CVPR'18 Salt Lake City
Notes

David Stutz
davidstutz.de

Note: The notes found in this document are not necessarily correct or complete. Please always refer to the corresponding paper for details. Furthermore, the views expressed in this document are my personal ones and are not intended to offend anyone; they also do not reflect the views of my employer.

1. Monday – Interpreting and Explaining Deep Models in Computer Vision

Organizers: Wojciech Samek, Grégoire Montavon, Klaus-Robert Müller
Webpage: http://interpretable-ml.org/cvpr2018tutorial/
GCPR 2017 slides: http://iphome.hhi.de/samek/pdf/GCPR2017.pdf

Introduction

- Two views for explaining deep models:
  - The mechanics view, i.e., understanding the inner workings of the network.
  - The functional view, i.e., taking the network as representing a mapping, without knowing the internal workings exactly.
  - Here, the functional view is adopted.
- Sensitivity analysis (Baehrens et al., 2010, Simonyan et al., 2014):
  - Explains a variation of a function, not the function itself.
  - It basically asks "which pixel should I change to make the image a car".
  - It tells how the function value changes if pixels are changed slightly.
  - Shattered gradients problem: gradients become increasingly varying with network complexity and depth.
- Layer-wise relevance propagation (Bach et al., 2015):
  - Not suffering from the shattered gradients problem as it does not correlate gradients with input!
  - Which pixels contribute how much to the classification?
- Deconvolution methods (e.g., Zeiler and Fergus, 2014).
- Layer-wise relevance propagation:
  - Given the network's output, how to go backward?
  - Note that this is simple for linear models, but non-trivial for non-linear models.
- Historical remarks:
  - Gradient approaches: sensitivity.
  - Decomposition approaches: layer-wise relevance propagation and derivatives.
  - Optimization approaches.
  - Deconvolution approaches: guided back-propagation, deconvolution.

Techniques for Interpretability

- Evaluation of techniques:
  - First attempt: distance to some ground truth. Can also evaluate explanation for ground truth class, not for others.
  - Try to evaluate explanation axiomatically; it must pass a number of unit tests.
  - Four properties: conservation, positivity, continuity and selectivity.
  - Continuity: if input and predictions are almost the same, explanations should also be almost the same.
  - Continuity can be tested on real data by following a path on the input domain.
  - Sensitivity analysis does not seem continuous.

  - Selectivity: model must agree with the explanation; when removing relevant features, the evidence (output) should reduce.
  - Selectivity can also be tested on real data by destroying pixels that are relevant
  - All four properties can be tested for various method.
  - Question: can these properties be deduced from the equations?

- Reminder: backpropagation and layer-wise relevance propagation (LRP):
  - Conservation (i.e., sum of relevances at specific layer should be a constant fixed value equal to the function value).
  - Conservation can be deduced for LRP.
  - Simple sensitivity via "gradient x input" does not have the conservation property.
  - Continuity can also be deduced analytically.

- Conclusions:
  - Axiom-testing better than ground truth evaluation;
  - Some properties can be deduced from equations.
  - LRP satisfies key properties, sensitivity and "gradient x input" does not.

- From LRP to Deep Taylor Decomposition:
  - Proposition: relevance at each layer is a product of the activation and an approximately constant term.
  - Thus, relevance can be seen as a neuron with the activations as input and a constant term plus ReLU activation.
  - Then, do a Taylor expansion for the relevance.
  - This gives a relevance backpropagation rule, the so-called generic Deep Taylor rule.
  - Problem: how to choose the root point for the Taylor expansion?
  - Verifying the product structure: shown via induction.
  - For input layers, the input box constraint (e.g., for images) need to be considered.
  - Pooling layers can be handled like ReLU layers.
  - Also applicable to kernel methods.
  - Implementation allows backpropagation in linear time.

- Conclusions:
  - Ground truth explanations are elusive.
  - Some properties that techniques should exhibit can be deduced from the equations.
  - LRP basically performs deep Taylor decomposition.
  - The deep Taylor decomposition can be extended to other models and new types of data.

- Question:
  - The network decision might be very spiky, so the continuity property is not meaningful.
  - But, they model the score (i.e., evidence), instead of the decision itself.

Applications for Interpretability

- So far: explain black-box models by LRP, i.e., decomposing the function into relevances:

$$\sum_i R_i = f(x) \tag{1}$$

- Data types: applicable to images, text, molecules, games, VQA, video, EEG, fMRI and more.

- Models: applicable to LSTMs, bag-of-words models, SVMs, Fisher Vector models),

- So what insights do these data types and models offer?

- How good are the explanations and what can we use with them?

- How to compare techniques:
  - Compare selectivity: destroying or randomly setting/flipping heat map pixels and see how evidence/output decreases.
  - Area over the curve of the decreasing output can be used as measure (called AOC).
  - On images, LRP outperforms sensitivity analysis (Simonyan et al., 2014) and deconvolution (Zeiler and Fergus, 2014).
  - On text, pixel flipping is basically word deletion (i.e., setting word vector to zero).
  - Note that LRP gives both positive and negative explanations (i.e., what speaks for the class and against it).
  - So, the least relevant words (i.e., words speaking against the true class) can also be removed (cannot be done for sensitivity analysis).

  - (LRP is compared against a number of recent methods.) New Keras toolbox available.

- Questions:
  - How the evaluation is implemented.
  - All pixels (or patches, words) are sorted by relevance and iteratively flip or destroy this information (e.g., sample from uniform distribution).
  - Connection to adversarial attacks.
  - Destroying is non-specific, i.e., not bias, onlz using random.
  - Is it possible to identify adversarial attacks using explanations?
  - Question: can one fool explanations?
  - Can different models be compared?
  - Second part of talk.
  - Can this be applied to improve networks?
  - Some examples in the second part, but there is still a lot of potential for work there.

- Application: compare classifiers.
  - Classifiers might have similar performance, but explanations can suffer significantly.
  - Classifiers with better explanations might be preferable.
  - Example: relevant words identified by the model should be meaningful.
  - Example: relevance heat maps might be sparse, which can be preferable.
  - There seems to be a relation between the structure (architecture, etc.) the heatmap and the performance. But these relations are nor clear yet.

- Application: measure context use.
  - Importance of context = relevance outside bounding box / relevance inside bounding box.
  - Also allows to compare difference architectures.

- Application: compare configurations (pre-training, not pre-training etc.).
  - Example: interpretations are more meaningful for pre-trained models.
  - Example: can also reveal unwanted biases in data (e.g., young people always lough – loughing speaks against old people). Pre-training can help to avoid bias.
  - Different models have different strategies for the same problem!

- Application: learn new representations.
  - Get better representations using the explanations of models.

- Understand models and get new insights:
  - Example: In videos, model might focus on beginning and end of videos. Video could be played in fast-forward ...
  - Example: correct objects are identified in VQA tasks.

Wrap-Up

- Perspectives:
  - Is the generalization error all we need?
  - Assumes the standard model of machine learning; learn on training set and validate, without looking at the test set.
  - Some models with excellent generalization error might actually have learned to cheat.
  - Generalization error does not distinguish how the models achieve a specific performance.

- Machine learning in the sciences:
  - Neuroscience: brain-computer-interfacing (BBCI);
  - Machine learning educed the training time for the patient from several hours to minutes.
  - Chemical compound space: learn the Schroedinger Equation.

- Take-home messages:
  - Sensitivity analysis is not what you like to ask.
  - Explanations for simple models do not work for deep models.
  - LRP avoids the shattered gradients problem and works for a variety of models.
  - Explanations can be evaluated, e.g., using pixel flipping methods.
  - Explanations help to improve models.
  - Now what?

