Nine problems for statisticians in the deep learning era

Rama Chellappa

Johns Hopkins University

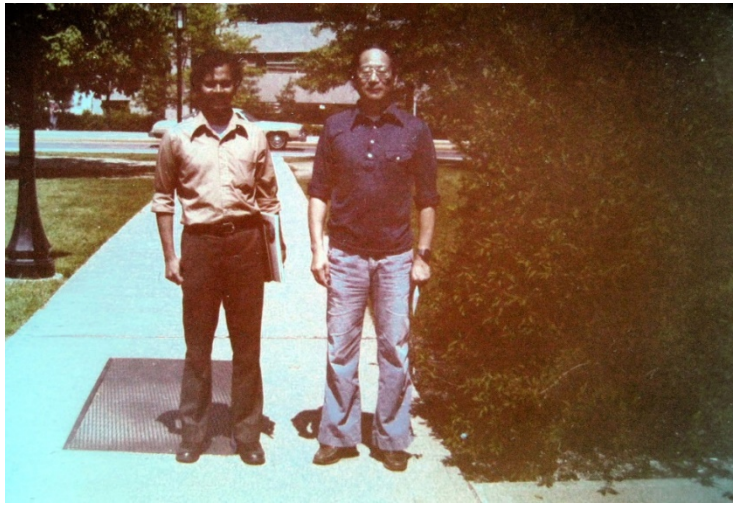
Outline

- Impact of statistics on image representations
- Impact of statistics on performance bounds for computer vision algorithms
- Shape statistics
- Action detection
- Bayesian inference
- Ten problems for statisticians

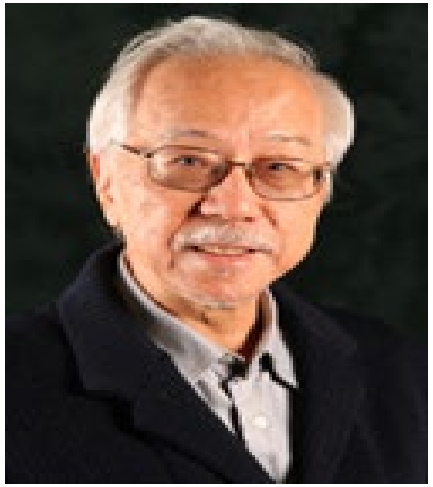
Doctoral mentors

Prof. Azriel Rosenfeld

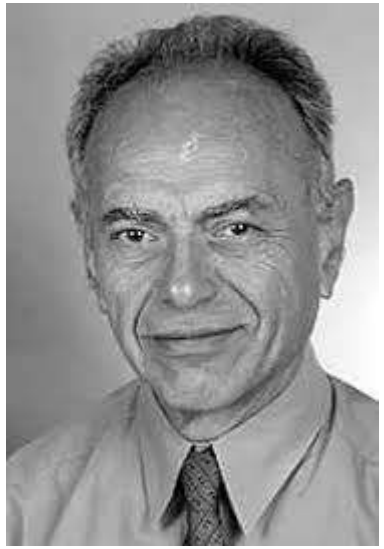
Prof. R.L. Kashyap



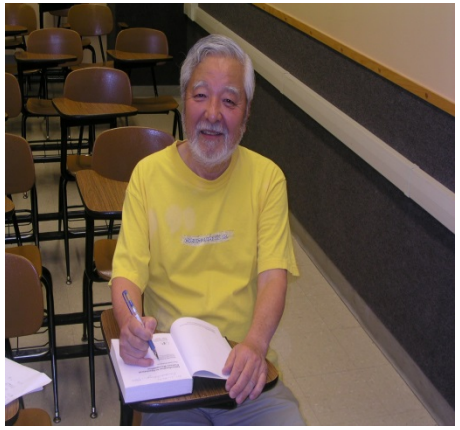
With Prof. K.S. Fu



Prof. Thomas Huang



Prof. Jack Skalnsky



Prof. K. Fukunaga

Neural networks, deep learning networks

- NNs have been around for a while: Second life in the 80's
- Good examples of NNs for vision
  - Autonomous driving (Dean Pomerleau)
  - Face detection (Tommy Poggio)
  - OCR (mostly in NIPS)
- Our work: Stereo, optic flow computation and texture segmentation using NNs
  - Artificial neural networks for computer vision, Springer (with Y.T. Zhou)
  - MRF-based algorithms and ANNs
  - MRFs and Boltzman machines
  - CNN for ATR, jointly with Army Research Lab sciences, Computer Vision and Image Understanding, 2001.
- Face recognition using a version of dynamic link architecture (CVPR 1992)
- Revenge of the networks
  - By adding more layers, deep learning networks are beating SVMs
  - LeNets, (1989, 1998), AlexNet (2012)
  - Turing award for Hinton, LeCun and Bengio

Statistis and image representations

- Peter Whittle, Biometrika - 2-D non-causal autoregressive models
- Second-order homogeneous random fields, in Proc. 4th Berkeley Symp. Mathematical Statistics and Probability, vol. 2. Berkeley, Calif.: Univ. California Press, 1961 – A.M. Yaglom
- On Gaussian fields with given conditional distributions, Theory of Probability and Its Applications, 1967, Y.A. Rozanaov
- R.L. Dobrushin's work
- Hammersley-Clifford theorem establishing the equivalence of MRFs and Gibbs distributions (1971)
- Spatial interaction and statistical analysis of lattice models - Julian Besag, Jl. Royal Stat. Society. Ser. B, 1974
- P.A. P. Moran and Julian Besag, Estimation in GMRF – 1975
- Julian Besag, Error-in-variable formulation, 1977
- W. E. Larimore, Statistical inference on stationary random fields, PIEEE 1977
- M. M. Ali. Analysis of stationary spatial-temporal processes: Estimation and prediction. Biometrika, 66:513–518, 1979 – Dynamic texture models

80's and 90's –The golden decades for statistical models of images

- Estimation methods, neighborhood selection rules, texture synthesis, classification, image restoration
  - Besag, Geman, Geman
- Julian Besag – Statistical analysis of dirty pictures - Iterated conditional mode (same as the energy function in a Hopfield neural model) - 1986
- MRFs were quite popular
- Whittle's simultaneous 2-D noncausal autoregressive models did not take off!
  - To my surprise, used in predicting the value of your home based on neighbors' homes.

Other tried/dropped models

- Fractals- Mandelbrot
- Simultaneous 2-D noncausal autoregressive models
  - White noise instead of the correlated noise as in Gaussian MRFs
  - Originally suggested by Peter Whittle in 1954
  - Not Markov wrt to the neighbor set, but Markov wrt a higher order neighbor set.
  - To my surprise, used in predicting the value of your home based on neighbors' homes.

Drawbacks of MRFs

- Sensitivity of parameters to transformations (illumination, rotation, resolution)
- The number of parameters were typically less than 20
- Extensions to videos harder
- Hierarchical representations harder to analyze
- Discriminative methods performed better
- Absence of non-linearities
- Emergence of better approaches
  - Normalized cuts for image segmentation
- These drawbacks point to why statistics may be “absent” in deep learning despite being data driven!

