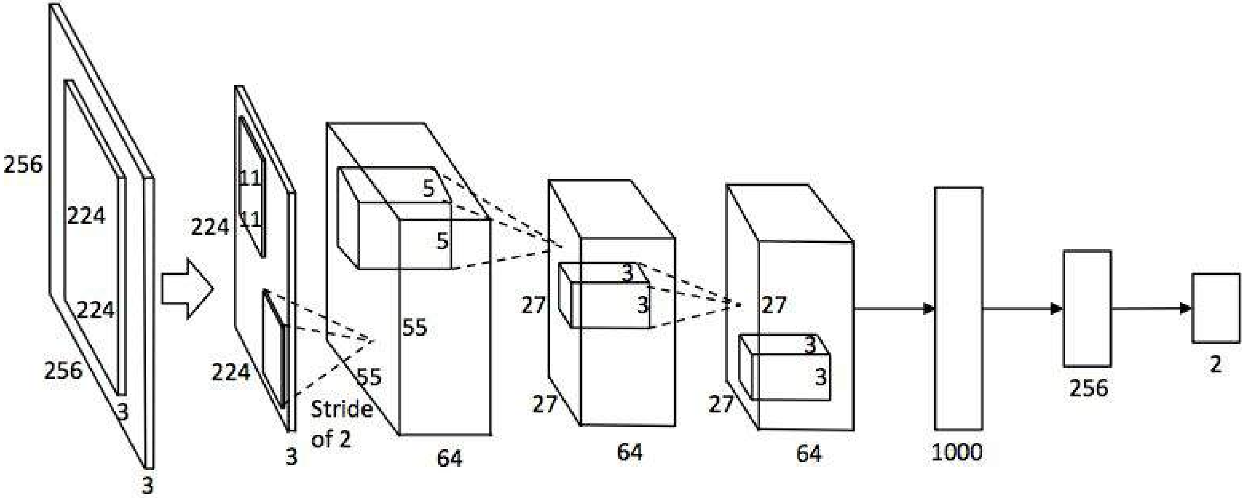


Global views

Local views



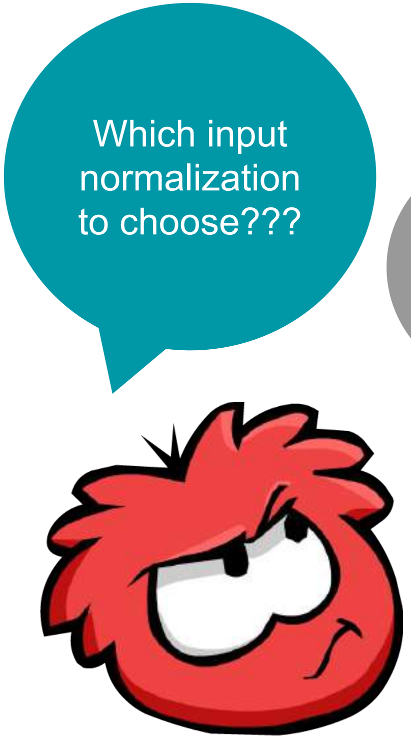
Data augmentation



SCNN architectures

Table 1: Accuracy for Different SCNN Architectures

	conv1 (64)	pool1	rnorm1	conv2 (64)	pool2	rnorm2	conv3 (64)	conv4 (64)	conv5 (64)	conv6 (64)	fc1K	fc256	fc2	Accuracy
Arch 1	√	√	√	√	√	√	√	√			√	√	√	**71.20%**
Arch 2	√	√	√	√	√	√					√	√	√	60.25%
Arch 3	√	√	√	√	√	√	√				√	√	√	62.68%
Arch 4	√	√		√	√		√	√	√	√	√	√	√	65.14%
Arch 5	√	√	√	√	√	√	√	√				√	√	70.52%
Arch 6	√	√	√	√	√	√	√	√	√		√	√	√	62.49%
Arch 7	√	√	√	√	√	√	√	√	√	√	√	√	√	70.93%