Personal Conclusion Interpretable deep learning, i.e. explaining predictions, is still very much dominated by visualization techniques. All discussed methods, e.g. sensitivity analysis, LRP or deconvolution, produce heat maps visualizing relevance. While this approach also works for text, medical imaging, videos etc., it is not clear whether visualization (or relevance) is enough for, e.g., verifying and explaining results in some applications such as autonomous driving or medical imaging. Still, the presented results are impressive, especially as the explanations allow to improve models or draw conclusions regarding the data. Additionally, the proposed axioms for evaluating explanations, as well as the methods for using explanations to compare classifiers seem useful tools for deep learning.

2. Monday – Robust Vision Challenge

Organizers: Andreas Geiger, Matthias Niessner, Marc Pollefeys, Carsten Rother, Daniel Scharstein, Hassan Alhaija, Angela Dai, Katrin Honauer, Joel Janai, Torsten Sattler, Nick Schneider, Johannes Schoenenberger, Thomas Schoeps, Jonas Uhrig, Jonas Wulff, Oliver Zendel
Webpage: http://www.robustvision.net/

Uwe Franke

- Topics:
  - What was done pre-deep-learning to achieve robustness.
  - How deep learning changed everything.
  - What still bothers us.
- Iterative semi-global matching and stixels:
  - Note that robustness here means robustness with respect to lighting, whether, daytime etc.
  - Started in the 90s with stereo vision.
  - Near real-time at around 2008 and it worked quite well.
  - Iterative semi-global matching won robust vision challenge at ECCV 2012.
  - Still not robust for very difficult conditions.
  - Then, a compact representation of $> 500k$ 3D points is needed.
  - Stixels: vertically aligned, rectangular areas summarizing points of same depth.
  - Stixel world with confidence can take care of noisy observations, reflections etc.
  - Still not robust for rare street configurations.
  - Solution: pixel labeling, i.e., semantic segmentation.
  - Stixel representation can also be used for semantic representations.
  - What did we learn?
    - ∗ Bigger receptive fields are better,
    - ∗ Linear operations are good, non-linear operations are better.
    - ∗ Time regularization is better.
    - ∗ Confidence at all stages are necessary.
- Deep learning:
  - Deep learning allowed better semantic segmentations than before.
  - Tremendous progress on CityScapes in the last few years, even in difficult situations.
  - State-of-the-art methods still have problems, which are partly caused by the labeling.
  - There are also still configurations/cases which are not observed in the training set.
  - Stixels for semantic segmentations reduce complexity and noise.
  - Similar observations hold for instance segmentation results.
  - What did we learn? We need to quantify robustness.
- What bothers us today?
  - Labeling is not an easy task.
  - Label definition is unclear.

3. Monday – Interpretable Machine Learning for Computer Vision

Organizers: Bolei Zhou, Laurens van der Maaten, Been Kim, Andrea Vedaldi
Webpage: https://interpretablevision.github.io/

Been Kim: Introduction to Interpretable Machine Learning

- When and why we need interpretability:
  - It is not true that simple models, linear classifiers or decision trees, are necessarily more interpretable.
  - Is interpretability possible at all? For example, for super-human performance.
  - Interpretability is not about understanding any bit for any data point.
  - Interpretability should help to use machine learning responsibly.

- Interpretability is by definition under-specified.
- Interpretability is even hard for humans, decisions cannot always be explained.
- Misconception: more data or better models will solve interpretability.
- There are also cases where interpretability are not needed:
  - For applications without significant consequences;
  - When the approach is well-studied and verified.
- Interpretability is not fairness, accountability, trust or causality.

- Interpretable models:
  - Options:
    - Before building any model (understanding data etc.).
    - Build an interpretable model.
    - Make model interpretable after training.
  - Before building a model:
    - Exploratory data analytics.
    - Many more choices than simple data statistics; including distribution analytics.
    - More participation from HCI, data visualization, psychology etc. needed.
    - Use examples to explain data, e.g., using k-means or k-nearest-neighbor.
  - Build a new model:
    - Fit rules for a learned classifier.
    - Fit a simpler model per feature, a human can then parse each feature at a time.
    - (Remark: unclear how this is suitable for computer vision, natural language or text applications.)
    - Example-based methods; it has been shown that humans also think in examples, expecially experts.
    - Example-based approaches also applicable to text and other data modalities.
    - Other options: sparsity, distillation, mimicing models, enforce monotonicity.
  - Build interpretable models:
    - Lots of interesting papers at CVPR 2018.
  - Make models interpretable after training¿
    - Ablation test: train models without specific features to see their importance; but very expensive when done naively.
    - Fit linear models to see importance of features (locally to samples).
    - These local explanations might be contradictive.
    - Saliency maps are not always meaningful: they might look similar even for randomized networks.
    - it is unclear what these saliency maps represent and visualize.
    - Concept-based approaches.

- How to evaluate interpretability methods:
  - Do human experiments (that are measurable).
  - For example, give explanation and ask people whether the model did the right thing.
  - Formulate experiments with ground truth if possible.

- Questions:
  - How to use insights to improve models.
  - Several works on that, so it is definitely possible.
  - Are the local importance methods meaningful? A feature can be unimportant locally, but important globally.
  - There are cases where interpretability is not wanted, examples?
  - Example: Geico, because people would try to "game the system".

Laurens van der Maaten: Dos and Don'ts when using t-SNE to understand Vision Models

- Introduction to t-SNE:
  - Goal: build a map of given data in two- or three-dimensional space.
  - Simple approach: principal component analysis (PCA).
  - But PCA is not helpful when classes/categories are unknown.
  - Why is PCA not the right method?
    - Linear mapping, pretty limiting.
    - PCA basically minimizes a sum-of-squared error distance.
    - It will focus on the large distances (because of the squared distance).
    - But large distances or not necessarily trustworthy or meaningful.

        - ∗ Goal of t-SNE and other methods: capture local structure better, not only large distances.
    - Compute pairwise similarity between data with normalized Gaussian kernel.
    - Gives a distribution over pairs of points¡ probability proportional to similarity,
    - Measure normalized student-t similarities in the t-SNE map (i.e., of the dimensionality-reduced points).
    - Minimize the KL-divergence between low- and high-dimensional distributions.
    - The KL-divergence preserves local data structure.
    - The heavy-tail student-t distribution corrects volume differences between the two spaces.
    - Naive implementations are quadratic in the number of data points; not great for large datasets.
    - There are very effective approximations in $\mathcal{O}(N \log N)$ or $\mathcal{O}(N)$.
    - Approximations are possible by grouping interactions of groups of points, e.g., using the center of masses for different groups.

- Do's of using t-SNE:
    - Use t-SNE to get some qualitative hypothesis on what features capture. Meaning, do the clusters found by t-SNE represent the classes/categories the features are supposed to represent.
    - Be creative regarding the inputs to t-SNE.
    - Be creative in how you visualize the outputs of t-SNE.

- Don'ts of using t-SNE:
    - t-SNE cannot proof any point or used to compare methods.
    - It is important to not forget alternative explanations/hypotheses.
    - Do not assign meaning to the distances across empty space.
    - The distance between clusters is not meaningful, does not say anything!
    - t-SNE cannot find outliers and local density of points is not meaningful, it's there by construction.
    - t-SNE is specifically designed to remove outliers.
    - Don't forget that scale / perplexity matters!
    - Perplexity is effectively the number of neighbors each data point want to have.
    - There are local minima in the objective, implementations may also be approximate. This will manifest as splitted clusters.
    - Remember that low-dimensional metric spaces cannot capture non-metric similarities.

- Conclusion: t-SNE never produces conclusive evidence!

- Questions:
    - People use t-SNE for showing that features are discriminative. Is there a way to put a number on that?
    - Well, accuracy seems most meaningful.
    - Clustering is often applied on top of t-SNE, how meaningful is that?
    - t-SNE does not generalize, it does not even allow sample extension. So clustering on top does not make a lot of sense apart from data analysis.
    - There was a paper on automatic selection of perplexity, any thoughts on that?
    - Automatic selection should be used with a sceptical view. Often the hyper-parameter is only hidden. Setting perplexity fully automatical might not even be possible.
    - Can t-SNE be applied on tp of lower-dimensional representations?
    - Yes, pixel distances do not always make sense and it might be inefficient.