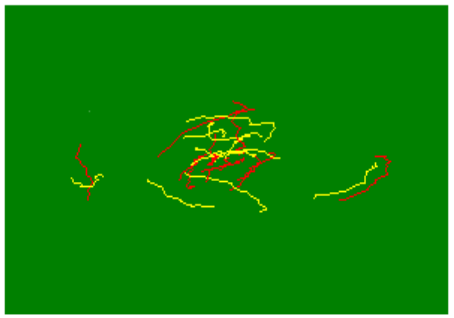
Impact of statistics on performance bounds for computer vision algorithms

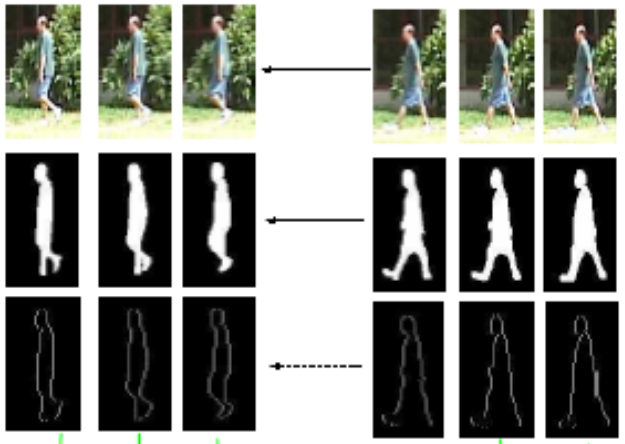
- Performance characterization of regression-like methods
  - Haralick's work in the 90's
- Cramer-Rao bounds for structure and motion parameters using point features and optic flow
  - Broida, Chellappa, JOSA 1989, PAMI 1991, Young and Chellappa, PAMI 1990, 1992
- For linear methods perturbation analysis without needing a distribution.
  - Roy-Choudhury and Chellappa, IJCV 2003
- Robust statistics
  - M estimation, least Median squares (Rousseeuw)

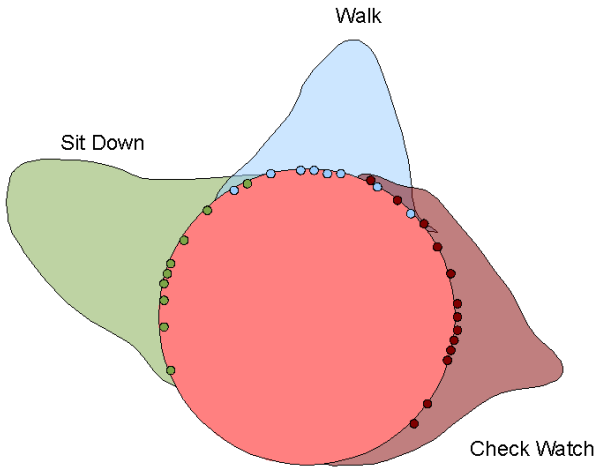
Statistics on manifolds

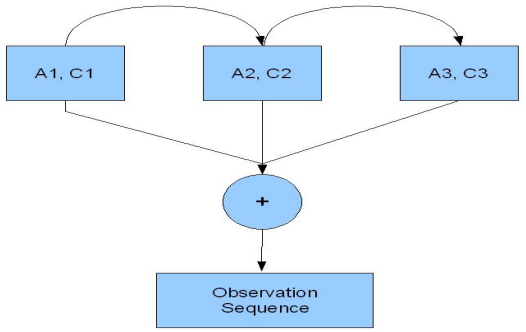
- Fisher-Rao metric (Veeraraghavan et al., TIP 2008)
- Shape statistics (Mardia, Srivastava)
- Circular statistics for vehicle orientation estimation (Hara and Chellappa, IJCV 2017)
- Dictionary learning on statistical manifolds

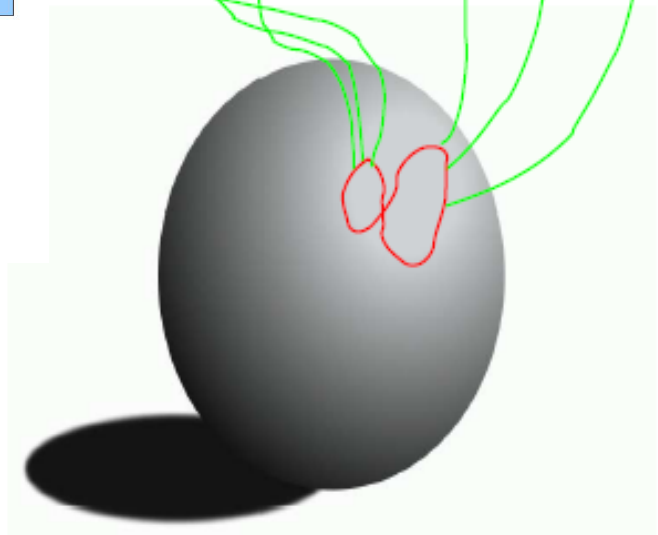
Statistics on manifolds











Bayesian inference and computer vision

- Model order selection
  - Bayes information criterion (Kashyap 1977, Schwartz, 1978, Kashyap and Chellappa, IEEE Trans. IT 1983))
- Object recognition (David Cooper)
- Bayesian graphical models – shallow hierarchy
- Simulated annealing (Geman and Geman, 1984), particle filter (Miller, Srivastava and Grenander, 1995, Isard and Blake, 1996), MCMC (too many to list)..
- Relevance vector machines did not take off

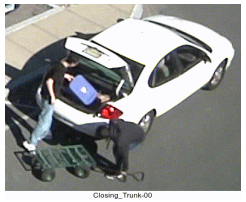
Statistics is struggling ..

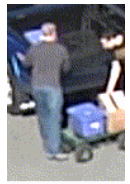
- Statistical methods were mostly absent when compressive sensing and sparse representations were popular (2005-2012)
  - Statistics likes *l2* more than *l1* and *l0*!
- When hierarchical models are considered
  - Multi-resolution time series models, MRFs have challenging inference problems (learnt the hard way in 1982!)
  - Wavelets dominated in the 90's and statistical methods were hard to come by for hierarchical representations
- Statistical methods for hierarchical and non-linear models (Deep learning) are even more challenging!

Deep learning miracle or mirage

- Since 2012, computer vision has become a one-trick pony
- Impressive performance on many tasks
  - Object/ face detection, classification, verification
  - For face verification at 10-7 false acceptance rate, > 90% true acceptance rate on faces in the wild (IARPA JANUS program)
- Not there yet
  - For action detection, probability of miss for the best systems are 0.52 and 0.71 for known and unknown facilities and 37 actions.
  - Deep intermodal video analytics

Deep intermodal video analytics





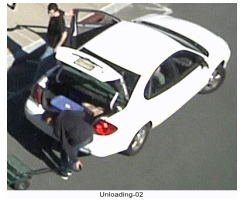


Figure 3. On the left, the DIVA action `Closing` makes up only a small portion of the image, and the surrounding context has no value for the action classification task. The THUMOS action `Cricket` on the right is much larger in the image, and the entire image's context is useful for classification.

