Conference Report - ICCV 2017

Written by: Moin Nabi, Isil Pekel, Anoop Katti, Hoang-Vu Nguyen, Julian Stoettinger

Venice, Italy October 22 - 29

- General
- Interesting Papers
  - Best Papers
  - Main Interesting Topics
    - Generative Adversarial Networks
    - Reinforcement Learning
    - Weakly-supervised Learning
    - Domain Adaptation
    - Efficient Deep Learning
    - Interpretable Deep Learning
    - Multimodal Deep Learning
    - Few-shot Learning
    - Privacy and Security in ML
    - Learning with Noisy Labels
  - Other Interesting Topics
- Conclusion

General

The **International Conference on Computer Vision (ICCV)** took place between October 22-29, 2017, in Venice, Italy. Also this year, ICCV, along with **CVPR**, remains at the forefront of conferences in the computer vision field.

As CVPR is moving toward being more like a general machine learning venue, ICCV still preserves a part of the traditional computer vision problems. The problems such as 3D geometry, computational photography, low-level vision etc, which get much less attention after the success of deep learning for visual recognition.

This year's conference registered full success, receiving a vast number of submissions, papers, and participants. In particular, the later has more than doubled compared to the previous ICCV edition, increasing from about 1400 to around 3200 participants. This growing crowd of participants shows the increasing level of awareness towards machine learning (ML) and particularly the importance of ML for computer vision applications.

Besides being an official sponsor of ICCV, SAP's ML researchers, data scientists and team were at the venue to participate in the talks, tutorials and workshops, here are our highlights.

General Trends and Discussions

Similar to other recent ML conferences, a large body of works was on **unsupervised learning** and **Reinforcement Learning (RL)**. In the field of unsupervised learning, the community continued to explore generative models and in particular **Generative Adversarial Networks (GANs)**. Thus, a large number of ICCV papers proposed different modifications of GANs by introducing different training techniques and showing its applications for different vision problems. Apart from this, there were several papers around the topic of RL. These lead on showing RL applications for robotics and particularly the navigation and visual planning. An interesting suggestion to point out was a new pathway combining RL with ideas coming from imitation learning.

Due to the cost of annotation collection in supervised learning and the difficulty of training with RL and GANs, the community also seems to regain interest to learn with minimal supervision. This was reflected by ICCV's large body of works on **weakly-supervised learning, semi-supervised learning, learning with noisy labels**, as well as active learning for different tasks like object detection, activity recognition and visual relation extraction among others.

Not surprisingly, **multimodal learning** drew considerable attention. Based on the technology's success gained through the integration of vision and language, several works suggested scaling up the modality integration to audio and other types of signals.

Following the recent trend of using gaming environments for different machine learning problems (e.g. AlphaGo), several researchers from different fields, such as robotics vision, visual dialogue or 3D geometry, pointed out the importance of applying **simulation environments** for computer vision. Consequently, the community is pushing towards replacing static datasets with simulation environments of both, training and evaluation. Among the reasons advertised for using simulation environments to learn visual tasks is the availability of free well-defined ground-truth labels, interpretability and focus on a particular aspect of tasks, as well as the low costs of failure.

Another important trend apparent at ICCV was the increasing thoughts put on **privacy and security in ML**. This concern holds especially true for computer vision applications, which seems to arise from the success of deep learning in the field. There were several papers proposing different adversarial attacks, evaluating the behavior of black-box models against these attacks, or suggesting some defenses for ML models in different vision tasks, such as autonomous driving.

One of the important open problems, particularly in fine-grained recognition, was learning with a small number of samples, known as **one-shot learning**. Several works in the conference addressed this problem from the perspective of meta-learning and data hallucination.

Finally, many papers presented core ideas like **curriculum learning**, self-paced learning and, in general, ranking-aware learning. In this context, the scholars suggested that for complex visual tasks the order from easy to difficult, strongly matters.