Bolei Zhou: Revisiting the Importance of Single Units in Deep Networks

- What is a unit doing?
    - Gradient-based visualization.
    - Iteratively optimize an image to activate a unit.
    - Reveals the patterns activating the unit.
    - Use test images to record activations for specific units.
    - For example, rank images by activation value.
    - Also visualize the unit's receptive field.
    - Allows to come up with different interpretations of units; meaning to link activations to concepts, which is non-trivial.

- How to compare units and interpret all units?

  - Amazon Mechanical Turk study with humans.
  - Identify units that "detect objects".
  - Link units to concepts using semantic segmentation datasets.

- How are interpretable units relevant for the prediction?
  - Remove relevant units and see how accuracy changes.
  - The unit importance is the relative change in accuracy or performance.
  - Here, interpretability seems to mean that a unit is correlated to the final classes.
  - There is no correlation between interpretability and unit importance. What does that mean?
  - Analyzing highly selective units might be misleading regarding the overall classification.
  - Dropout helps to scatter information of classes over all units.
  - Otherwise, individual units might be very important for accuracy, resulting in a large drop in accuracy when removed.
  - The conclusion might be that dropout and batch normalization somehow influence the learned representations.

- From interpretable units to explainable models:
  - Generate explanations using another module, i.e., another black-box model.
  - Instead, we would prefer some self-explanation of the model.
  - Self-explanation by highly activated units.

- Future directions:
  - Network compression.
  - Network defenses.

- Conclusion: why care about interpretability?
  - From alchemy of deep learning to chemistry of deep learning.

Andrea Vedaldi: Understanding Deep Networks using Natural pre-Images, Meaningful Perturbations, and Vector Embeddings

- What does the network actually learned to do?
  - Deep networks contain several encoders, encoding an image into some representation or code.
  - Generic iconic examples:
    * How to get a sense about what information the code has about the image.
    * Look at images that are mapped to the same/similar code?
    * The network is expected to build up invariances when going from image to deeper lazers.
    * So find the "pre-images" of a single code; a set of images mapped to the same code.
    * Starting from random noise, match the code via direct optimization.
    * This way, we can sample from the set of pre-images.
    * However, does not work very well when done naively.
    * Because neural networks are meaningless outside their training domain, so obtained images are often random noise; need to restrict to natural images.
    * We need to know what a natural image is ...
    * Constrain to pseudo-natural images instead.
    * For example, using deep image prior or generative networks. Focus on deep image prior.
    * (Short introduction of deep image prior.)
    * The deep image prior basically favors natural images over random noise (because the structure in natural images is fitted more easily).
    * Goal: inverting codes via deep image prior.
    * The reconstructions of individual layers basically shows what invariances the layers learn or can represent.
    * Implies that the information in high fully-connected layers is both visual and semantic.
    * Same framework can be used for activation maximization.
    * Using the deep image prior is just one option.
    * Alternative: use a generative adversarial network. Or different types of priors, for example empirical priors.
  - Attribution:
    * What parts of an image are salient for a network.
    * Saliency by backpropagation; basically saliency analysis.
  - (Talk not complete.)

Personal Conclusion This tutorial gave a broader overview over what interpretability means and in which cases we should be interested in having interpretable models. Still, when it comes to applications on deep networks, as discussed by Vedaldi and Zhou, we are back at visualizing units or reconstructing activations – very similar to the approaches discussed in the earlier tutorial. Personally, I found the talk on t-SNE most useful for practical research in deep learning and interpretability.

4. Tuesday – Session 1-1C: 3D Vision I

Rotation Averaging and Strong Duality

- Estimate absolute camera rotations from relative ones.
- Without noise, a solution is guaranteed to exist; with noise that is not guaranteed.
- Dual formulation of rotation averaging results in a max-min formulation in the rotations and the Lagrange multipliers of the rotation constraints (e.g., for simplicity, an orthogonality constraint).
- The primal is non-convex in general, while the dual is convex; weak duality always holds, i.e. the dual problem will always be an under-estimator of the primal problem.
- If we can establish strong duality, the problem could be solved efficiently as equality between dual and primal holds.
- Does strong duality hold? Only if noise levels are not too severe.
- Contributions: sufficient conditions for strong duality; simple and scalable algorithm for dual problem to solve rotation averaging.
- Conclusion: rotation averaging can be solved with global optimality without too high noise levels.
- Bounds could be strengthened even further.
- Question:
  - What can be defined as high or low noise here?
  - There is no relation between the angular residuals and the noise levels yet, but working on it.
  - What about outliers?
  - There needs to be a pre-processing step to filter out outliers. The current theory does not handle outliers; strong duality will than not hold.

Hybrid Camera Pose Estimation

- Task: camera pose estimation.
- Given an SfM model and its images.
- Find the 6 DOF pose of a query image.
- Important for AR, VR and mobile robotics.
- Most common approach is using 2D-3D key point matching, providing geometric constraints that can be solved efficiently.
- Disadvantage: good matches and 3D points needed.
- Structure-less camera pose estimation using 2D-2D matches only.
- Disadvantage: very slow.
- Idea: mixing matches between 2D-3D and 2D-2D.
- Is this possible and how to integrate this into RANSAC?
- Yes it is possible, they propose 9 novel solvers for different combinations and settings.
- Propose H-RANSAC to handle hybrid matches.
- In each iteration, H-RANSAC first decides which solver to use (i.e., 2D-3D or 2D-2D). The probability distribution over these solvers favors exploration and variety to avoid local minima.
- Termination criteria similar to RANSAC but taking into account the different solvers.
- In practice, more inliers can be found.
- The hybrid approach allows a better distribution of inliers over the image, both foreground objects (2D-3D) and background (2D-2D).

Certifiably Globally Optimal Solution to the Non-Minimal Relative Pose Problem

- Problem: Given 3D points (meaning the corresponding projection rays in images), find relative pose up to scale.
- Existing approaches: 8-point, normalized 8-point, some iterative approaches.
- None of these approaches is able to solve the problem with globally optimal solutions (certified).
- Suboptimal solutions might be prone to local minima.
- Why is relative pose optimization hard.

- Quadratic, highly non-convex.
- Contribution: Convex relaxation, with empirically tight bound allowing to recover an optimal certificate of the problem.
- Quadratically constrained quadratic problem (QCQP) formulation.
- Relax the problem, hope that the relaxation is tight in terms of the optimal solution.
- Using redundant constraints to obtain tighter relaxations.
- (See paper for details.)
- Conclusion: traditional solvers are suboptimal. A certifiable global solver base don tight relaxations was proposed.

Single-View Stereo Matching

- Monocular depth estimation.
- But supervision is very difficult to optain.
- unsupervised and semi-supervised approach provide suboptimal performance.
- Motivation: spatial transformation network, stereo matching usually has higher quality and better generalization.
- Approach: monocular depth estimation can be reformulated as stereo problem with randomly generated second image.
- Allows to enforce geometric constraints and allow better generalization.
- First, synthesize right image from left image. Then, perform stereo.
- Quantitative and qualitative results show improvements, e.g., on KITTI.
- Also better generalization from training on KITTI but testing on other datasets.
- Also compared against other stereo methods.

Fight Ill-Posedness with Ill-Posedness: Single-Shot Variational Depth Super-Resolution from Shading

- Problem: depth super-resolution.
- Motivation: tackle depth super-resolution and shape-from-shading jointly.
- A variational formulation of the joint problem is proposed.
- Also comparison to multi-view approaches.
- Allows fine details compared to other approaches.