- DIVA actions are very small
  - The average activity is 150x300 resolution
- Limited data, actions of variable lengths

In Phase 3
UMD (Lead) with CMU, Columbia, JHU and UCF as partners

Despite being successful, deep learning-based methods are weak

- While seen as a non-linear mapping between data and labels, lack of analytical results is worrisome.
- Learning millions of parameters from relatively small data is statistical blasphemy!
- Tightly clings to training data and does not generalize well
- No performance measure to say why and when it works
- We can pile on…

Problem1: Unsupervised domain adaptation

Source domain
Data: X, Labels: Y

Target domain
Data: X', Labels: Y'

Transfer Learning[1]
- ❖ $P(Y|X) \neq P(Y'|X'), P(X) \approx P(X')$

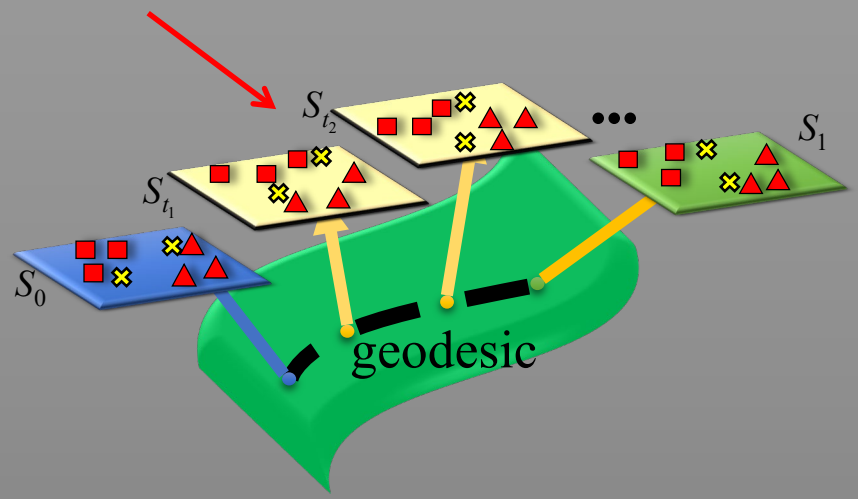
Domain adaptation
- ❖ $P(X) \neq P(X'), P(Y|X) \approx P(Y'|X')$

Image credit: Saenko et al., ECCV 2010, Bergamo et al., NIPS 2010

[1] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Trans. Knowledge and Data Engineering, 22:1345 –1359, October 2010.*

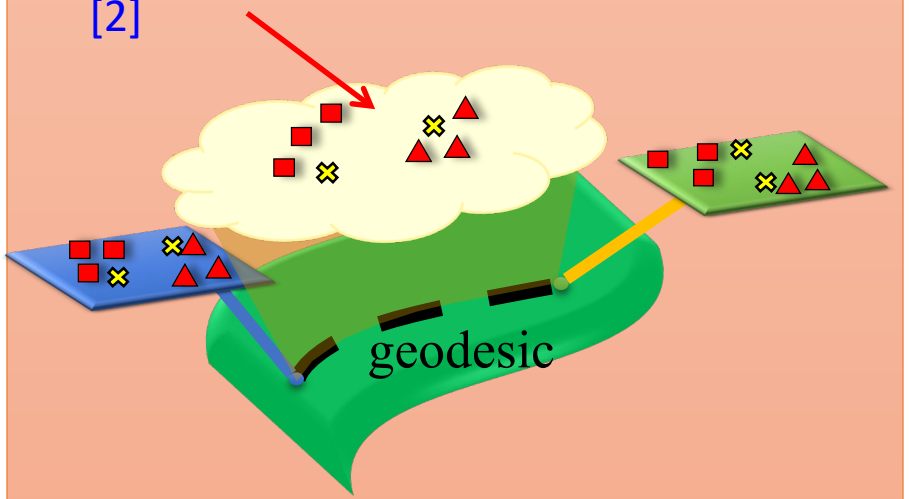
Finite vs infinite intermediate subspaces for domain adaptation

Finite intermediate subspaces [1]



- samples a limited number of intermediate subspaces
- concatenates the subspace projection as the final features for learning.
- train a discriminative learner on the projected source data

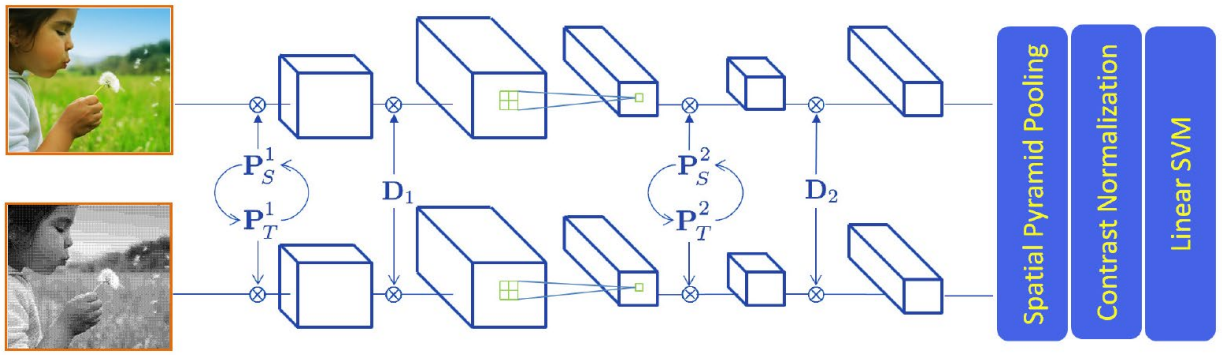
Infinite intermediate subspaces [2]



- Samples infinite intermediate subspaces
- Integrates the distance of sample projections along the geodesic

[1] R. Gopalan, R. Li, and R. Chellappa, "Domain Adaptation for Object Recognition: An Unsupervised Approach", ICCV, 2011, PAMI 2014
[2] Gong et al., Generalized Kernel flow, CVPR 2012

Hierarchical dictionaries (Left) and GANs (Right)



Domain-specific transformations
Joint training of P_S and P_T

Sparse coding
Shared dictionary

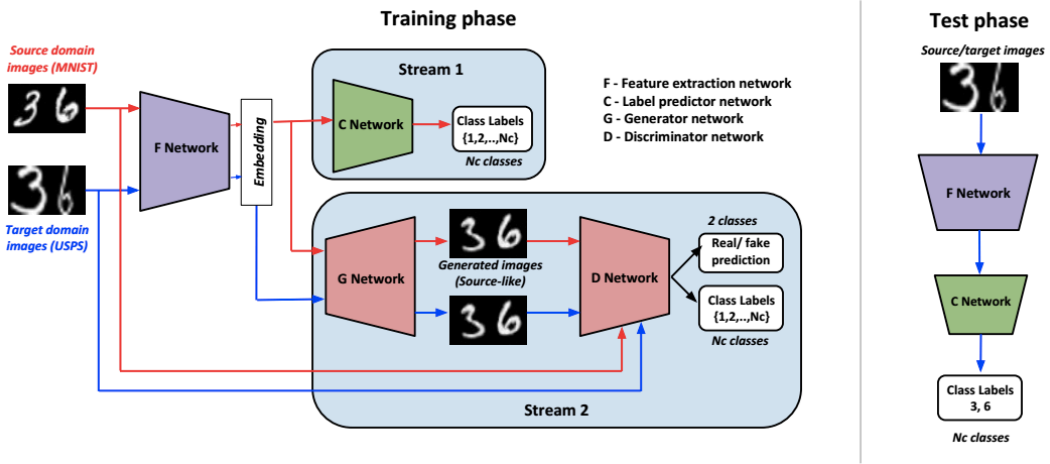
Max-pooling across 4x4 pixels for each dictionary atom