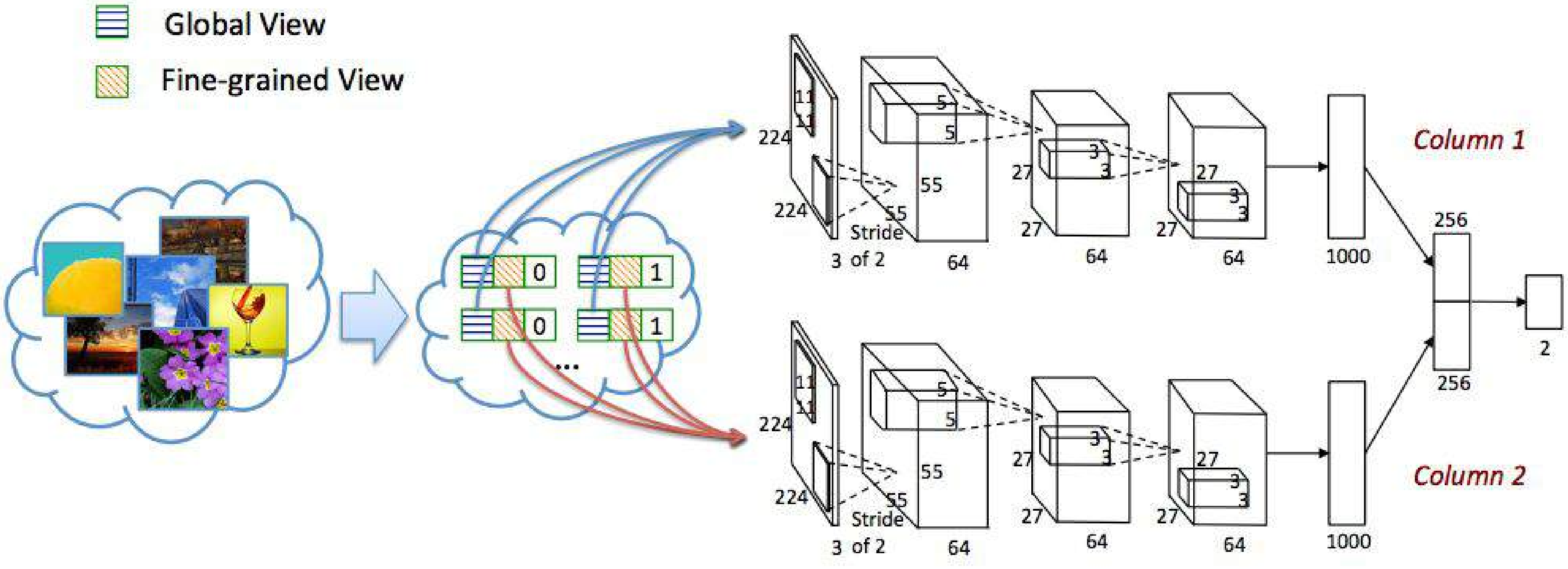
60.43%	65.48%		67.79%		71.2%
Padding	Center crop	<	Warp	<	Random crop

Global view (Padding, Center crop, Warp)

Local view (Random crop)

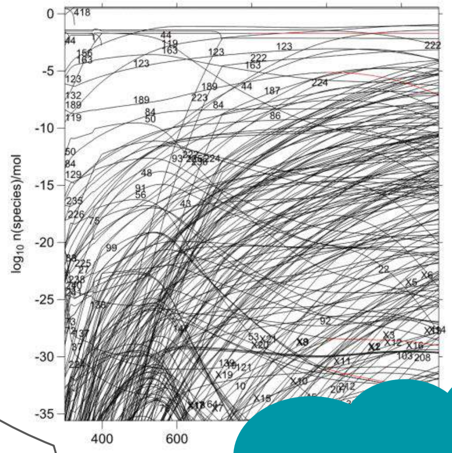
Parallelization

DCNN (Double column CNN)



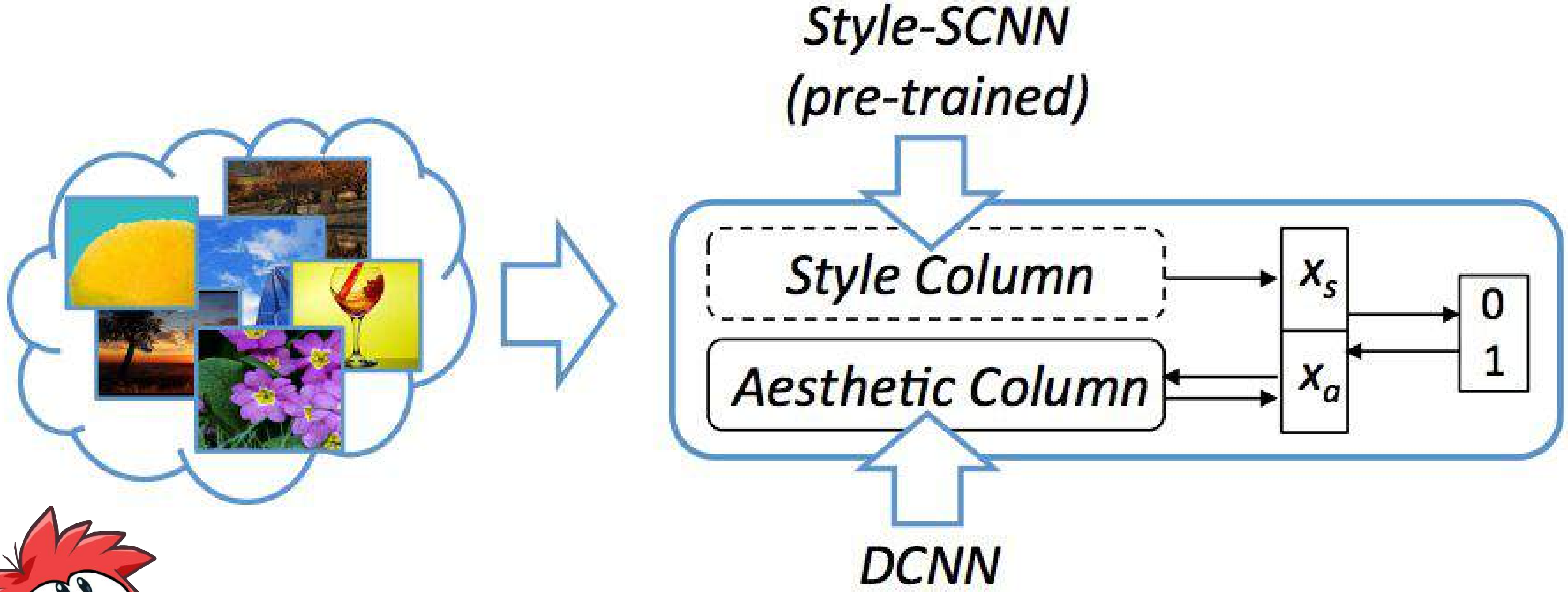
AVA provides semantic and style tag. Can these information improve the accuracy???