Deep Depth Completion of a Single RGB-D Image

- Problem: depth from RGB-D sensors is not complete due to lighting, reflections, complex geometry etc.
- Goal: depth completion taking care of larger holes.
- Depth completion requires a combination of local and global features.
- Approach: FCN to estimate the surface normal map first, because it is easier compared to absolute depth estimation, and normal vectors are easier to regress.
- Then, estimating depth from normals is more robust.
- Qualitative results look impressive, allowing to complete large holes.

PPFNet: Global Context Aware Local Features for Robust 3D Point Matching

- How to find local correspondences in 3D point clouds?
- There are many hand-crafted features, but performance is not good on cluttered, noisy scenes.
- Related work: 3DMatch.
- Proposed approach uses deep learning but operates on point clouds directly.
- (See paper for architecture details.)

FoldingNet: Point Cloud Auto-Encoder via Deep Grid Deformation

- Encoder, directly operating on 3D point clouds allowing to extract code words for shape classification.
- Assumption: 3D point clouds are often sampled from surfaces; these can be constructed by folding papers.

- The network essentially deforms a flat paper to mimic specific shapes (the animation in the presentation is quite illustrative).
- FoldingNet's decoder can also be used for interpolation, codewords can be used for classification.

A Paper-Mache Approach to Learning 3D Surface Generation

- Goal: learn to generate meshes.
- Basically predicts multiple patches/surfaces parameterized by points constrained to the surface.

LEGO: Learning Edge with Geometry All at Once by Watching Videos

- Unsupervised 3D geometry learning from videos.
- At testing, given an image, depth, geometric edges and normals are predicted.
- Motivation: Zhou et al., CVPR 2017.
- Different networks, estimate depth, edges and camera poses from source and target image.
- Object boundaries are preserved better.
- Predicted normals are less noisy and also adhere boundaries.
- In the shown video, however, the normals did not look very noise-free.

Personal Conclusion Surprisingly, the three full orals in this session were focused on classical optimization problems in 3D vision, without any machine learning. In the spotlights, as well, some papers were not using any deep learning. The remaining spotlights combine tasks or disregard classical task boundaries by, e.g., using stereo to solve monocular depth estimation, solving depth super-resolution together with shape-from-shading or predicting surface normals for completing depth maps. Deep learning on point clouds also seems to become very interesting for several tasks. Personally, my favorite work is the FoldingNet paper.

5. Tuesday – Session 1-2A: Machine Learning for Computer Vision I

Deep Layer Aggregation

- Two trends: better building blocks and skip connections.
- Question 1: how to make these trends compatible?
- Skipped connections as layer aggregation.
- Hierarchical deep aggregation (in a tree structure, see paper for illustration).
- (Deep layer aggregation seems to be the main contribution, was not clear at the beginning of the talk.)
- Results in a better parameter-performance tradeoff.
- Iterative deep aggregation (see paper for illustration).
- Improves segmentation performance.
- Better transferability across datasets.

Convolutional Neural Networks with Alternately Updated Clique

- Background: deeper and deeper networks in computer vision, up to ResNet networks, always implying better performance.
- How to further maximize information flow in deep networks.
- Do feedback connections help?
- Motivation: Attention mechanisms.
- New block: Each two layers have both forward and backward connections.
- New update rule for alternating updates of these connections, after initializing the network using feed-forward only.
- Allows to reduce parameters while holding performance.

Practical Block-Wise Neural Network Architecture Generation

- The trainability of networks highly depends on the architectures.
- Skip connections result in smoother loss functions.
- Currently: hand-crafted networks.
- Auto-generated networks: have been researched before, by grid search, reinforcement learning etc.
- BlockQNN: block-wise design for CNN, find optial block with Q-learning and an early stopping strategy.
- A CNN is represented by a DAG.

- Construct network by stacking blocks sequentially.
- Designing with Q-learning: current state is the status of the current layer, action is the decision for the next successive layer.

Residual Dense Networks for Image Super-Resolution

- Propose a residual dense block for image super-resolution.
- (See paper for details.)

Attentive Generative Adversarial Network for Raindrop Removal from a Single Image

- Problem difficult because raindrop locations not given and information at randrops completely lost.
- Attention injected into generator and discriminator.
- Generator: Attentive-recurrrent network and contextual auto-encoder.
- Dataset: 1000 image pairs of vraious rain types and scenes.
- Looks quite convincing compared to Photoshpp and pix2pix.
- Attention mechanism is shown to focus on the raindrop locations.

FSRNet: End-to-end Learning Face Super-Resolution with Face-Priors

- First network with facial geometry prior for super-resolution.
- Coarse super-resolution network, followed by fine super-resolution network and a prior estimation network (both basically encoders) then a decoder reconstructs high-resolution image.
- Also use adversarial training.
- Looks OK, still a bit blurry; quantitatively not necessarily better.
- Landmark and parsing results (used intermediate in the network) are competitive to state-of-the-art approaches, although they are based on lower-resolution images.

Bursed Denoising with Kernel Prediction Networks

- Trained exclusively on synthetic data, by synthetically creating burst noise.
- Kernel prediction architecture: Network predicts local kernels, result is the sum or kenel-products with local patch.
- It is argued that kernel prediction networks are interpretable.

Unsupervised Sparse Dirichlet Net for Hyperspectral Image Super-Resolution

- (See paper.)

Dynamic Scene Deblurring Using Spatially Variant Recurrent Neural Networks

- Problem: dynamic scene blur.
- Motivation of the approach: deconvolution.
- (See paper for more details.)

Crafting a Toolchain for Image Restoriation by Deep Reinforcement Learning

- Images often subject to a sequence of blur and compression.
- Contribution: dynamic toolchain for efficient, and transparent image restoration.
- Image restoration as decision-making process, the agent can apply different tools to the image.
- 12 restoration tools, using CNNs against Gaussian blur, Gaussian noise and image compression.
- Tool selection as reinforcement learning problem.
- For training, loss is computed throughout the full toolchain applied.
- Competitive performance with less computations.

Personal Conclusion The three orals were mostly concerned with devising novel and better-performing architectures for standard computer vision tasks such as image recognition. Interestingly, the performance-complexity trade-off (e.g., in terms of the number of parameters) was mentioned several times. This might indicate a slight change in architecture philosophies towards lighter models (possibly easier to train) while retaining state-of-the-art performance through innovation regarding the network structure. The spotlights had a similar theme: while not being focused on the performance-parameters trade-off, novel architectures for various tasks such as super-resolution and image restoration were presented. None of the presented papers really convinced me that the presented architecture concept is “the next big development in computer vision”.

6. Tuesday – Poster Sessions P1-1, P1-2 and P1-3

The Unreasonable Effectiveness of Deep Features as a Perceptual Metric

- Perceptual metrics such as PSNR and SSIM do not capture the nuances of human perception.
- Deep features have been found to improve image generation when employed as “perceptual loss”.
- How meaningful are these perceptual losses?
- Experiments show that deep features from networks trained for complex tasks show close to human performance in rating similarity of images.

Deep Sparse Coding for Invariant Multimodal Halle Berry Neurons

- Top-down feedback and sparse coding seem to be two elements of more interpretable models in terms of what bottleneck neurons actually learn.

LDMNet: Low Dimensional Manifold Regularized Neural Networks

- Assumption: Data actually lives on several low-dimensional manifolds.
- General networks do not learn feature that correspond to these manifolds.
- A reason might be overfitting.
- This happens if the learned manifolds have higher dimensions than the true ones.
- One solution: regularize the manifold dimension.

What do Deep Networks Like to See?

- An auto-encoder is trained to good reconstruction performance.
- The encoder is then fixed.
- The decoder is fine-tuned by training the auto-encoder together with a classifier, the classifier taking the reconstructed image as input.
- The fine-tuned decoder allows to reveal what the classifier “likes to see”.
- Could that be an interesting approach to interpretability?