Contrast-normalization

Higher layer adaptation

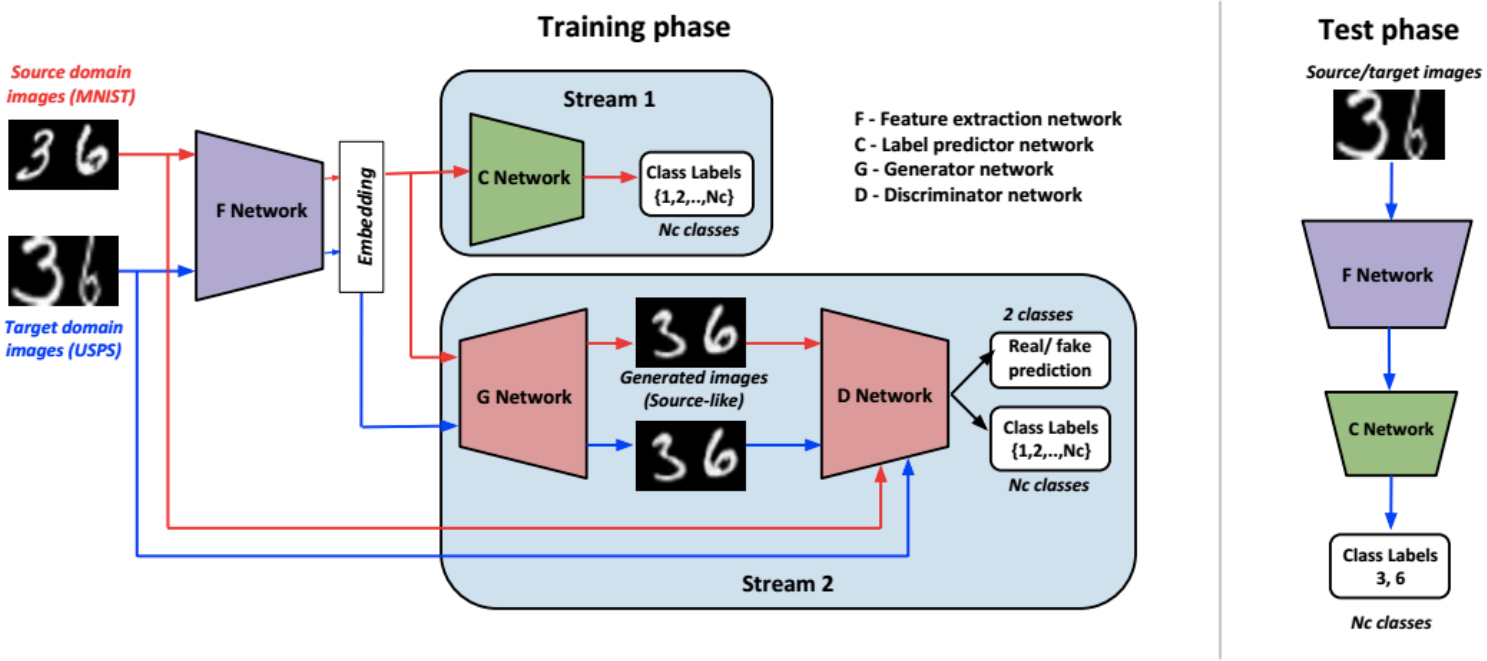
Multi-level feature aggregation with 1x1, 2x2, 3x3 spatial blocks

Nguyen et al., TIP 2015



Sankaranarayanan et al., CVPR 2018

Generate to adapt for classification: Overall approach



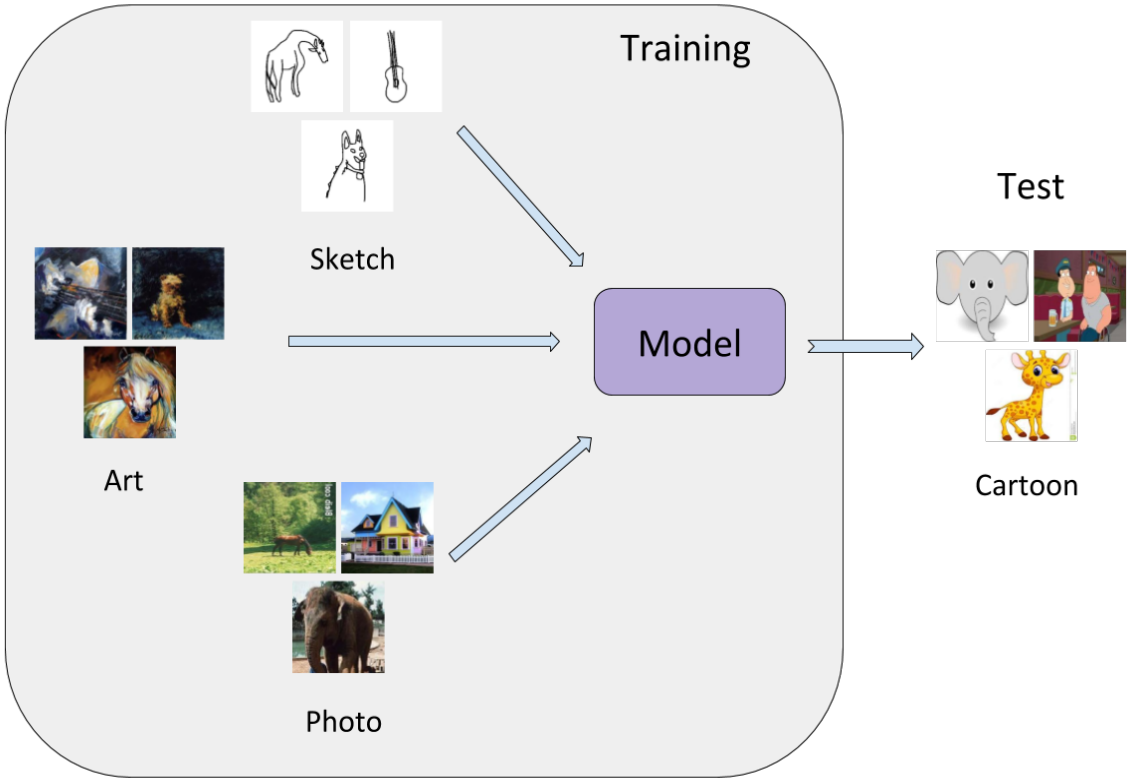
Why are DCNNs and GANs popular?