What if we train individual DCNN for each semantic category???



Nope, this is time consuming

RDCNN (Regularized double column CNN)



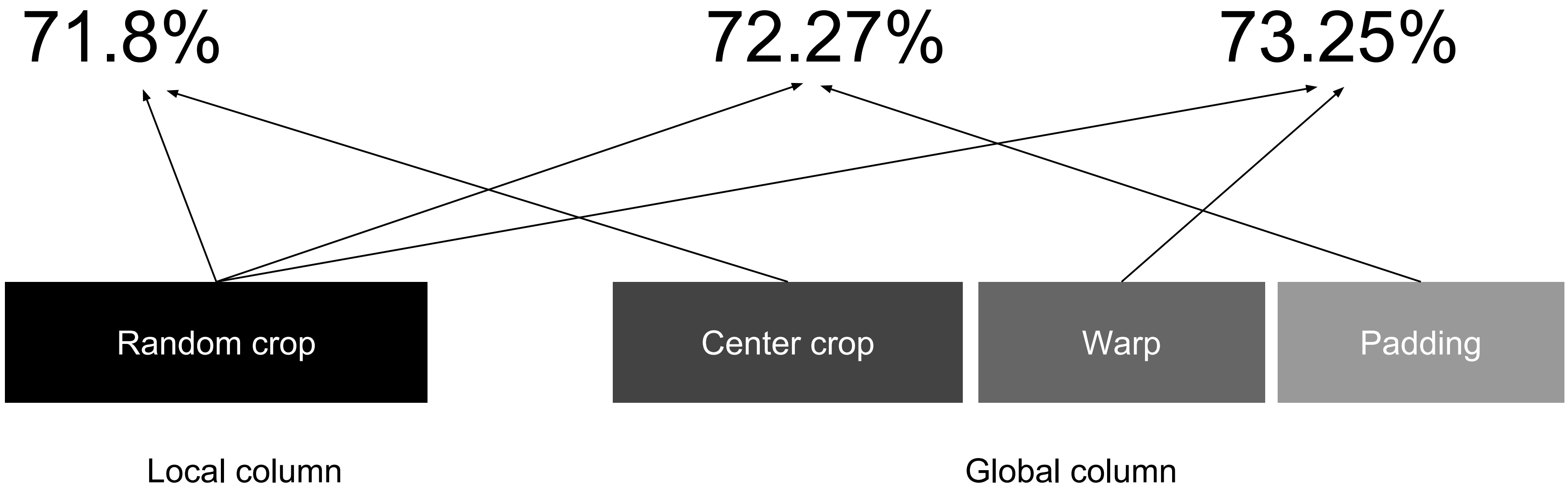
Remarks for RDCNN

- Only **1.4K** images out of **230K** images in AVA dataset contains the **style labels**
- Due to small number of training example, the number of filters are reduced by half in Style-SCNN
- Style attributes are extracted from **fc256** in Style-SCNN
- The input for style column is random crop image
- Fine tune only the parameters in aesthetic column in backpropagation
- Learning process is supervised by aesthetic label only
- Due to small number of training sample for semantic tag, the **Image Net** model used as pre-trained column

Results (SCNN)

δ	I_l^r	I_g^w	I_g^c	I_g^p
0	71.20%	67.79%	65.48%	60.43%
1	68.63%	68.11%	69.67%	70.50%

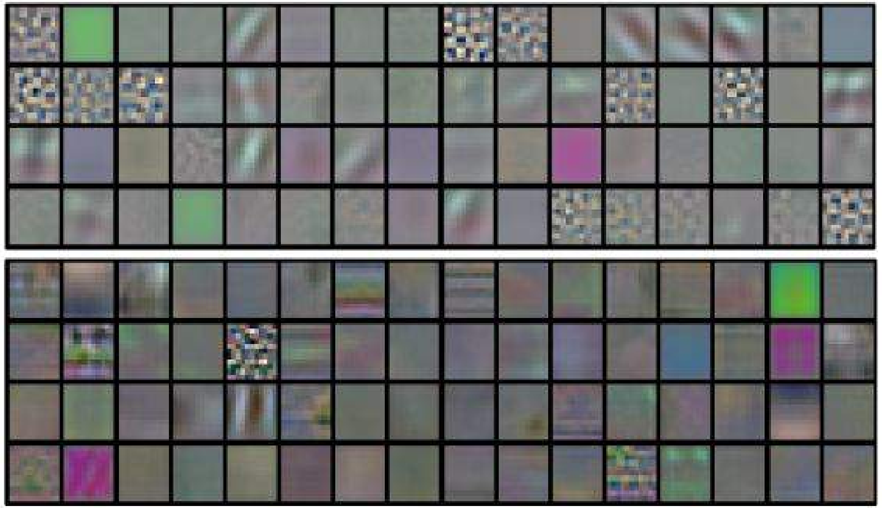
Results (DCNN)