Lightweight Probabilistic Deep Networks

- Each layer should generate a distribution, in their case, a Gaussian distribution.
- To ensure that, commonly known activation function can be re-formulated and simple be replaced; no other changes to parameters or architectures are needed.

Defense Against Universal Adversarial Perturbations

- Essentially a detection scheme for universal adversarial examples.

Personal Conclusion Personally, I had the impression that the following topics were addressed by many posters: 3D vision in general, but using different shape representations in particular (e.g., surfaces, point clouds etc.); attention mechanisms and using attention mechanisms in applications; generative adversarial networks and adversarial losses in various applications and configurations; block-based or module-based neural networks, e.g., where tasks are addressed jointly or networks for different tasks are combined to achieve an overall task.

Wednesday – Session 2-1B: Machine Learning for Computer Vision III

Efficient Optimization for Rank-based Loss Functions

- Problem: ranking, rank relevance of images according to query.
- Collect dataset, learn ranking model, which scores each sample and sorts the samples by score.
- Ideally one would like to directly optimize the rank loss.
- Learning to rank: computing the gradient of the ranking loss involves solving an optimization problem.
- Contributions: properties of efficient optimization of ranking loss, efficient algorithm.
- Negative decomposability property and interleaving dependence property.
- (See paper for details, these properties impose a partial ordering on the scores and the ranking used for efficient optimization).

- Algorithm:
  - Induce the partial-ordering structure by sorting positive and negative samples.
  - Compute the optimal interleaving rank for each negative sample independently.
  - These steps can be optimized to achieve a runtime in $\mathcal{O}(NlogP)$ with N the number of negative samples and P the number of positive ones.
- As result, performance increases over a simple 0-1 loss while not increasing the runtime significantly.
- Conclusion: optimizing ranking-based loss functions improves performance.

Wasserstein Introspective Neural Networks

- Generative models: maximum likelihood framework, generative models from discriminative models (e.g. introspective networks), generative models with discriminative models (e.g., generative adversarial networks).
- Goal: reduce the number of cascade to a reasonable number (see related introspective networks paper).
- Endow modern CNN classifiers with generative capabilities that enhance robustness in classification.
- WINN (Wasserstein Introspective Neural Networks):
- Train a classifier to distinguish positives from pseudo-negatives.
- Generate samples from the model and add them to the data.
- Repeat until model learns the target distribution.
- Wasserstein loss with gradient penalty is used for classifier training; means to maximize the energy between real and pseudo negative samples.
- Applications: image generation, texture synthesis

Taskonomy: Disentangling Task Transfer Learning

- Are all our computer vision tasks related or not?
- Show: task relationships exist and can be measured.
- Is any set of task related?
- Redundance and recycling allows efficiently in needing less labeled data.
- If we could quantify these relationships would allow to use these redundancies for tasks.
- Would allow to solve many tasks with labeled data for few, or solve new tasks without labeled data.
- Result: Taskonomy.
- Set of tasks selected (not comprehensive).
- Dataset: 4 million real images with ground truth for all tasks.
- Train task specific networks.
- Transfer modeling: train all possible transfer functions; provides complete directed graph, but weights need normalization.
- Normalize adjacency matrix.
- Most of the relationships are weak, but there are also som really strong ones.
- Graph allows to deduce source tasks for a target task, or source tasks for a new task, also including higher-order relationships.
- For testing, some tasks have very few data only used for training the relation networks, not the task-specific networks.
- Using the found source tasks allows a significant gain for transfer learning.

Maximum Classifier Discrepancy for Unsupervised Domain Adaptation

- Problem: cost to collect labeled samples.
- Solution: transferring knowledge between different domains; difficulty: difference of domains.
- Tackle unsupervised domain adaptation; labeled source and unlabeled target samples.
- Popular approach: shared feature generator for both domains, and put a classifier on top. The feature generator is trained to fool a domain classifier.
- Known to work well for various tasks.
- One problem: generator can predict features that are close to the decision boundary, these are not discriminative.
- Proposed method: task specific classifier (on top of feature generator) distribution alignment.
- Idea: features that overlap between both domains may be mis-classified, so the goal is to align the feature distributions better.

- Two training steps:
  - Maximize discrepancy for fixed feature generator.
  - Minimize error on source images.
  - Minimize discrepancy for fixed classifiers.

‘

- Applications: classification and semantic segmentation.

Unsupervised Feature Learning via Non-Parametric Instance Discrimination

- Motivation: human labels hard to obtain.
- Currently: clustering, generative modeling, self-supervised feature learning.
- Goal: learn a feature representation useful for applications.
- Past work faces limitations: trained with static losses on low-level cues; inconsistencies between training and testing.
- (Exact approach not clear.)
- Result is lower-dimensional features with better classification accuracy; also show nearest neighbor results on these low-dimensional features.

Multi-Task Adversarial Network for Disentangled Feature Learning

- Motivation: the image generation process consists of several independent factors, only some of them of interest.
- What if we are only interested in one primary factor, but for learning also need to tackle the others.
- Do we need to really consider all factors to be able to generalize.
- Learn from two adversarial tasks, feature should e good for content recognition but not for differentiating style.
- Results in more robust models against the secondary factors (e.g., illumination for face recognition).

Teaching Categories to Human Learners with Visual Explanations

- Goal: teaching categories to humans, so an algorithm for automatic teaching to humans in expert settings.
- Current approaches: human needs to find features for himself, is not given explanations.
- The proposed model is used to select teaching examples and also shows an explanation (by highlighting a region).
- Comparison against teaching baseline algorithms without explanations.

Personal Comments The best-paper award “Taskonomy: Disentangling Task Transfer Learning” was definitely a highlight in this session. Personally, I found the paper very interesting as it systematically addresses the transferability of models across tasks. To date, this has mostly be done by “common sense” or based on observations of the human visual system. This is, to the best of my knowledge, the first work systematically investigating which combinations of problems should be learned jointly or benefit from each other. Other interesting work was using machine learning to improve automatic learning environments for humans in expert domains.

7. Wednesday – Session 2-1C: 3D Vision III

Modeling Facial Geometry using Compositional VAEs

- So far: global, PCA-like models.
- Better: local models to also consider high frequency structure.
- Problem is among others the single bottleneck preventing to capture high frequency details.
- Solution: multiple layers of latent variables.
- U-Net like architecture, with stochastic skip connections, basically has multiple latent variable spaces for lower and higher resolution.
- Allows to interpolate high or low-level features separately.

Tangent Convolutions for Dense Prediction in 3D

- Goal: semantic segmentation of 3D data.
- Should be able to be efficient for million sof points, indoor or outdoor.

- Introduce tangent convolutions directly operating on surfaces and implemented efficiently.
- Locally approximate observations with planes.
- (Seems to assume pretty dense point clouds.)
- Local structure is projected onto these planes, resulting in so-called tangent images.
- Results compared to CNNs, OctNets, PointNets, ScanNet.

Neural 3D Mesh Renderer

- 3D to 2D projection for neural networks, allows backpropagation through the 2D rendering process.
- Meshes would be the best suited 3D representation, but they are not yet commonly used for neural networks.
- Backpropgation of a mesh-renderer is non-trivial due to standard rasterization which is highly non-differentiable.
- Solution: replace the sudden change in the rasterization with a gradual change, allowing a gradient.
- Application: single-image 3D reconstruction; more accurate than voxel methods.
- Application: 2D to 3D style transfer.

Structured Attention Guided Convolutional Neural Fields for Monocular Depth Estimation

- Depth estimation from one image.
- Based on deep convolutional neural fields.
- Also motivated by multi-scale prediction of depth maps.
- Use attention as guidance for feature learning?
- Model: encoder, and an attention guided CRF for multi-scale feature learning.
- (See paper for details, slides unclear.)
- Mean field approximation allows end-to-end learning.