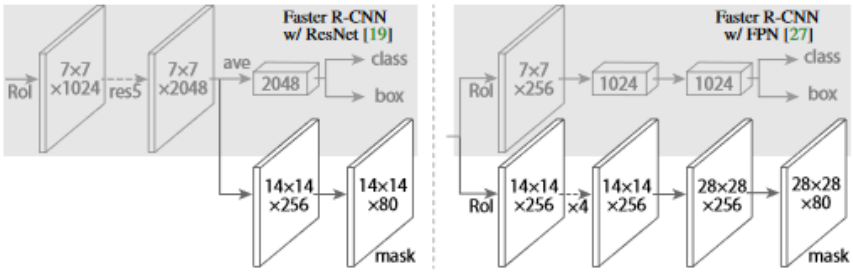
Interesting Papers

Apart from the best papers, we found some interesting papers worth mentioning and describing briefly. For illustration purposes, selected figures are taken from the papers. ## Best Papers - **Mask R-CNN**

General idea He et al. present a conceptually simple, flexible, and general framework for object instance segmentation. The approach efficiently detects objects in an image while simultaneously generating a high-quality segmentation mask for each instance. The method, called Mask R-CNN, extends Faster R-CNN by adding a branch predicting an object mask in parallel with the existing branch for bounding box recognition.

General architecture Mask R-CNN adopts the same two-stage procedure, with an identical first stage (which is RPN) as Faster R-CNN. In the second stage, in parallel to predicting the class and box offset, Mask R-CNN also outputs a binary mask for each region of interest. Separating mask prediction from classification pipeline keeps the complex multi-stage architecture simple in contrast to original R-CNN.

What is special about masks? A mask encodes an input object's spatial layout. Thus, unlike class labels or box offsets that are inevitably collapsed into short output vectors by fully-connected (fc) layers, extracting the spatial structure of masks can be addressed naturally by the pixel-to-pixel correspondence provided by convolutions. In the RoIAlign layer, they do not use any quantization, like in the RoIPool layer, and thus the extracted features can be aligned with the input layer.



Why is it interesting: Mask R-CNN is simple to train. It improves the accuracy in the instance segmentation with a very

simple extension on Faster R-CNN and adds only a small overhead to Faster R-CNN, running at 5 fps. Moreover, Mask R-CNN is easy to generalize to other tasks like estimating human poses in the same framework and shows top results in all three tracks of the challenges; instance segmentation, bounding-box object detection, and person keypoint detection

- **Focal Loss for Dense Object Detection**

General idea: The highest accuracy object detectors to date are based on a two-stage approach popularized by R-CNN, where a classifier is applied to a sparse set of candidate object locations. In contrast, one-stage detectors that are applied over a regular, dense sampling of possible object locations have the potential to be faster and simpler, but have trailed the accuracy of two-stage detectors thus far. In this paper, Lin et al. investigate why this is the case and they discover that the extreme foreground-background class imbalance encountered during training of dense detectors is the central cause.

Methodology: They propose to address this class imbalance by focal loss. Focal Loss focuses training on a sparse set of hard examples and prevents the vast number of easy negatives from overwhelming the detector during training. Mathematically it reshapes the standard cross entropy loss with a new modulating factor such that it down-weights the loss assigned to well-classified examples. To evaluate the effectiveness of the loss, they introduce RetinaNet, which is a simple one-stage object detector. The experiments with RetinaNet shows that when trained with Focal Loss, RetinaNet is able to match the speed of previous one-stage detectors while surpassing the accuracy of all existing state-of-the-art two-stage detectors.

Why is it interesting: RetinaNet shares many similarities with the existing dense detectors. However, the power of RetinaNet does not come from the innovations in the network design, but from the novel and robust loss function; focal loss. With the small change in the cross entropy function, the authors of the paper have achieved top accuracy and high performance.

Main Interesting Topics

Apart from the best papers, we found that a number of papers are worth taking a deeper look. Here is a list of our favorite examples from different topics:

Generative Adversarial Networks

- DualGAN: Unsupervised Dual Learning for Image-to-Image Translation

This paper deals with the problem of translating images from one domain to the other, for example converting day-time images to night-time images, paintings to images etc. The previous approaches based on conditional GANs required paired images from both the source and the target domain. However, such labeled pairs are expensive to collect. In this paper (and a related work dubbed CycleGAN), the authors propose a way to perform domain translation with two independent sets of images from both domains. In particular, the model has two complementary GANs, the primal and the dual GANs. The primal GAN learns to translate images from domain U to domain V, while the dual GAN learns to invert the task. This creates a closed loop and allows us to use the reconstruction loss to train the model. Figure 1 in the paper shows the architecture in more detail:

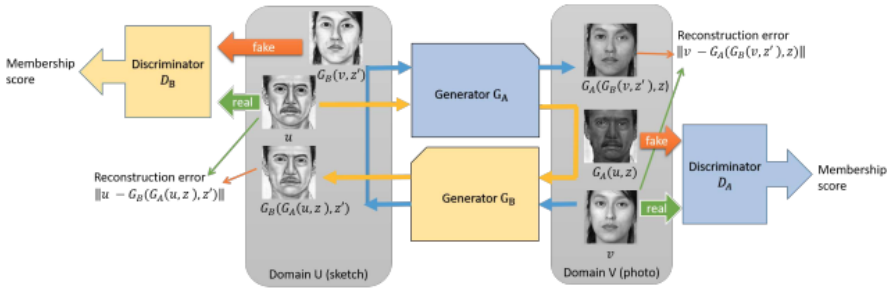


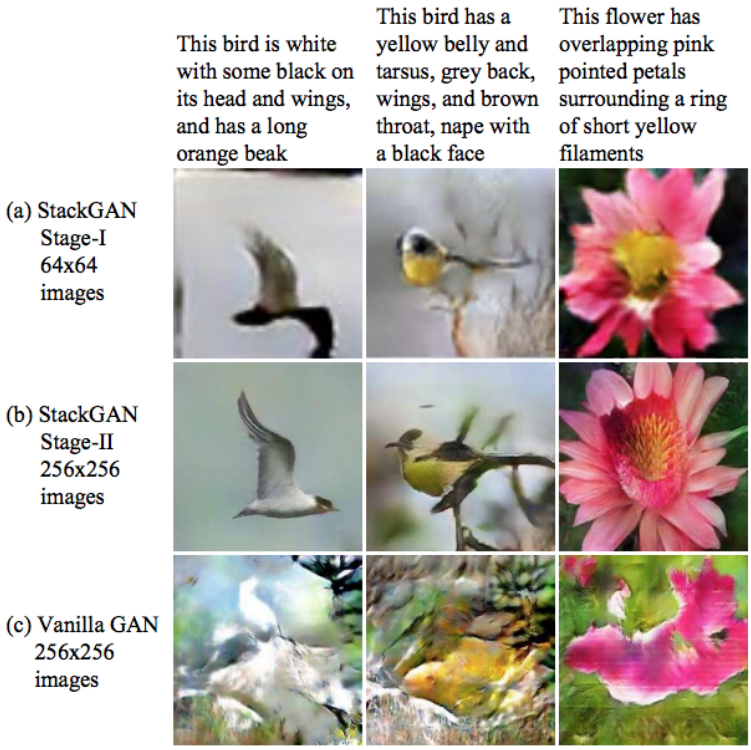
Figure 1: Network architecture and data flow chart of DualGAN for image-to-image translation.

Why is it interesting: This work is interesting for two reasons: (1) It attempts to solve Domain Translation without needing any explicit labeled data. It is a step towards unsupervised learning. (2) By successfully training such an indirect formulation, it demonstrates the power of the latest machine learning models. This motivates to make the models work harder with less and less explicit labels, while still achieving successful results.

- StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks

Generating high resolution images with GANs is a challenging problem. Aiming to find a solution to this problem, the authors propose a two stage process that progressively increases the resolution of the generated image. In Stage-I, the GAN sketches the primitive shape and colors of the object based on the given text description, yielding an initial sketch in low-resolution. In Stage-II, the GAN takes the Stage-I results and text descriptions as inputs, and generates high-resolution images with photo-realistic details. To improve the diversity of the generated images, the authors introduce some randomness in the text feature conditioning.

Rows 1 and 2 of Figure 1 of the paper visualize the outputs of Stage-I and Stage-II GANs. As it can be seen, the output of Stage-I captures the color and the shape but is blurry and the output of Stage-II is much sharper and corrects the mistakes of Stage-I.



> **Why is it interesting:** Generating images from descriptions is a very interesting problem with many potential applications. However, it is critical for any application that the generated images appear realistic. This work is a step in this direction.