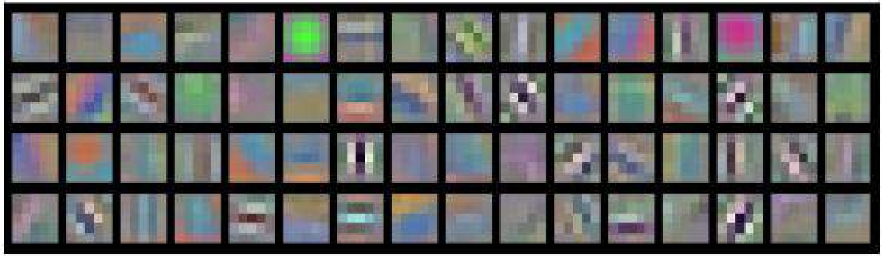
Results

δ	[10]	SCNN	AVG_SCNN	DCNN	RDCNN_{style}	$\text{RDCNN}_{semantic}$
0	66.7%	71.20%	69.91%	73.25%	74.46%	75.42%
1	67%	68.63%	71.26%	73.05%	73.70%	74.2%

- AVG_SCNN = the average result of SCNN where input was random crop and warp



Aesthetic filters



CIFAR filters

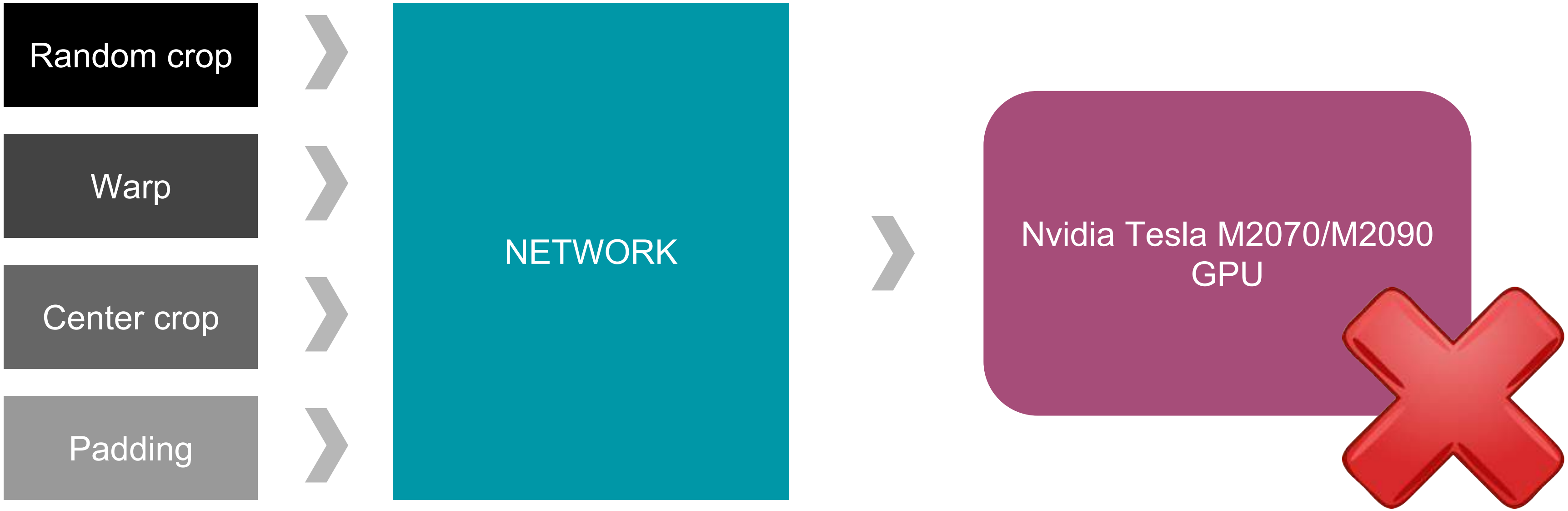
Aesthetic filters look smoother and cleaner!

Correctly classified by DCNN but misclassified by SCNN

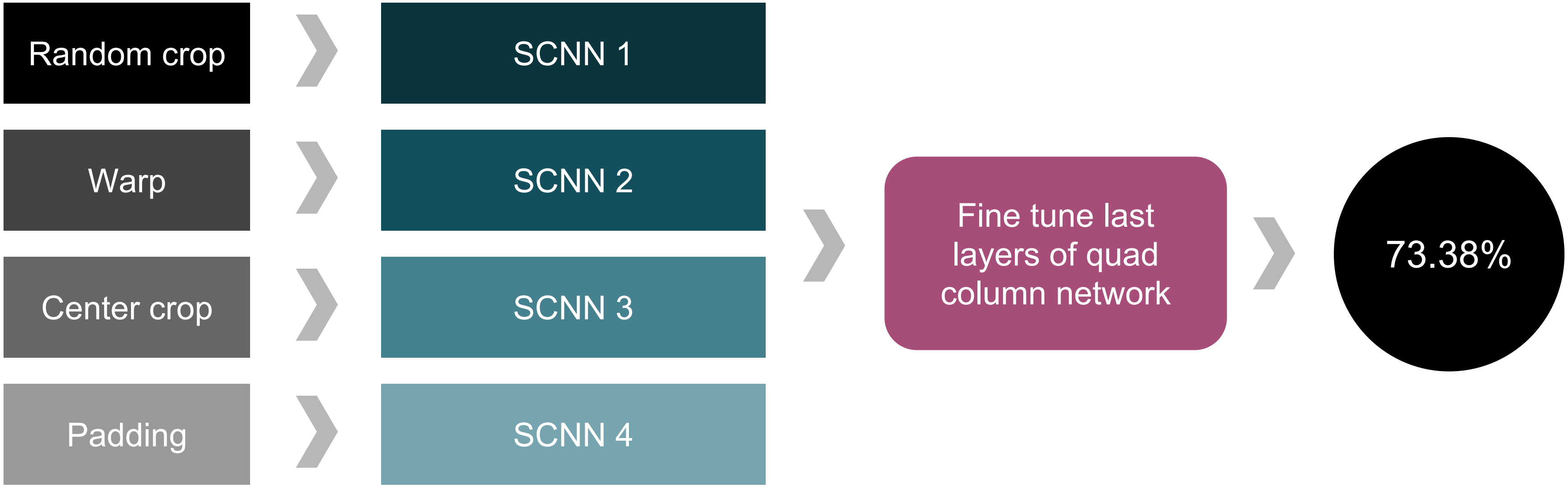
Input as **local** random crop. Most of misclassified images dominated by an **object**