Personal Comments Personal favorite work is the compositional VAE. Although I can imaging that this is not the main contribution as there have been hierarchical versions of VAEs before, it is simple, has a clear motivation and allows to capture multi-level (i.e., high- and low-resolution) elements separately. This session also included the "Neural 3D Mesh Renderer" paper which I recommend to everyone intending to use deep learning for 3D tasks.

8. Wednesday – Session 2-2C: Computational Photography

Illuminant Spectra-Based Source Separation Using Flash Photography

- Problem basically intrinsic image decomposition using flash photograph up to specific light sources.
- Extended image formulation to contain flash image.
- Still results in a highly under-determined and non-linear.
- However, image and flash image can together be used to identify the flash-induced shading.
- This allows to easily solve for the reflectance and receive the so-called gamma image.
- This leaves to identify illumination from the shading for both images.
- Also allows to estimate the light sources, up to three.
- Results look quite impressive (without knowing related work).

Trapping Light for Time of Flight

- Problem with capturing a complete model: every camera can only get one view per image.
- Solutions: multiple cameras, rotating the object, registration.
- Or: generate multiple views with mirrors.
- Still needs to find correspondences.
- Time-of-Flight imaging to get more detailed depth and complete reconstructions.
- But with a single time-of-flight sensor, there is still the problem of a single view.
- Solution: Create a light trap, at some point the ray will hit the object.
- The ray can be traced back through the light trap to obtain the depth, i.e. the location of the object.
- What is the best geometry of light traps?
- Impressive results for some objects, when using a pyramid as light trap.
- Questions:
  - It is important to use a time-of-flight sensor, where only the first ray is considered!

The Perception-Distortion Trade-Off

- Image restoration: obtain high quality version of an image, e.g., super-resolution, denoising etc.
- Goals: reconstruction should be similar ground truth and good perceptual quality.
- Contribution: it is impossible to achieve both low distortion and god perceptual quality.
- Empirically, no algorithm is good at both.
- Usually people assume squared error not to be a good perceptual metric between images.
- But the empirical evidence shows that it might be independent of the used metric.
- This trade-off can be formalized and mathematically proven.
- (See paper for derivation.)
- So is it possible to traverse along the lower bound characterizing the trade-off?
- Yes, already done by GAN-based image restoration.
- Implication: no single algorithm performing well in all cases, it is application dependent.

Label Denoising Adversarial Network for Inverse Lighting of Faces

- Applications: lighting transfer and image forensics.
- Task: Lighting estimation.
- Problem: not enough training data, so use a state-of-the-art estimatore; ground truth needs to be denoised.
- Idea: transfer the features of real images to the synthetic images without noise.
- Train a network on synthetic data, split into feature network and task network.
- Fix lighting net, and train a new feature net matching the feature distribution of the synthetic data.

Optimal Structured Light a la Carte

- Multi-pattern light triangulation is the most robust way for obtaining point clouds.
- What is the optimal k-pattern sequence to use?
- Contribution: optimal pattern generator and corresponding decoder.
- Instead of proposing a new pattern, an algorithm is proposed taking arguments such as # patterns or other specifications as input.
- It outputs the optimal k-pattern.
- Allows to tune pattern to specific settings.

Tracking Multiple Objects Outside the Line of Sight

- (See paper.)

Inferring Light Fields from Shadows

- Goal: infer objects and their depth from observing only shadows.
- The hidden elements are assumed to be planar and diffuse, smooth and occlusions are negligible.
- But setting still constrained (see paper for examples).

Personal Conclusion Personally, my two favorite papers from this session are "Trapping Light for Time of Flight" and "The Perception Distortion Trade-Off". The former uses simple mirror setups (e.g., a simple pyramid) and a time-of-flight sensor to get accurate point cloud representations of shapes. The advantage is that only one sensor is needed and the object can still be reconstructed with impressive detail. The latter provides theoretical and experimental evidence of the so-called perception-distortion trade-off for image restoration tasks. Specifically, for image restoration tasks where exact inversion is not possible, there will also be a trade-off between removing the actual distortion and generating perceptually appealing results – algorithms cannot achieve both perfectly.

9. Wednesday – Session 2-2B: Object Recognition and Scene Understanding III

Cascade R-CNN: Delving into High-Quality Object Detection

- Current object detection can be quite noisy.
- How to train a high quality object detector.
- Simple answer: increase IoU threshold.
- Two problems: training overfitting because of vanishing positive examples.
- Inference time quality mismatch.
- Multi-stage framework, increasing IoU threshold in a cascade, from stage to stage.
- Makes the detector gradually more selective, but it reduces overfitting and avoids inference-stage quality mismatch as the same model is applied at test time.

Functional Map of the World

- Classification of satellite image with bounding boxes and temporal views.
- Allows temporal reasoning, e.g., construction site or office buildings.
- Also allows to consider temporal information.
- Created a dataset, but data is inherently difficult to annotate.
- 1M+ images with 4/8 band images, 63 categories.
- Experiments show that using temporal information is required.

MegDet: A Large Mini-Batch Object Detector

- Batch size in training is slower for detection than for classification.
- This leads to unstable training.
- Also leads to inaccurate BN statistics.
- MegDet: the first large batch detector.
- Trained on multiple-devices, paper shows how to split the BN statistics over devices.
- Also consider different learning rate policy.
- Faster training, but also multiple devices.

10. Thursday – Session 3-1C: Applications

Direction-Aware Spatial Context Features for Shadow Detection

- Shadow detection is a fundamental problem, goal is to get a shadow estimation in the form of a segmentation map.
- Widely studied since 1990.
- Data/driven approaches learn features using deep neural networks, but results not perfect.
- Global image context is important for shadow detection.
- The context can ba analyzed in a direction aware manner.
- Spatial recurrent neural network approach, generating spatial context features allowing to propagate features from all directions to each pixel.
- Additionally, an attention mechanism is used to focus on specific directions.
- Outperforms related work by quite a margin.
- Qualitative results seem improved, specially for shadows covering multiple backgrounds.
- In the ablation study, however, a large part of the improvement comes from the network, even without using the spatial context modules ...

Discriminative Learning of Latent Features for Zero-Sot Learning Recognition

- (See paper.)

Learning to Adapt Structured Output Space for Semantic Segmentation

- Cross-domain semantic segmentation, between source and target domain.
- For example, source is synthetic and target is real.
- In practice, neural networks do not generalize well.
- Observation: large gap in appearance, but smaller gap in the semantic segmentations.
- Feature space adaptation: shared feature network, loss for semantic segmentation for synthetic data and discriminator to fool (adversarial loss).

Multi-Task Learning using Uncertainty to Weight Losses for Scene Geometry and Semantics

- Scene understanding is a multi-task learning problem, including geometry and semantics.
- Simple network with shared encoder and individual decoder for individual tasks.
- Really important to weight the losses appropriately.
- Want to learn the weighting, considering task uncertainty.
- Weight should depend on magnitude of output and difficult of task, uncertainty captures both.
- Improves performance over uniform weighting or finding weights by grid search.

Jointly Localizing and Describing Events for Dense Video Captioning

- Dense video captioning: video captioning but with longer text *and more details?).

Going from Image to Video Saliency: Augmenting Image Salience with Dynamic Attentional Push

- Explicitly modeling attentional push. objects pushing the viewers attention to specific other objects.
- So far: static saliency models base on state-of-the-art saliency detectors and an attentional push module.
- Problems for videos: long-term temporal fusion of features/information and also changes the way humans attend to objects.
- Attentional push seems to be stronger in videos.
- Actor gaze, cuts in videos and bounces of attention are important in videos.
- LSTM model with good static saliency model.