- A Generative Model of People in Clothing

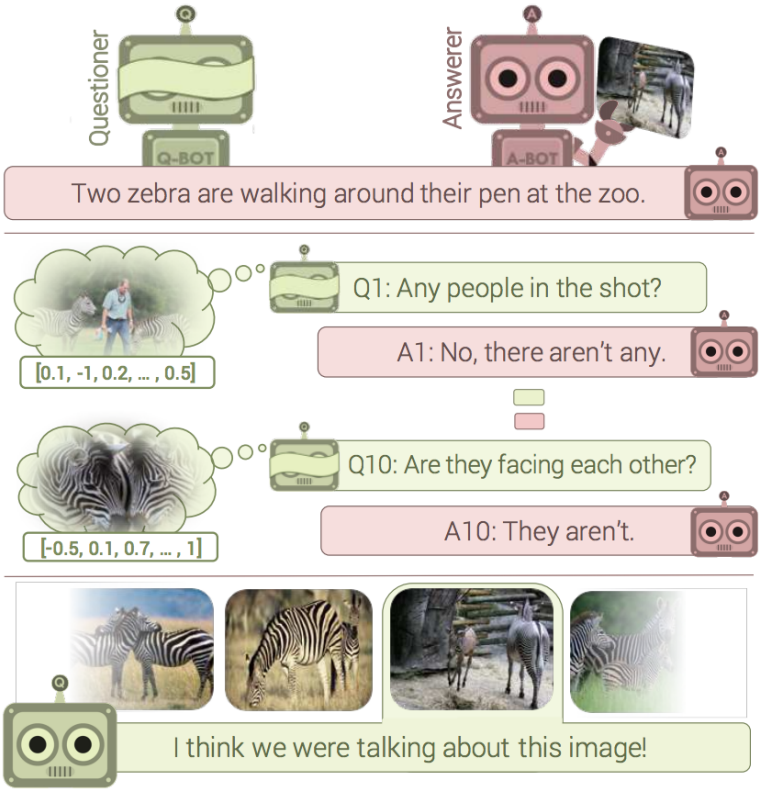
This work develops a method to generate samples of full-body people in diverse clothing. Traditionally, this task required complex graphics rendering pipeline and 3D scans of dressed people. In this work, the authors build a generative model based solely on images. The main challenges are large variations in pose, shape and appearance. The authors cope with this by splitting the generation process in two steps: generating a semantic segmentation of the body and clothing and generating realistic images conditioned on the outputs of the first step. The generated images can be conditioned on pose, shape and color and this makes the model highly versatile.

> **Why is it interesting:** Generating realistic images of fashion products is key to high sales in e-commerce. Further, being able to condition sample generation on pose, shape and color makes this model highly flexible and attractive.

Reinforcement Learning

- Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning

The authors introduce a goal-driven training for visual question answering and dialog agents. Specifically, they pose a cooperative 'image guessing' game between two agents – Question-BOT and Answer-BOT– who communicate in a natural language dialog so that the Q-BOT can select an unseen image from a lineup of images. The proposed deep reinforcement learning (RL) approach is used to learn the policies of these agents end-to-end, from pixels via multi-agent multi-round dialog to game reward.

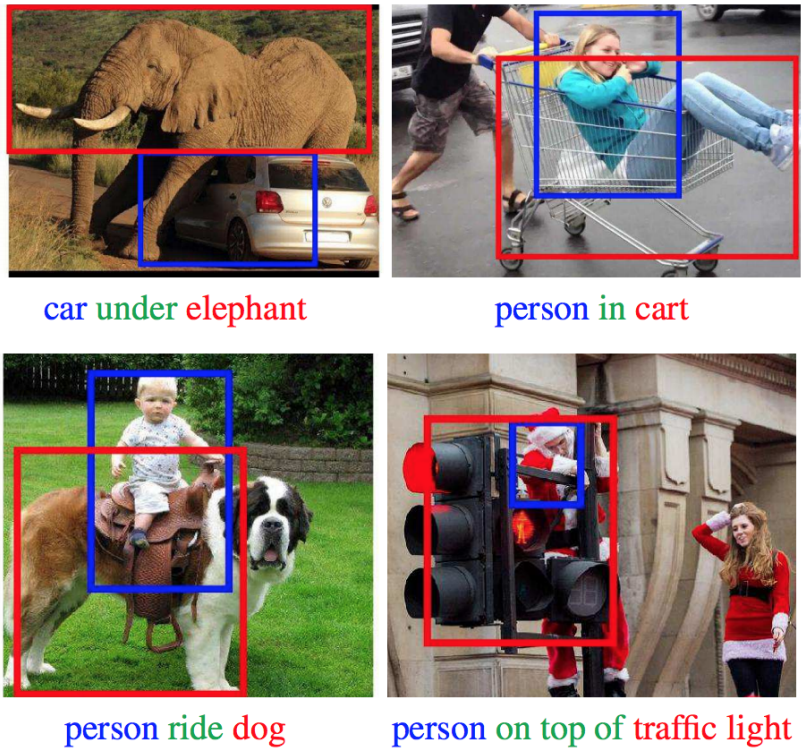


Why is it interesting: All previous approaches for visual question answering and dialog agents, have been based on supervised learning, and this paper is the first goal-driven training paradigm for these tasks. The authors succeed to demonstrate the emergence of grounded language and communication among 'visual' dialog agents with no human supervision. More interestingly, the RL Question-BOT learns to ask questions that the Answer-BOT is good at, ultimately resulting in more informative dialog and a better team.