Source	Target	Manifolds 2012	Dictionaries 2016	Deep features 2017	Deep features and /GANs 2018
Webcam	Dslr	71.2		99.5	99.8
Dslr	Webcam	68.8	72	98.2	97.9
Amazon	Webcam	55.6	72	62.4	86.5
Amazon	Dslr			64	87.7
Dslr	Amazon		48.9	52	72.8
Webcam	Amazon		49.4	48.4	71.4

Unsupervised domain adaptation results for office data set

Problem 2: Domain generalization

Domain generalization involves generalizing to novel test domains using variations in multiple source domains



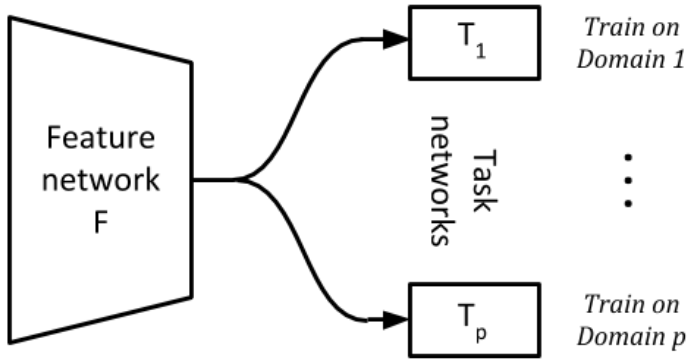
Y. Balaji, S. Sankaranarayanan and R. Chellappa, "MetaReg: Towards Domain Generalization Using meta-regularization", Proc. Neural and Information Processing Systems, Montreal, Dec. 2018.

Our approach

- Use of regularization to enable generalization to novel test conditions
  a. A parametric regularizer acting on the weights of neural network
  b. Parameters of the regularizer should capture the notion of domain generalization
- Use of meta-learning for estimating the parameters of the regularizer
- After estimating the regularizer, a domain invariant model is trained using regularized cross-entropy loss on the source domains.
- Examples
  - Face recognizer by training on different domains representing pose, illumination, expression, resolution, etc.
  - Other examples (multi-sensor based target recognition, wine tasting, …

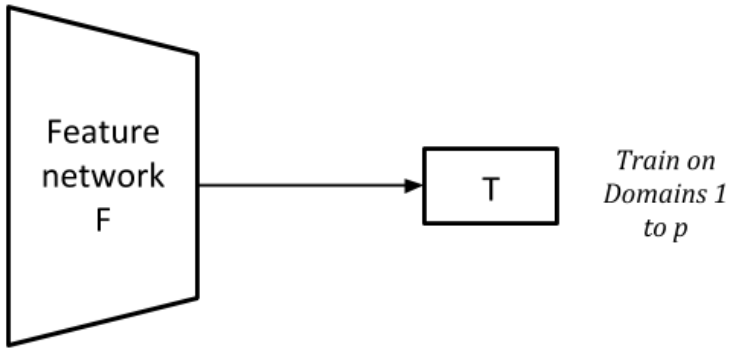
Network design - Step (i)

- Decompose the neural network into shared feature network and multiple task networks
  a. Feature network shared across all source domains
  b. Task network T_i trained on source domain i



- *Number of source domains: p*

- Once the regularizer is estimated, a single F-T network is trained on all source domains using the regularized cross-entropy loss



Learning the regularizer

- The figure shows the regularizer update for generalizing from domain i to j

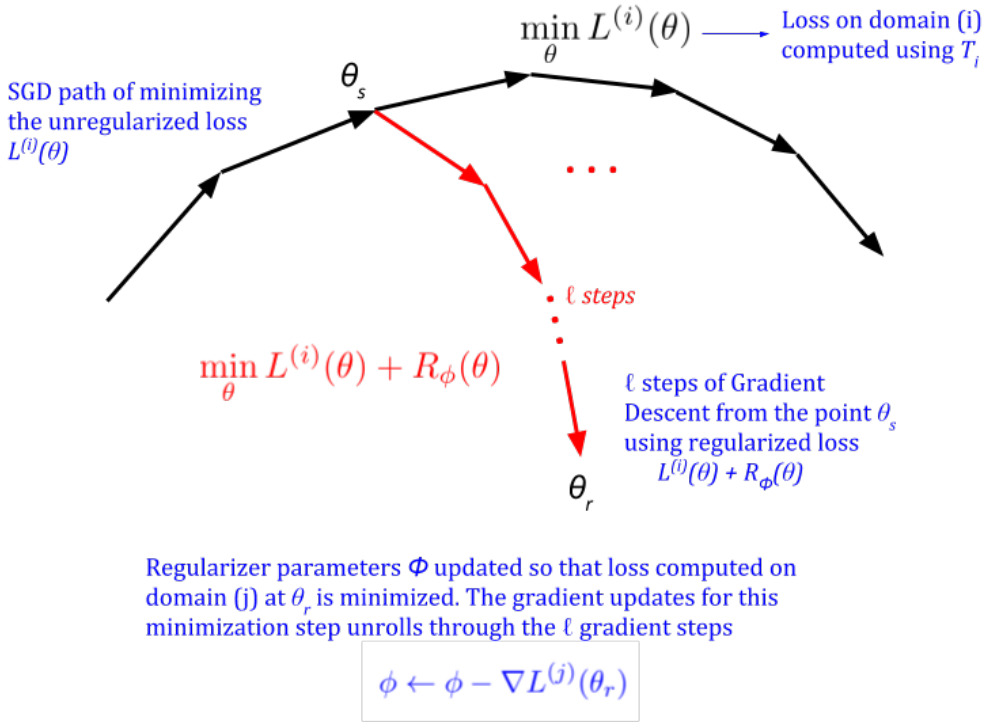
$$R_\phi(\theta) = \sum_i \phi_i |\theta_i|$$

- These updates are repeated for every (i, j) source domain pair

$L^{(i)}(\theta)$ - *cross-entropy loss of F-T_i networks on source domain i*

$R_\Phi(\theta)$ - *regularization function*

- We use a weighted L_1 loss as regularizer



(b) Learning a regularizer for generalizing from domain (i) to (j)

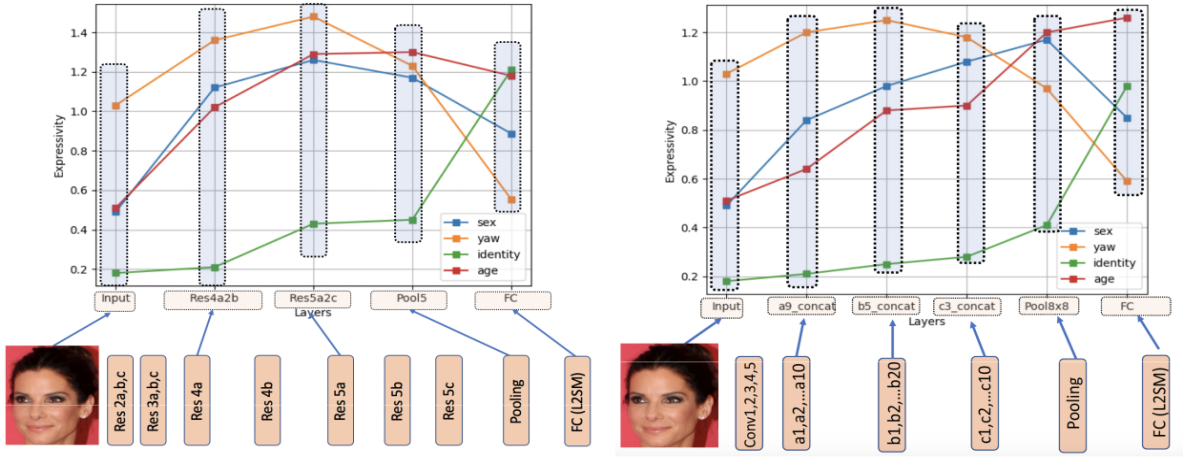
At each iteration we sample a pair of domains (i, j) from the training set. The black arrows are the SGD updates of the task network Ti trained on domain i. From each point in the black path, we take l gradient steps using the regularized loss and the samples from domain i to reach a new point ∗. We then compute the loss on domain j at ∗. The regularizer parameters ϕ are updated so that this meta-loss is minimized. This ensures that the task network Ti trained with the proposed regularizer generalizes to domain j

Problem 3:Model pruning/optimization

- Existing methods for model pruning/optimization are heuristic
- Is there a BIC for deep networks?
- Which parameters are statistically insignificant and what harm they would cause if removed?
- Learning from noisy data (errors-in-variable formulation)
- We need rigorous hypothesis testing procedures.