M3: Multimodal Memory Modeling for Video Captioning

- Automatically generate a sentence to describe a video.
- Challenges: learn good features and good language model.
- CNN video encoder, multimodal memory module allowing to read and write elements, and a LSTM text decoder. Both encoder and decoder can read and write form the memory.
- (See paper for equations.)
- Multimodal memory is just modeled as matrix.
- Performance over related work improves slightly.

Emotional Attention: A Study of Image Sentiment and Attention

- Visual attention is important for several applications, HCI, robotics etc.
- How do sentimental properties influence attention?
- Developed eye tracking dataset with images with strong emotional properties.
- Images annotated with 33 attributes, positive, negative and neutral ones.
- Attention is higher for positive or negative images compared to neutral ones.
- Also propose a new network architecture to take into account the context (meaning sentiment, I assume) of the image.
- Model achieves state-of-the-art performance. It captures the emotional elements of images, but only qualitatively.

A Low Power, High Throughput, Fully-Event Based Stereo System

- Stereo correspondence system implemented on event-based digital hardware.
- 2000 disparity maps per second, with lower power requirements.
- Implemented on neuromorphic chips.
- (See paper for approach details.)

VITON: An Image-Based Virtual Try-On Network

- Virtual try on of clothes.
- Three desired properties: pose should be preserved, clothing items deform naturally and details of clothing should be preserved.
-
- Lacking full supervision and no 3D data of cloth.
- Approach: clothing agnostic person representation; encoder/decoder network for coarse result: shape context for refinement.

Multi-Oriented Scene Text Detection via Corner Localization and Region Segmentation

- (See paper.)

Multi-Content GAN for Few-Shot Font Style Transfer

- Data dropout in font domain, i.e., only few characters per font given and goal is to generate text in specific fonts.
- Glyph network first generates rough estimates of all glyphs although the data doe snot contain all characters.
- The Ornamation network transfers color to the predicted glyphs.
- Additional regularizers penalize strong deviation and artifacts.
- Results look quite interesting.

Audio to Body Dynamics

- Generate arms and fingers corresponding to piano music.
- Output is a moving avatar playing the music.
- Trained the network on recital videos on YouTube, 5hours, various lighting, resolutions and music styles.
- MaskRCNN to obtain pose on these videos, per frame.
- Results don't look really impressive, but the application idea is quite interesting.

Weakly Supervised Coupled Networks for Visual Sentiment Analysis

- Propose weakly supervised learning only requiring a image-level label for training.
- Predicts a sentiment map (similar to saliency map) first and then uses this to predict sentiment.
- Unclear how the sentiment prediction is supervised.

Personal Conclusion This session was mainly interesting because of its variety in different computer vision applications. Personally, I found the presented work on font style transfer and audio to body dynamics most interesting. For the latter, however, the results look very pre-mature.

11. Thursday – Session 3-2C Object Recognition and Scene Understanding V

Learning Descriptor Networks for 3D Shape Synthesis and Analysis

- Availability of 3D datasets spurred work in discriminative and generative work on 3D shapes.
- Not much work on energy-based models (descriptive models), as alternative to 3D generative adversarial networks.
- Decoder/generator consisting of several convolutional stages; model represented with an energy that is optimized using maximum likelihood.
- However, the derivative of the normalization constant is intractable.
- Thus, MCMC is used as alternative.
- (See paper for derivation of the sampling process.)
- Applications: synthesis, shape recovery/completion, shape super-resolution, shape classification.
- Note: seems similar to 3D ShapeNets.

Neural Kinematic Networks for Unsupervised Motion Retargetting

- Problem: motion retargeting from human to virtual model.
- Previous work uses heuristics, which should be avoided by motion synthesis.
- Problem: these models fail to preserve the target structure (of body parts).
- Motion is split up into global motion and relative motion (motion with respect to root joint).
- (See paper for architecture details.)
- Use RNN to predict retargeted motion; then apply it again to re-retarget the motion to the original structure for supervision.
- Some regularization losses assume smoothness and ensure proper joint angles.
- Does only slightly outperform a simple baseline where the angles are copied.
- Human to model retargeting does not look convincing.

Group Consistent Similarity Learning via Deep CRF for Person Re-Identification

- Given a probe of a person image, find similar persons in a gallery.
- Current limitations: only local constraints are used (in the sense f possible pairs), e.g., triplet loss, quadrupel loss etc.
- New loss to learn more accurate similarities between image pairs within groups.
- Neural network for similarity estimation and a new group similarity refinement model that is supervised by group similarities.
- Refinement model implemented as CRF.
- (See paper for unary and pairwise terms.)
- Optimized using mean field approximation; optimization is then unrolled as network.

12. Friday – How to be a Good Citizen of the CVPR Community

Organizers: Devi Parikh, Dhruv Batra
Webpage: https://www.cc.gatech.edu/~parikh/citizenofcvpr/

How to Write a Good Paper

- Only very good papers have a significant impact on the career
- There is an opportunity cost for writing only good or OK papers
- The scientific community today is like a crowded marketplace, lots going on, everyone wants attention
- Teaser figure seems to be expected in CVPR and SIGGRAPH.
- Introduction: write a dynamite construction.
- Bill likes one section or paragraph on the main idea - using an illustration maybe.
- Simple examples are important
- Conclusion:
  - What does this work open up?
  - How not to end a paper: future work section
- General writing tips:
  - Anticipate the needs of the reader
  - Omit needless words
  - Write short and concise: more direct and more space
  - Figures and captions: only the figures and captions should also tell the story, so caption should be standalone
  - The caption should point the reader to the main points of the figure that the reader should notice, so why the figure is there
  - Text should still flow when equations are ignored
  - Be kind in talking about competing work
- Be clear and reliable and honest, build a reputation.
- The area chair's game is how to reject papers, so don't give them reasons.

Rights and Obligations

- Rights: Do research, publish it
- Obligations: Volunteer for conferences, good reviewer
- Reviewing is voting deciding about the community characteristics and it's growth.
- It is about inlier and outlier papers
- Graduate students tend to reject outlier papers, with new or different ideas, but that's not good to the community.
- It needs special care for outlier papers.
- What we are, where we come from and where we go - need to judge these points.
- There are two class of knowledge bases:
  - Books, which are dictatorships, read more than one book.
  - Encyclopedias: Wikipedia or computer vision reference guide.
- Where do we come from: Dartmouth conference, and splitting AI into separate conferences.
- We need to keep the applications and our origins in mind to not get extinct.
- Sounds a bit like DNN is the current state-of-the-art but to not get extinct we need to revisit core questions and goals of the community.
- What kind of AI do we want?

How to Avoid a Clique Culture

- Social interactions: Posters, coffee, lunch, reception, parties ...
- These informal meetings result in collaborations.
- Affects what you read and cite, who you invite, who you want to recruit etc.
- There is a tendency to over emphasize big institutions.
- Try to broaden who you interact with.
- Assume that everybody here has something good to offer.
- Try to introduce people to other people.

Strengthening our Community

- Being a young researcher is stressful, everyone else is doing better – mentors can make a difference.
- Service and Leadership:
  - Volunteering
  - Reviewing
  - Chairs
  - Tutorials, workshops ...

- But don’t over commit to too many roles
- Inclusiveness:
  - Welcome less developed communities
  - Inclusiveness over other disciplines

How to do Research?