Weakly-supervised Learning

- Weakly-supervised learning of visual relations

This paper introduces a novel approach for modeling visual relations between pairs of objects. The specified relation is a triplet of the form (subject, predicate, object), where the predicate is typically a preposition (eg. 'under', 'in front of') or a verb ('hold', 'ride') that links a pair of objects (subject, object). Learning such relations is challenging as the objects have different spatial configurations and appearances depending on the relation in which they occur. The difficulty to get annotations for all possible triplets, especially at box-level, makes it complex for both learning and evaluation. In this paper, the authors propose a weakly-supervised discriminative clustering model to learn relations from image-level labels only.



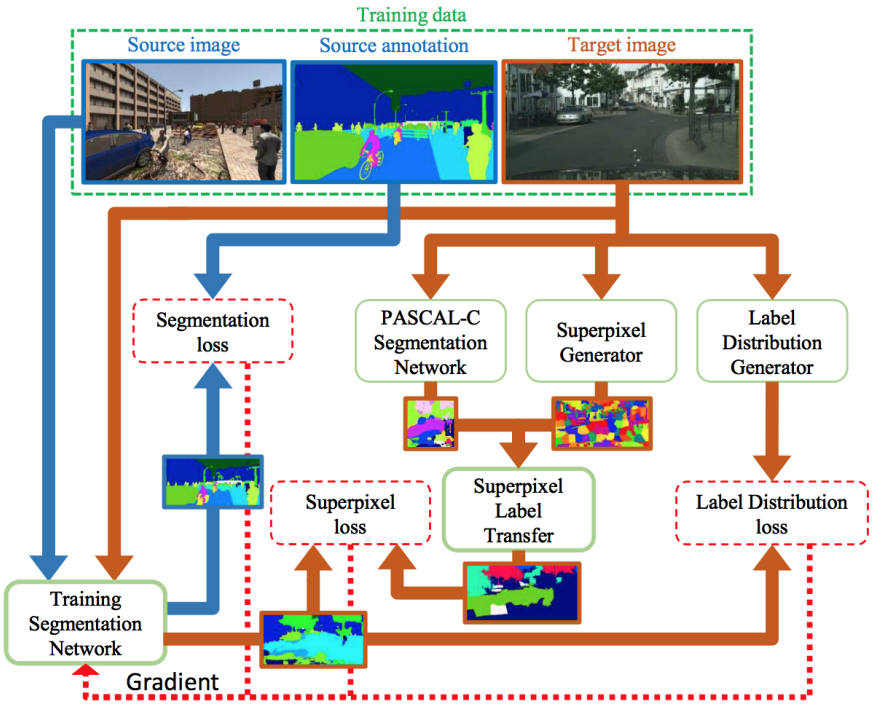
Why is it interesting: Learning a large vocabulary of visual relations directly from large-scale Internet collections is the goal of this paper. Specifically, the authors proposed a weakly-supervised

model for learning object relations and have demonstrated that, given pre-trained object detectors, object relations can be learnt from weak image level annotations without a significant loss of recognition performance. A new dataset for visual relation detection is also presented that enables to evaluate retrieval without missing annotations and to assess generalization to unseen triplets. This opens up the possibility of studying on "causality" and "correlation" in computer vision as well.

Domain Adaptation

- Curriculum Domain Adaptation for Semantic Segmentation of Urban Scenes

In this paper, a curriculum-style learning approach is proposed which minimizes the domain gap in semantic segmentation. The curriculum domain adaptation solves easy tasks first in order to infer some necessary properties about the target domain. They then train the segmentation network in such a way that the network predictions in the target domain follow those inferred properties.



Why is it interesting: Recent advances in computer graphics make it possible to train CNN models on photo-realistic synthetic data with computer-generated annotations. Despite this, the domain mismatch between the real images and the synthetic data significantly decreases the models' performance. Following the assumption that some tasks are 'easy' and therefore suffer less due to the domain discrepancy, in this paper a curriculum-based domain adaptation method is proposed. After high success gained through the ranking-aware model in different machine learning problems, this paper is the first showing that curriculum learning can also be applied in domain adaptation problems.