Problem 4: Analysis of hierarchical non-linear models

- How does information flow from data to labels? Are convolutions (correlations?) the best to leverage on?
- What else is coded in the deep features other than ID information?
- Influence of nuisance factors on the main task
- Expressivity of yaw, sex, age and identity



Mutual information

$$I(V_1, V_2) = D_{KL}(\mathbb{P}_{V_1,V_2} || \mathbb{P}_{V_1} \otimes \mathbb{P}_{V_2}) \tag{1}$$

$\mathbb{P}_{V_1,V_2}$ = Joint probability distribution

$\mathbb{P}_{V_1} \otimes \mathbb{P}_{V_2}$ = Product of marginals

Donsker-Varadhan representation of KL divergence

$$D_{KL}(p|q) = \sup_{T:\Omega \longrightarrow \mathbb{R}} \mathbb{E}_p[T] - \log \mathbb{E}_q[e^T]$$

Where p,q are probability distributions, and T is a function that maps a set of parameters to a real number. Plugging this in eq. (1) we get

$$I(V_1, V_2) = \sup_{T:\Omega \to \mathbb{R}} \mathbb{E}_{\mathbb{P}_{V_1,V_2}}[T] - \log(\mathbb{E}_{\mathbb{P}_{V_1} \otimes \mathbb{P}_{V_2}}[e^T])$$

Computing supremum using gradient descent

V_1 = F = Set of 512 dimensional face descriptors
V_2 = A = Attribute labels for face descriptors

$$I(F, A) = \sup_{T:\Omega \to \mathbb{R}} \mathbb{E}_{\mathbb{P}_{F,A}}[T] - \log(\mathbb{E}_{\mathbb{P}_F \otimes \mathbb{P}_A}[e^T]) \quad (3)$$

Let's consider T as a neural network with a parameter set Ω ,

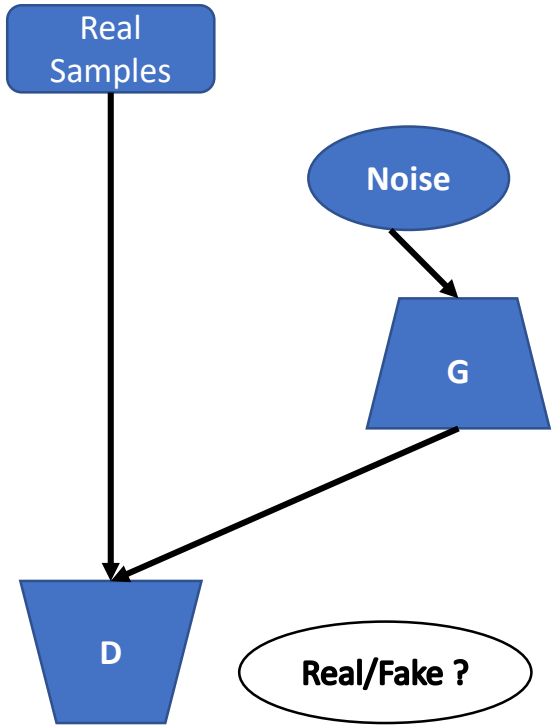
Finding the supremum of argument Eq (3) is same as finding minima of L_Ω defined below:

$$L_\Omega = \log(\mathbb{E}_{\mathbb{P}_F \otimes \mathbb{P}_A}[e^T]) - \mathbb{E}_{\mathbb{P}_{F,A}}[T] \quad (4)$$

We train a network T to minimize L_Ω , in order to find optimal parameter set Ω^*,
which maximizes the argument of Eq 3. This maximum value approximates mutual information I(F,A).

Problem 5: Analysis of generative adversarial networks

1. Traditional GAN architecture involves a Generator (G) and Discriminator (D) pair.
2. Both are modeled as deep networks (since DCGAN)
3. Optimize D to identify the generator's fakes compared to real samples
4. Optimize G to fool the discriminator in thinking that G produced a real sample
5. Min-max adversarial game between G and D



$$\min_G \max_D \mathbf{E}_{x \sim p_{data}}(\log(D(x)) + \mathbf{E}_{z \sim p_{noise}} \log(1 - D(G(z)))$$

Goodfellow 2014

GANS are useful for….

- Vehicle re-identification using attention models – ICCV 2019 (oral paper), ECCV 2020
- Defending against adversarial attacks; Defense-GAN – ICLR 2018
- GAN-based reduction of metallic artifacts in CT images –CVPR 2019
- GAN-based transfer of visible to thermal images -Prof. Vishal Patel
- Text to video synthesis, IJCAI 2019
- Wasserstein GAN for domain adaptation, ICCV 2019
- GANs for restoring turbulence-degraded images, FG2020
- GANs for landmark extraction from low-resolution images –Lau, Kumar and Chellappa (Under review)
- GAN zoo!
  - https://github.com/hindupuravinash/the-gan-zoo

A generative model without statistical validity is an anomaly

Problem 6: Choosing the best subsets for training from a much larger pool of training data

- Setup
  - Given a **fixed** classifier architecture
  - A set of **labeled** training data points from ***L*** different classes
- Objective
  - Iterative algorithm
  - At teach time instance *t*, select a subset of the training data to resume training on
- Selection criteria
  - The samples in the selected batch must be such that the classifier is *uncertain* about classifying them (or *certain* but *wrong* in its classification)
  - The batch must have a *balanced* selection from all classes
  - The batch should be sufficiently *diverse*.
  - The batch should be *representative* of the training samples.

Classifier uncertainty and error

- At time instance *t*, the classifier produces *L* outputs for each training sample $X_{i,k}$ from class *k*:

$$p^t(X_{i,k}) = [\underbrace{p_1^t(X_{i,k})}_{\times}, p_2^t(X_{i,k}), \ldots, p_L^t(X_{i,k})]$$

classifier's probability that $X_{i,k}$ belongs to class 1 (correct class is *k*)

- Classifier uncertainty:

$$c^t(X_{i,k}) = -\sum_{l=1}^{L} \left(\beta^t \cdot \mathbb{1}\left[l = k\right] + (1 - \beta^t) \cdot p_l^t(X_{i,k})\right) \log p_l^t(X_{i,k})$$

  - Can be interpreted in two ways:
    (1) Weighted sum of error term (for classifier *error*) and entropy term (for classifier *uncertainty*)
    (2) Cross entropy between "weighted correct label" (in the case of label noise) and classifier predicted probabilities

Class balance

- At each time instance t, we select a total of M^t samples, distributed among all classes in a balanced way.
- We assign a budget M_k^t to each class depending on the classifier's average uncertainty on this class.
- We use a logarithmic objective function and formulate the problem as follows:

$$\max_{M_k^t \in \mathbb{Z}^+} \sum_{k=1}^{L} \log\left(1 + \alpha \cdot c_k^t \frac{M_k^t}{M^t}\right) \qquad \text{s.t.} \qquad \sum_{k=1}^{L} M_k^t \leq M^t; \quad M_k^t \leq |\mathcal{X}_k|,$$

- Since we are considering supervised learning settings, we can leverage the label information: Find a diverse and representative subset of each class separately.
- We can therefore solve L independent problems.