Input as **global** warp. Most of misclassified images dominated by **fine-grained details**

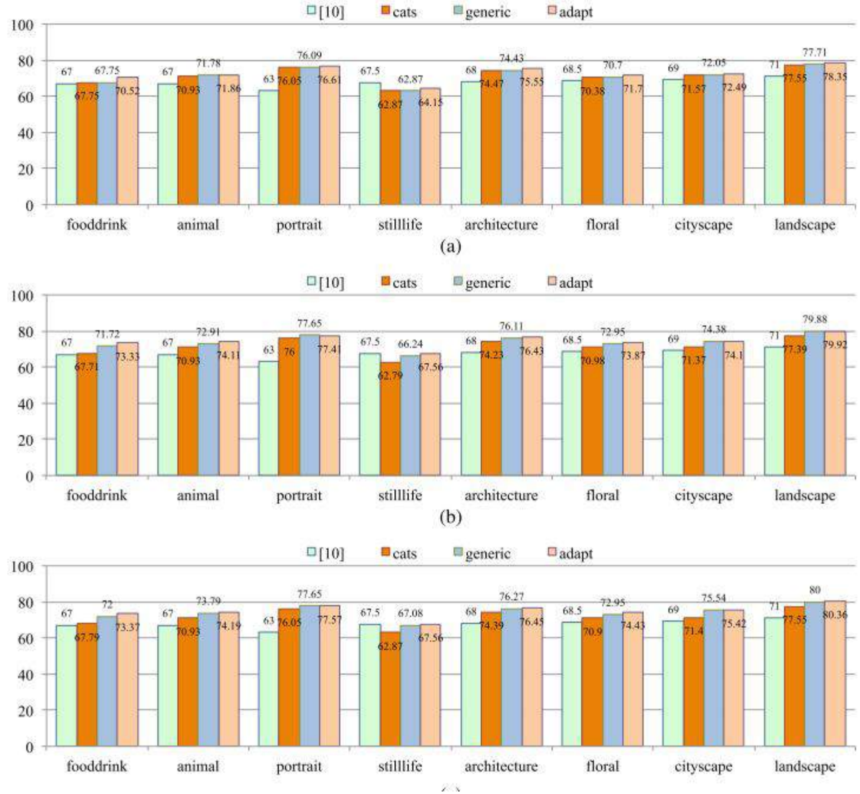
Attempt to use quad-column network...



Attempt to use quad-column network...



Content based image aesthetics



- Cat = trains network using categorized images
- Generic = trains network using AVA training set
- Adapt= proposed network adaptation approach

- once an image is associated with an obvious semantic meaning, then the global view is more important than the local view in terms of assessing image aesthetics
- global view and the local view contribute to the aesthetic quality categorization of content-specific images.

Style-CNN

	conv1 (32)	pool1	rnorm1	conv2 (64)	pool2	rnorm2	conv3 (64)	conv4 (32)	conv5 (32)	conv6 (32)	fc1K	fc256	fc14	mAP	Accuracy
Arch 1	√	√	√	√	√	√	√	√			√	√	√	**56.81%**	**59.89%**
Arch 2	√	√	√	√	√	√					√	√	√	52.39%	54.33%
Arch 3	√	√	√	√	√	√	√				√	√	√	53.19%	55.19%
Arch 4	√	√		√	√		√	√	√	√	√	√	√	54.13%	55.77%
Arch 5	√	√	√	√	√	√	√	√				√	√	53.94%	56.00%
Arch 6	√	√	√	√	√	√	√	√	√		√	√	√	53.22%	57.25%
Arch 7	√	√	√	√	√	√	√	√	√	√	√	√	√	47.44%	52.16%

Style-CNN

	I_l^r	I_g^w	I_g^c	I_g^p
AP	56.93%	44.52%	45.74%	41.78%
mAP	56.81%	47.01%	48.14%	44.07%
Accuracy	59.89%	48.08%	48.85%	46.79%

- mAP - mean Average Precision
- Average Precision

Style-CNN

Correctly classified by RDCNN$_{style}$ but misclassified by DCNN

- Rule-of-thirds
- HDR
- black and white
- long exposure
- complementary colors
- vanishing point
- soft focus