Efficient Deep Learning

- Learning Efficient Convolutional Networks through Network Slimming

The authors propose a network slimming technique to learn more effectively from compact CNN models. Their approach is called network slimming, meaning that it takes wide and large networks as input models, but during training automatically identifies and prunes insignificant channels. This technique yields thin and compact models with comparable accuracy, which can be achieved by directly enforcing channel-level sparsity in the network in a simple but effective way.

More specifically, the sparsity-induced regularization is applied as the scaling factors in batch normalization layers, which thus allows to detect unimportant channels during training and later to exclude them. The authors empirically demonstrate the effectiveness of their approach with several state-of-the-art CNN models, including VGGNet, ResNet and DenseNet, on various image classification datasets.

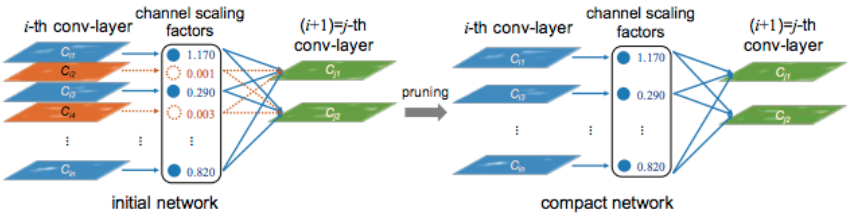


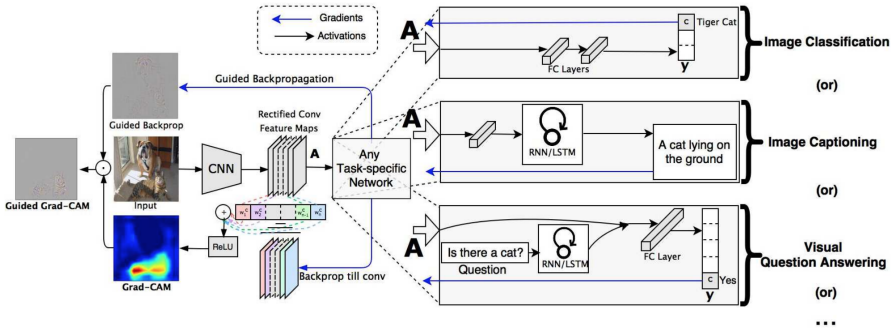
Figure 1: We associate a scaling factor (reused from a batch normalization layer) with each channel in convolutional layers. Sparsity regularization is imposed on these scaling factors during training to automatically identify unimportant channels. The channels with small scaling factor values (in orange color) will be pruned (left side). After pruning, we obtain compact models (right side), which are then fine-tuned to achieve comparable (or even higher) accuracy as normally trained full network.

Why is it interesting: The deployment of deep convolutional neural networks (CNNs) in many real world applications is largely hindered by their high computational cost. This paper proposes a scheme for CNNs to reduce the model size, decrease the run-time memory and lower the number of computing operations simultaneously. The appeal is that this method can be directly applied to modern CNN architectures, with minimum overhead to the training process, requiring no special software/hardware accelerators and more interestingly without sacrificing accuracy.

Interpretable Deep Learning

- Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization

The authors propose a technique for producing 'visual explanations' for decisions from a large class of Convolutional Neural Network (CNN)-based models, making them more transparent. The approach, called Gradient-weighted Class Activation Mapping (Grad-CAM), uses the gradients of any target concept (e.g. logits for 'dog' or even a caption), flowing into the final convolutional layer to produce a coarse localization map highlighting the important regions in the image for predicting the concept. Unlike previous approaches, GradCAM is applicable to a wide variety of CNN model-families: (1) CNNs with fully-connected layers (e.g. VGG), (2) CNNs used for structured outputs (e.g. captioning), (3) CNNs used in tasks with multi-modal inputs (e.g. visual question answering) or reinforcement learning, without architectural changes or re-training. They also combine Grad-CAM with existing fine-grained visualizations to create a high-resolution class-discriminative visualization, Guided Grad-CAM, and apply it to image classification, image captioning, and visual question answering (VQA) models, including ResNet-based architectures.