Diversity and representativeness

- **Diversity:** Seek to maximize the average distance between all selected samples ***(from each class)***

$$\max_{\mathcal{B}} \frac{1}{M^2} \sum_{X \in \mathcal{B}} \sum_{X' \in \mathcal{B}} d(X, X') = \max_{\mathbf{s} \in \{0,1\}^N} \frac{1}{M^2} \mathbf{s}^\intercal \mathbf{D} \mathbf{s}$$

- **Representativeness:** Seek to minimize the average distance between selected samples and non-selected samples ***(from each class)***

$$\min_{\mathbf{s} \in \{0,1\}^N} \frac{1}{M(N-M)} (\mathbf{1} - \mathbf{s})^\intercal \mathbf{D} \mathbf{s}$$

- Solve a separate optimization ***for each class*** to jointly maximize diversity, representativeness, and uncertainty

$$\max_{\mathbf{s} \in \{0,1\}^N} \quad \lambda_1 \cdot \underbrace{\frac{1}{M^2} \mathbf{s}^\intercal \tilde{\mathbf{D}} \mathbf{s}}_{\text{diversity}} - \lambda_2 \cdot \underbrace{\frac{1}{M(N-M)} (\mathbf{1} - \mathbf{s})^\intercal \tilde{\mathbf{D}} \mathbf{s}}_{\text{representativeness}} + \lambda_3 \cdot \underbrace{\frac{1}{M} \tilde{\mathbf{c}}^\intercal \mathbf{s}}_{\text{classifier uncertainty}}$$

- Subject to the budget constraint assigned to the class by the water-filling algorithm

$$\text{s. t.} \qquad \mathbf{1}^\intercal \mathbf{s} = M$$

- Approximate solution to this NP-hard problem obtained by semi-definite programming (SDP)

Results on VGG dataset

- We use our algorithm as a fine-tuning strategy on a CNN pre-trained on CASIA-WebFace. We start with pre-trained weights for the first 15 layers, and add two randomly initialized fully connected layers.
- We fine-tune on VGG Face using 20 non-overlapping subjects.
- Examples of images selected by our algorithm in the 1st loop, when representativeness weight λ_2 is large and the uncertainty weight $\lambda_3 = 0$

- Examples of images selected by our algorithm in the 13th loop after λ_2 has considerably decreased and λ_3 increased.

- VGG testing accuracies:

# of selected samples	0	100	500	1300	1600
Our approach	12.73%	**89.69%**	**93.23%**	**97%**	**97.15%**
Random	12.45%	80.05%	90.79%	94.89%	94.89%

Algorithm selections at the beginning of training and 75% through the training process: SUN397

Lighthouse

Landing deck

Library interior

Laundromat

Operating room

Office building

Islet

Iceberg

Dining room

Problem 7: Prediction of critical events from heterogeneous data

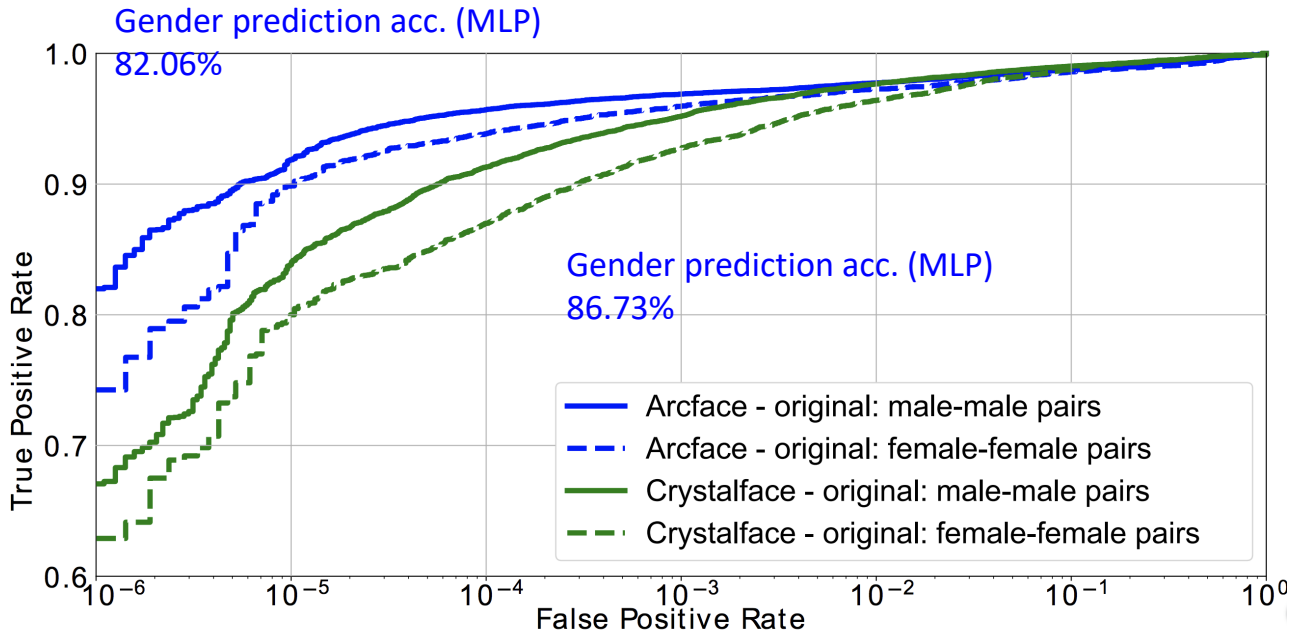
- Inputs: a) Clinical, claims and specialized JHM research data on the particular patient; b) Similar data from prior system (consortium) experience projected from PMAP onto a clinical cohort database (registry); c) Outputs of video and speech processing algorithms; and d) Expert knowledge about the etiology of the health or disease condition.
- Outputs: (1) the prediction, prevention, monitoring, and intervention of frailty and dementia, (2) the definition, measurement, and promotion of physical, physiological, and psychological well-being, and (3) the identification of robust signals, biomarkers, and processes of frailty and dementia.
- Bayesian hierarchical models (Zeger, Nishumura)
- More needs to be done!

Schematic of the information flow within the JHU Precision Medicine Analytics Platform (PMAP). The AI suite will analyze the integrated data and return the results to the clinician and patient to improve their interactive and collaborative shared decision making.