- Pick a problem:
  - Area with enough work, but not enough to leave room.
- Making a contribution:
  - Read a lot but don’t believe anything, think about when approaches break.
  - Re-implement the state of the art and look for surprises
  - Understand how and why a system works, do controlled experiments, swap or exchange components
- Many papers that you read were not the papers the authors meant to write
- You want to maximize your contribution, your value to the community
- Publication:
  - Quality over quantity
  - Be willing to bury drafts and move on
  - Do not compromise on methodology or ethics
  - No body will thank you for too many papers with insignificant contributions
  - Community needs meaningful contributions
  - Many papers add to the noise without significant contribution and end up misleading people
  - Publication is not the goals contribution is
  - Publication is a liability towards ones reputation
- Research over time:
  - Keep track of interesting problems
  - See the larger goals
  - Read. A lot.
  - Write down ideas, talk to people, take quiet time for thinking

How to Write a Good Paper

- Rule #1: do good research.
- Parts: Title, abstract, introduction, rest. Spend same time on these parts!
- Title:
  - Should capture what is special about the paper.
  - The title should only be applicable to this paper and shoul dbe memorable.
- Opening lines:
  - Start writing with section 2.
  - Get catchy opening lines.
- Introduction:
  - Most important section, should say everything and the rest of paper is just the evidence for the claims made.
- In general, good writing follows from good reading.
- Experiments:
  - Results and ablation, learn from good papers, e.g. Ross Girshick.
- Figures: should make a talk when considered in isolation.
- Unfortunately, in science the one with good presentation and expose has better chances.

Good Research and Evauation

- Proper baselines: avoid weak or trivial or carelessly implemented baselines.
- Proper evaluation includes standard train/test splits, multiple datasets and statistically significant results.
- Make results reproducible.

Principles to Thrive in the Research Community

- Research direction should be chosen as if in a large team – the CVPR community team.
  - “Don’t crowd the ball”, find your position.
- Think about what will be a popular topic once the current problems are solved.
- As an author, aim to team and surprise.
- Grow and adapt in your career:
  - Keep branching, exploring new topics and roles.
  - Choose activities with potential to learn.

Calendar. Not to-do lists.

- This is a really good talk, so my notes are rather short and I highly recommend looking at the slides!
- The idea is to use calendars to organize to-dos.
- Two calendars: a done calendar and a to-do calendar.
- The to-do calendar contains everything that has to be done. This includes literally everything because everything takes time (sleep, lunch, mails etc.).
- Every evening, done things are moved to the done calendar, everything else needs to be re-planned. For re-planning it is important to prioritize and focus on not wasting time.
- It is also useful to compute a multiplier factor base don the data that tells how far off your time estimates usually are.
- In her talk, she also includes leisure time, but in questions she suggests also using a lightweight version only for professional time.
- Two important tools: backtracking from important deadlines, and anticipating problems!

Being Open

- Principles of research:
  - Quality;
  - Honesty; (a bit utopian);
  - Openness (share code, datasets, papers and work together).
- Open-sourcing:
  - Datasets, libraries, models, ...
  - It allows comparison, baselines and research also for smaller groups with less resources
  - Incentives for open source: community awards for best open-source projects, citation/star counts for open source projects, professors, group leaders and companies should reward open-sourcing.
- Collaborations:
  - Open research horizons;
  - Different experience, different personalities;
  - Results in personal and professional growth.
  - Junior researchers get to learn different working styles, topics and environments.
  - Senior researches get to train the next generation of scientists.
  - Incentives for collaborations: student co-advising, reward collaborations in career evaluation.

Welcome Everyone

- Top 5 attendee countries are from Asia; same observation for sponsors.
- Asian research hubs are en par with the best in the world.
- Strong imbalances: Asia vs. western, men vs. women.
- Important difference: Asian ethnicity does not mean that they represent Asian labs/companies.
- Be aware of these imbalances!

What PCs told ACs for CVPR 2018

- CVPR is ranked first in computer science; we are it.
- Massive challenges from scaling up:
  - Demographics: balance is hard.
  - Money: lots of money, but also opens problems.
  - Loss of coherence: what are our key aspirational problems?
- Strategies:
  - We need regulation, codes of conduct, principles; boring but valuable.
  - Training: lots of people want to help, but don't know how.
  - Exposition: Industry needs to stop hiring senior vision academics right now! Need a big picture, a book or a sense of who we are.
- Principles:
  - Make decisions to help the community.
  - Authors should understand their decisions.
  - Minimize appeals.
- Practices:
  - All decisions based on reviews and discussions.
  - All decisions will have a summary and need to have consensus of two ACs.
  - Not need to tell authors how to write papers.
- In principle self-plagiarism is not allowed!

13. Panel

- Question: very few memory between conferences especially when chairs change drastically. Is that correct?
- Answer: yes and no, on the long scale we were not good in doing that (e.g., lists of bad referees); in the short term, information has been passed on. And it is reasonable that future conference committees do things slightly differently.
- Question: is there any reward system for reviewers?
- Answer: fair implementation might be difficult; try hard to acknowledge reviewers and volunteers, but it is never going to be a motivating awards. And area chairs will remember bad or good reviews.
- Question: how to force industry to be a better CVPR citizen?
- Answer: stop hiring senior academics; there seems to be an underlying assumption that industry just pais better, but that's not necessarily true. However, hiring strategies of companies can create an atmosphere of "cool kids and uncool kid".
- Question: what is the right entry bar for reviewing; and what materials to teach reviewing?
- Answer: examples of reviews will definitely be valuable; at smaller conferences, individual area chairs choose junior people for having an extra (fourth) review and give feedback personally, but that does not scale;
- Question: How to encourage better talks?
- Answer: reviewers already have to do enough to do, so it is hard to judge whether the author is capable of giving a good talk; another part is giving advice on giving good talks. Also the decision of oral is high-variance, so pretending that orals are better than posters should stop.
- Answer: What makes a good poster?
- Question: The poster needs to have the visual aids such that the author can easily explain the key points of the story.
- Question: For teaching reviewers it is necessary to give feedback, current material does not allow this.
- Answer: Giving feedback for reviews is difficult and might cause more confusion.
- Question: How is consensus actually achieved in a large community?
- Answer: The PAMI TC (technical committee) is usually the vehicle of this. The PAMI TC is usually at the first day of the conference. And the PAMI TC is very open in our community.
- Question: the best paper awards seem unreasonable.
- Answer: As the pool of papers scales up, the number of awards should also scale up. But popular vote is very biased towards popular labs and popular problems.
- Question: What to recommend for people at CVPR for the first time?
- Answer: meet people and ask them about their research, even if they are random. Hear random spotlights or talks. Conferences are inherently social events, meeting new people is core.

14. General Impression

Salt Lake City Personally, I found Salt Lake City well chosen for CVPR 2018. The conference center was large enough to hold the $\sim$ 6.5k attendees without being too crowded. Additionally, plenty of hotels are close-by and several restaurants, diners and shopping opportunities are in walking distance. However, the airport is rarely accessible via direct flights from Europe and the city did not offer much sight seeing apart from the Great Salt Lake[1] and the Temple Square[2].

Organization Overall, I think the conference was well organized. There was always enough to eat and drink, and because lunch and breakfast times partly overlapped tutorials, workshops or poster sessions, long queues were avoided. The poster and exhibition area was also large enough, and as each day's posters had unique IDs – all setup in a single hall – one could usually go through nearly all posters in a single hour. Personally, I am not the biggest fan of spotlights – I find them too short to get any value out of them. However, with more than 900 papers, I see the need for these "short orals". Additionally, I found the session topics, i.e., the grouping of papers, to be too arbitrary at times. When looking into the proceedings, it seems that there are only three topics in computer vision right now: 3D vision, image recognition and scene understanding and machine learning; I found this to be too coarse.

[1] https://en.wikipedia.org/wiki/Great_Salt_Lake
[2] https://en.wikipedia.org/wiki/Temple_Square

Companies Plenty of companies had booths in the exhibition hall; among them many big, western IT companies including Amazon, Google, Microsoft, Intel, Facebook, Intel, NVidia etc. as well as several startups – mostly focussed around autonomous driving or training data for autonomous driving. However, I also noticed that roughly half of the companies were Asian companies (mostly from China as far as I know). Overall, except for the big interest in autonomous driving, there was quite some variety of topics represented. For most big companies, as well as some smaller ones, researchers were present in addition to recruiters and other company representatives. Many companies were also presenting their own research work, either as posters at CVPR itself or at their respective booths.