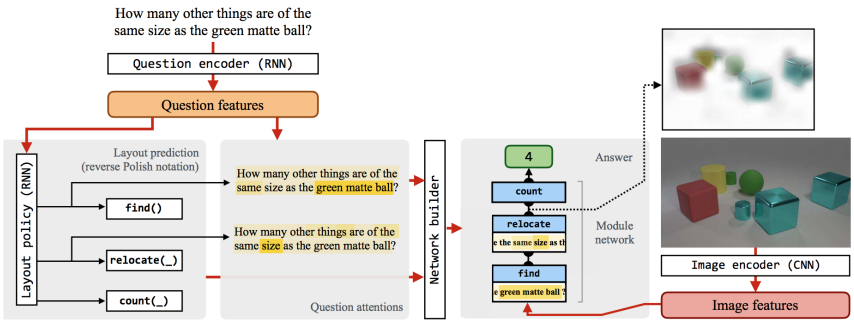
Why is it interesting: A true AI system should not only be intelligent, but also be able to reason about its beliefs and actions for humans to trust it. In this regard, visualization is known to be a very helpful tool toward interpretability and faithfulness aspects of the ML model. In this work, a novel class-discriminative localization technique (Grad-CAM) is presented. It provides the means to make any CNN-based models more transparent by producing visual explanations. The authors also combine the Grad-CAM localizations with existing high-resolution visualizations to obtain high-resolution class-discriminative Guided Grad-CAM visualizations. In the paper, they show the broad applicability of Grad-CAM to various off-the-shelf available architectures for tasks including image classification, image captioning and VQA providing faithful visual explanations for possible model decisions. This method can be employed as a diagnostic tool to discriminate more accurately between classes, better reveal the trustworthiness of a classifier, and help identify biases in datasets.

Multimodal Deep Learning

- Learning to Reason: End-to-End Module Networks for Visual Question Answering

In this paper, an End-to-End Module Network for visual question answering is presented. The model uses a set of neural modules to break down complex reasoning problems posed in textual questions into a few sub-tasks connected together, and learns to predict a suitable layout expression for each question using a layout policy implemented with a sequence-to-sequence RNN. During

training, the model can be first trained with behavioral cloning from an expert layout policy, and further optimized end-to-end using reinforcement learning. The results demonstrate that such model is capable of handling complicated reasoning problems. Also, the end-to-end optimization of the neural modules and layout policy can lead to significant further improvement over behavioral cloning from expert layouts. The authors obtained state-of-the-art results on the difficult CLEVR dataset by a large margin.



Why is it interesting: Natural language questions are inherently compositional, and many are most easily answered by reasoning about their decomposition into modular sub-problems. The interestingness is that the authors introduce a method for learning a layout policy that dynamically predicts a network structure for each instance, without the need of any external linguistic resources at inference time. They also propose a module parameterization that uses a soft attention over question words rather than hard-coded word assignments. Their results suggest that such model is capable of directly predicting expert-provided network layouts with near-perfect accuracy, and even of improving on expert designed networks after a period of exploration.

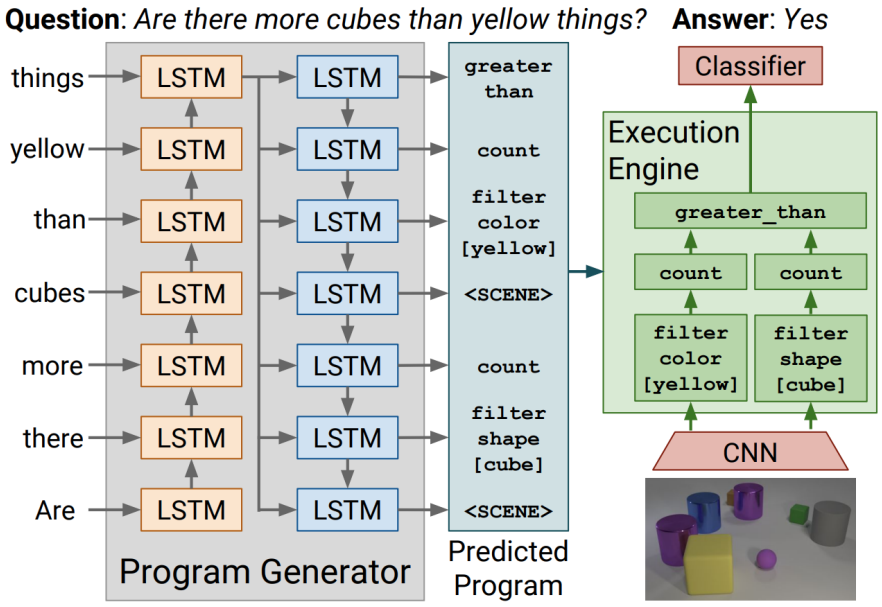
- Inferring and Executing Programs for Visual Reasoning

Modern approaches to visual recognition learn mapping directly from inputs to outputs; they do not explicitly formulate and execute compositional plans. Direct input-output mapping works well for classifying images and detecting objects for a small, fixed set of categories. However, it fails to outperform strong baselines on tasks that require the model to understand an exponentially large space of objects, attributes, actions, and interactions, such as visual question answering (VQA).