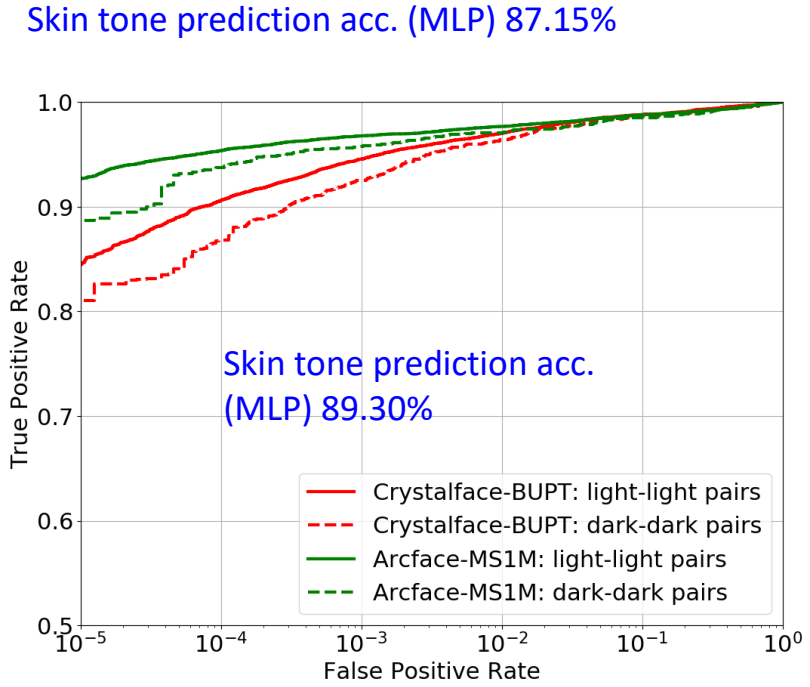
Problem 8: Mitigating AI bias

Hypothesis: Expressivity leads to bias

Low gender predictability → low gender bias



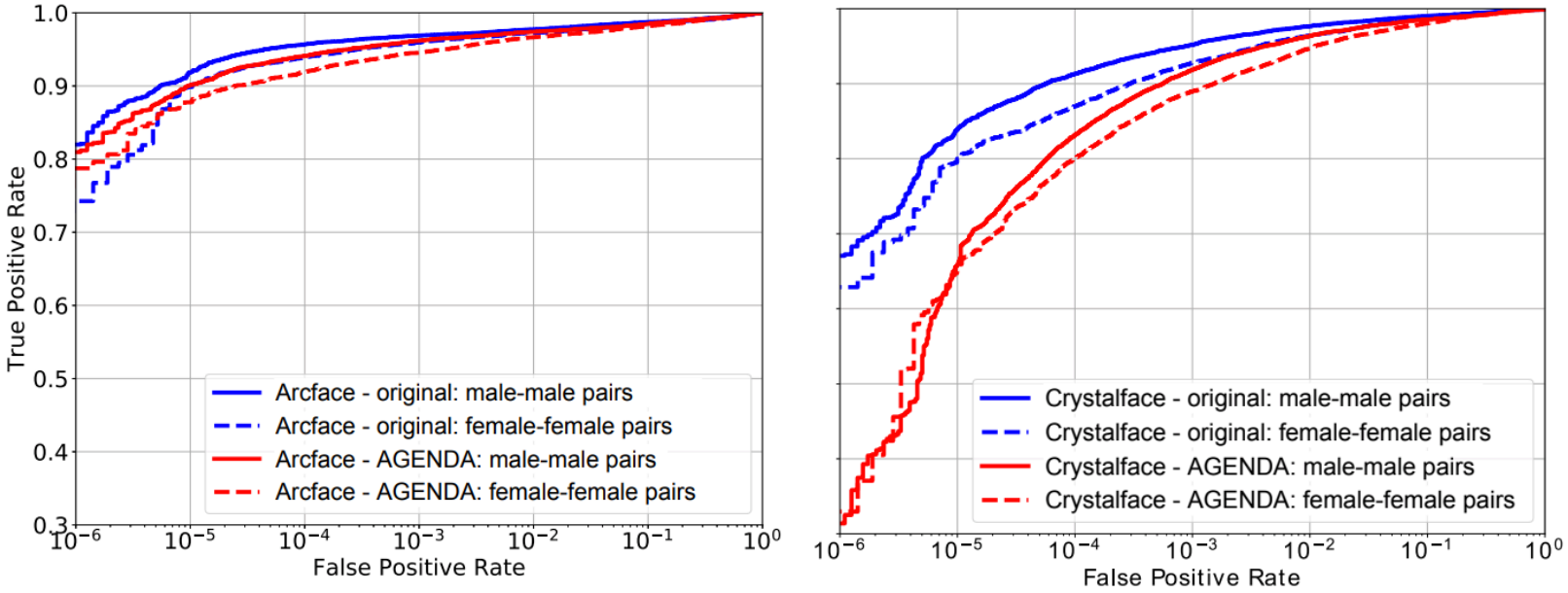
Low skin tone predictability → low skin tone bias



Bias mitigation using loss function

$$\text{Bias}^{(F)} = |\text{TPR}_m^{(F)} - \text{TPR}_f^{(F)}|$$

Bias mitigation loss reduces Gender bias in face descriptors from two SOTA network: Arcface and Crystalface



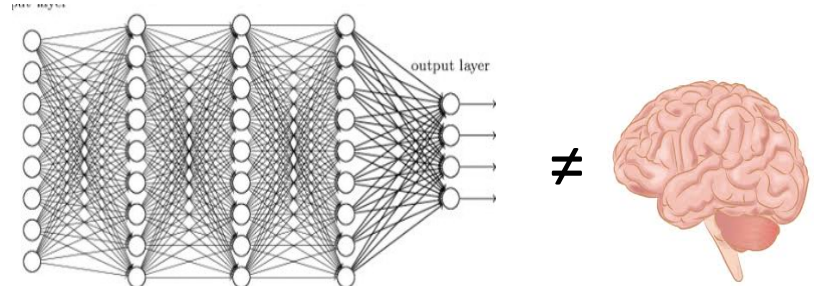
Dhar et al., Towards Gender-Neutral Face Descriptors for Mitigating Bias in Face Recognition. 2020. arxiv: 2006:07845

Problem 9: Robust statistics for handling adversarial attacks

- Huber's robust statistics, M and GM estimates
- Robustness in time series – Outliers at the

scale of noise can cause issues-Doug Martin

- White, black and Carlini-Wagner attacks
- Defense-GAN [Samangouei, et al., 2019]
- Modify the incoming data to lie on the manifold
- generated by the trained GAN.
- On manifold and off-manifold cases
- Cohen et al., Detecting Adversarial Samples Using Influence Functions and Nearest Neighbors. 2019; arxiv:1909:06872



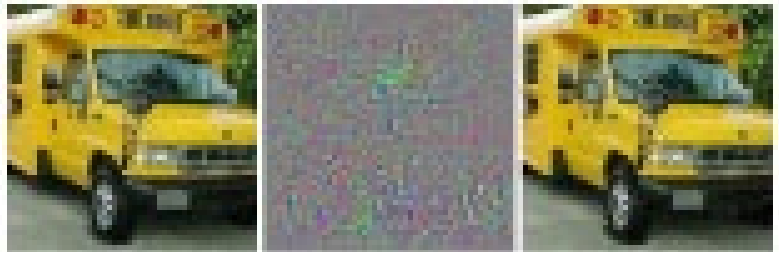


Figure: Courtesy. [Szegedy et al., 2013]

Conclusions

- Recent trends in machine learning, deep learning and AI have tons of problems for which we need rigorous solutions based on mathematical statistics.
- Lot of alchemy
- Alchemy is producing gold once in a while
- If this happens more, folks may not care how the gold is produced.
- Time is of the essence!