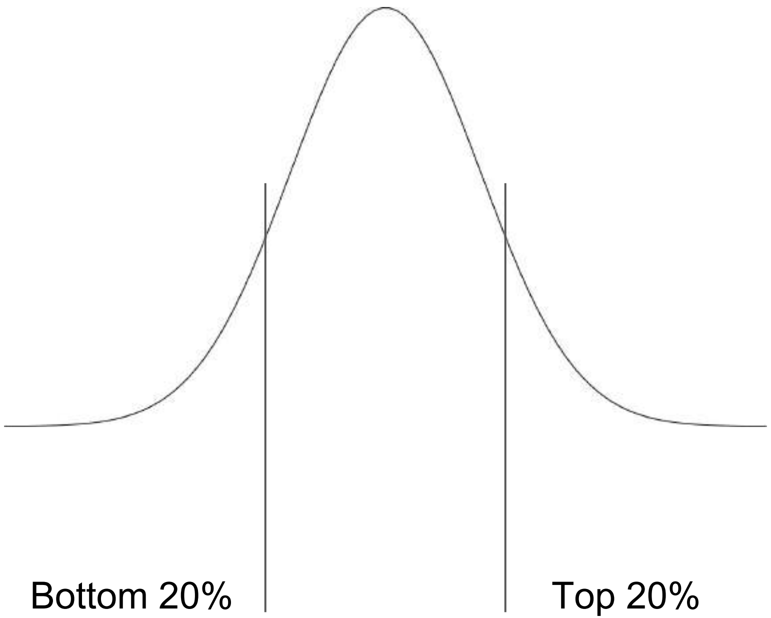

1.2 million
PHOTO.NET
300K
DPChallenge
1.5 million
IAD (Image Aesthetics Dataset)

Results on IAS dataset

SCNN (random crop)	73.21%
SCNN (warp)	73.65%
DCNN (random crop & warp)	74.6%

Results on IAS dataset

SCNN (random crop)	72.11%
SCNN (warp)	72.65%
DCNN (random crop & warp)	72.9%



- AVA test set contains images with rating in the middle
- Utilizing top and bottom 20% of images reduced the number of training data

Future studies (IAD)

- Aperture/FNumber
- ISO/ISOSpeedRatings
- Shutter/ExposureTime
- Lens/FocalLength

Details and tools

- ConvNet
- Logistic Regression cost function
- Nvidia Tesla M2070/M2090 GPU

Time

Style SCNN	SCNN	DCNN	RDCNN	SCNN
1 day	2 day	3 day	3-4 day	4 day
AVA	AVA	AVA	AVA	IAD

thank you

terima kasih

ধনয্বাদ

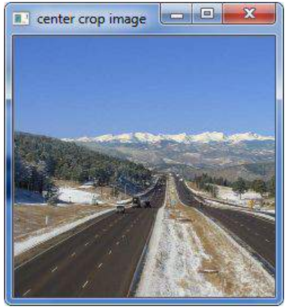
謝謝

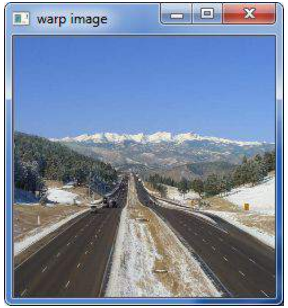
धन्यवाद

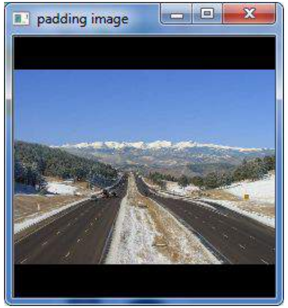
Рахмет

So, let’s start to discuss?

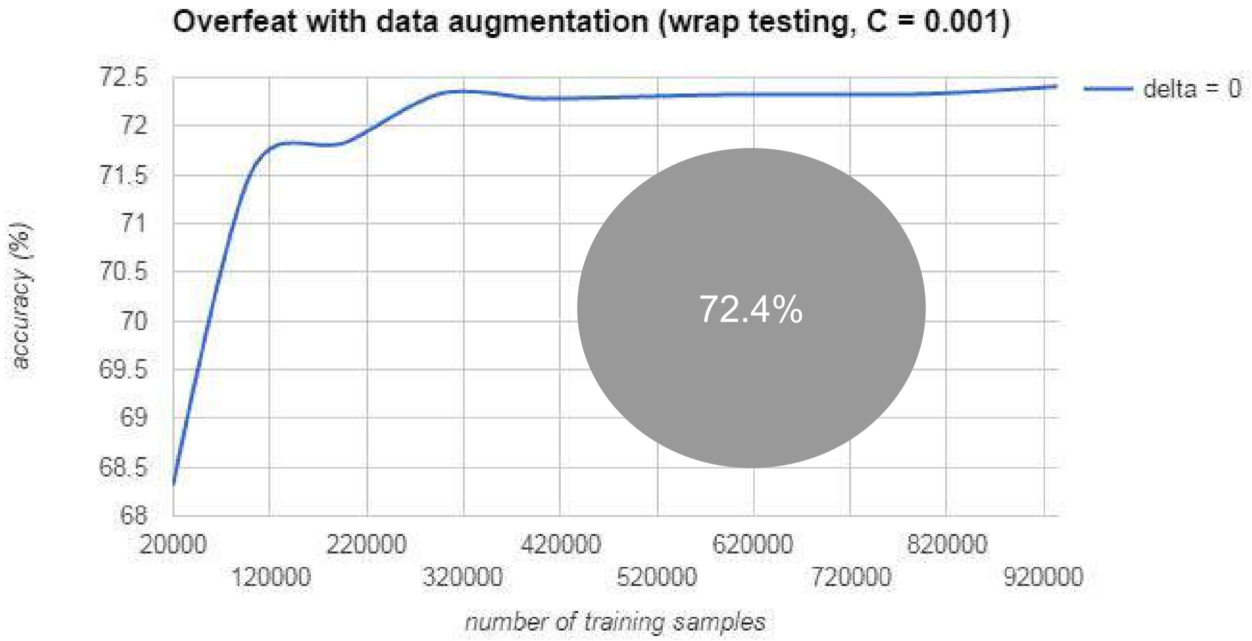
My experiments



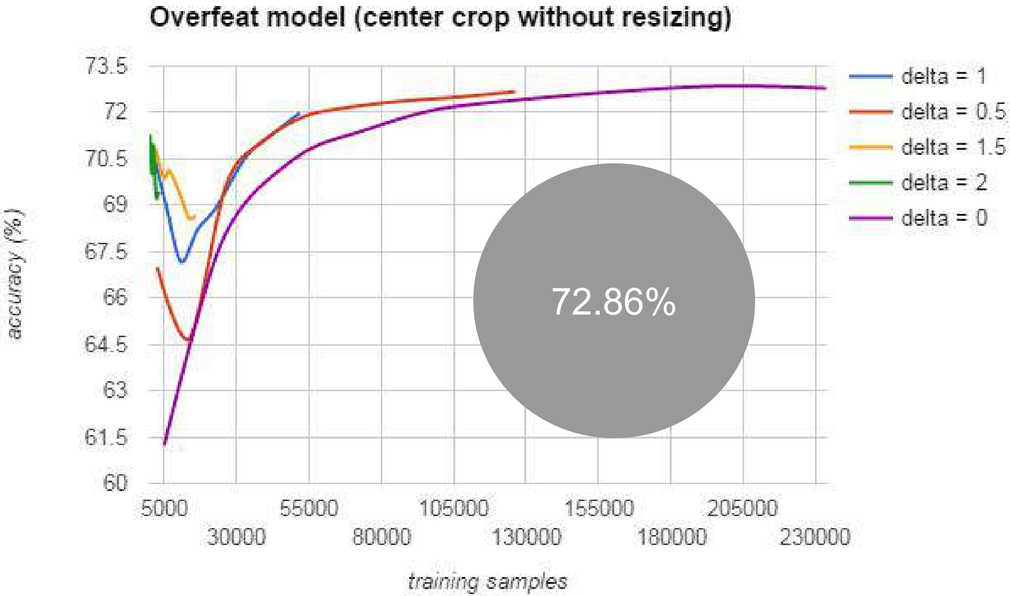




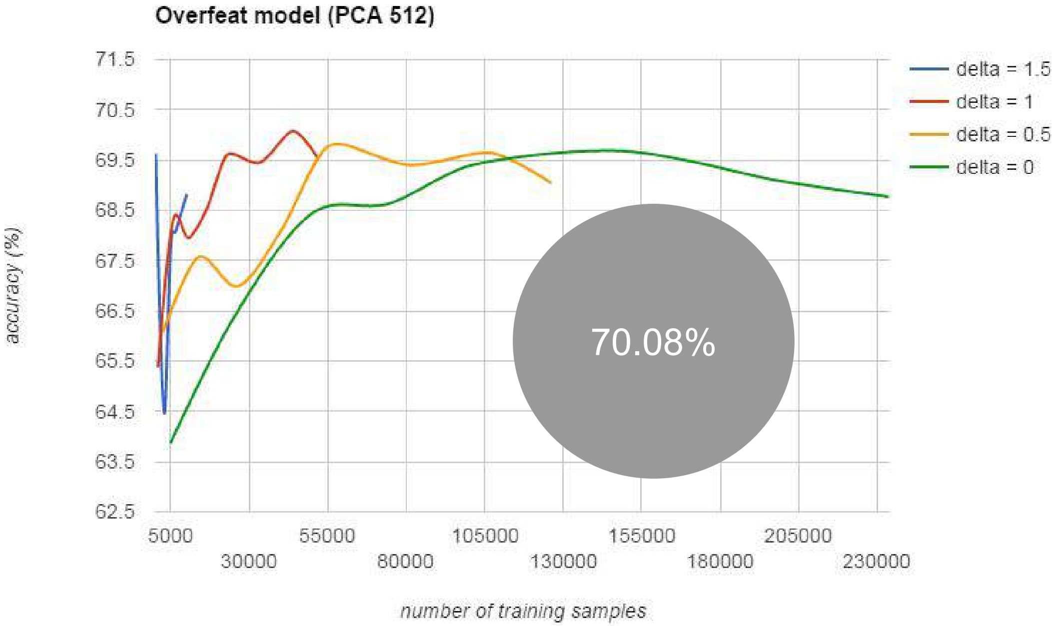
My experiments



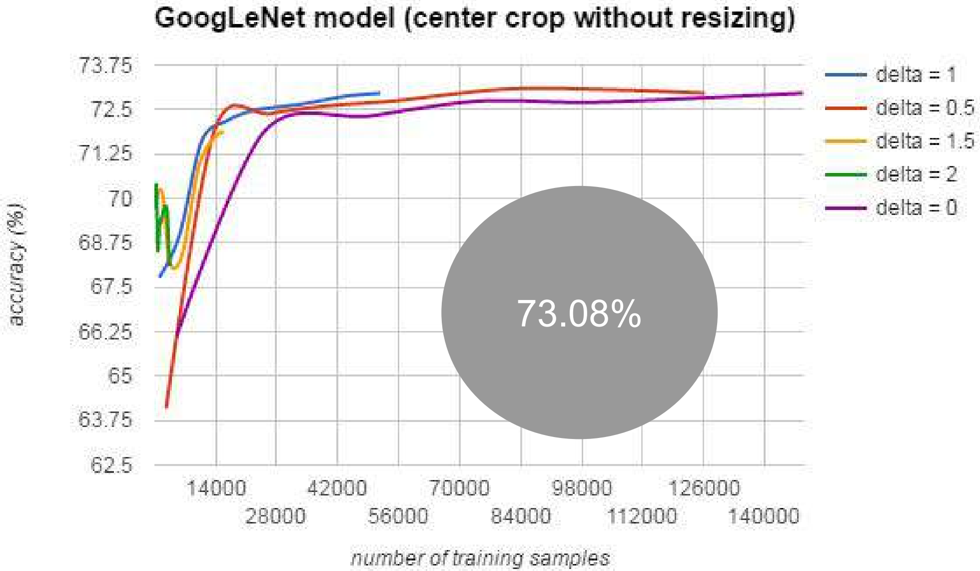
My experiments



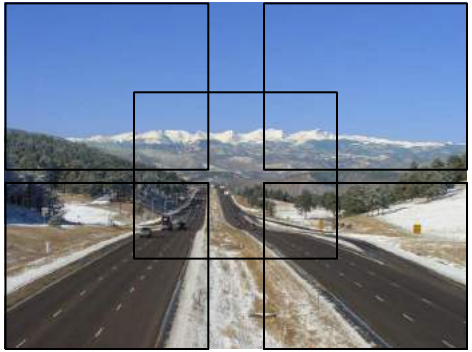
My experiments



My experiments



My experiments

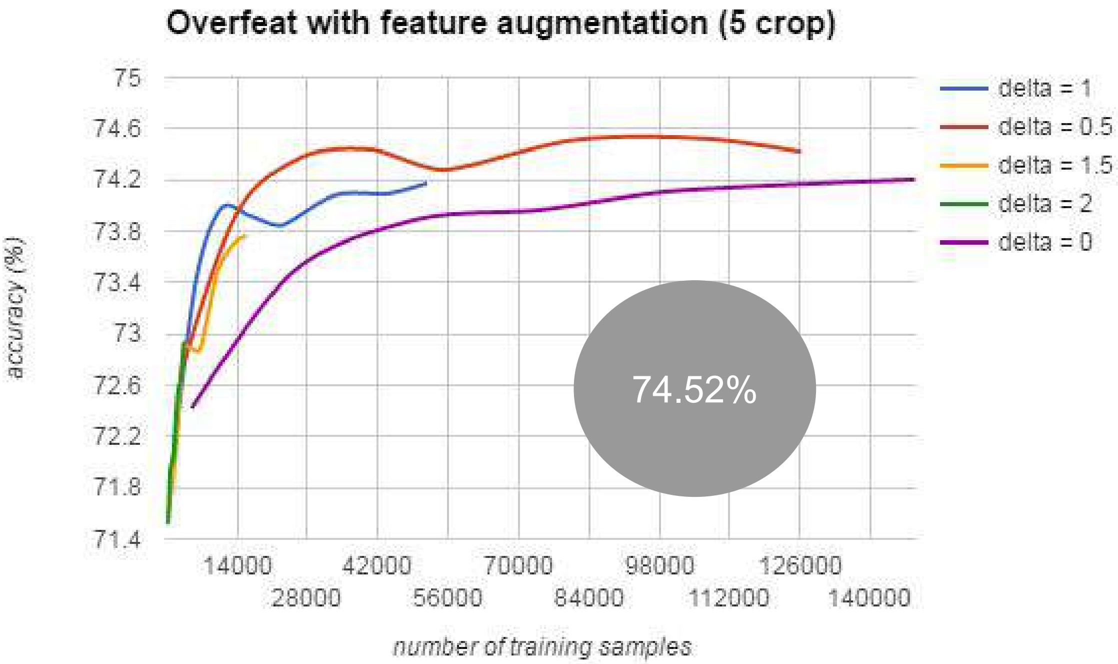


f (1), f (2), f (3), …, f (20480)

f1(1)...f1(4096) f2(1)...f2(4096) f3(1)...f3(4096) f4(1)...f4(4096) f5(1)...f5(4096)

1 2 3 4 5

My experiments



???

We took these normalized inputs (random crop, warp, center crop, padding) for SCNN training (2023 page, last sentences of first paragraph)

Using the selected network architecture, we trained and evaluated SCNN with four types of inputs (random crop, warp, center crop, padding) on the AVA dataset (2024 page, first sentences of second paragraph)