To address this problem, the authors investigate on a new model for visual

question answering that consists of two parts: a program generator and an execution engine. The program generator reads the question and produces a plan or program for answering the question by composing functions from a function dictionary. The execution engine implements each function using a small neural module, and executes the resulting module network on the image to produce an answer.

They evaluate the model on the recently released CLEVR dataset and find that with only a small amount of reasoning supervision ground truth programs, their model achieves very high precision in a very generalizable manner.



Why is it interesting: This paper fits into a long line of work on incorporating symbolic representations into (neural) machine learning models. The scholars show that explicit program representations can make it easier to compose programs to answer novel questions about images. The generic program representation, learnable program generator and universal design for modules make the model much more flexible than neural module networks and thus more easily extensible to new problems and domains.

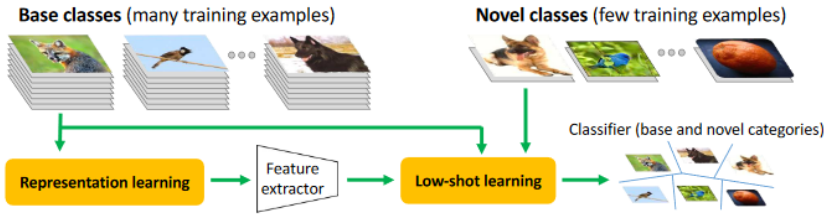
Few-shot Learning

- Low-shot Visual Recognition by Shrinking and Hallucinating Features

Low-shot visual learning, the ability to recognize novel object categories from very few examples, is a hallmark of human visual intelligence. Existing machine learning approaches fail to generalize in the same way. To make progress in this foundational problem, Hariharan et al. present a low-shot learning benchmark on complex images that mimics challenges faced by recognition systems in the wild.

The benchmark is implemented in two phases. In the representation learning phase, the learner tunes its feature representation on a set of base classes that have many training instances. In the low-shot learning phase, the learner is exposed to a set of novel classes with only a few examples per class and must learn a classifier over the joint label space of base and novel classes. To improve the learning ability, the authors formulate a different loss function, called squared gradient magnitude, that penalizes the difference between classifiers learnt on large and small datasets.This then allows to draw connections between this loss and regularization of feature activations.

The authors investigate further how to improve the learner's performance on the benchmark. They build on the intuition that certain modes of intra-class variation generalize across categories (e.g., pose transformations). A way of "hallucinating" additional examples for novel classes is achieved by transferring modes of variation from the base classes.

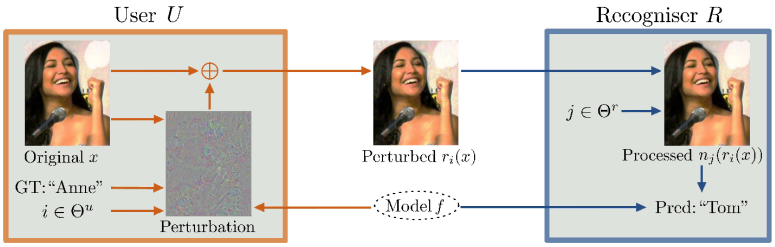


Why is it interesting: The authors approach the common class imbalance problem in many different ways. They not only create a low-shot recognition benchmark of realistic complexity, but also address the performance of this benchmak in two ways: (1) choosing the proper loss function to penalize the differeence between many and few example classifiers and (2) they propose a novel way of transferring modes of variation from base classes to data-starved ones.

Privacy and Security in ML

- Adversarial Image Perturbation for Privacy Protection: A Game Theory Perspective

Recent studies on adversarial image perturbations (AIP) suggest that it is possible to confuse recognition systems effectively without unpleasant artifacts. However, in the presence of counter measures against AIPs, it is unclear how effective AIP would be, in particular when the choice of counter measure is unknown. The authors introduce a general game theoretical framework for the user-recogniser dynamics, and present a case study that involves current state of the art AIP and person recognition techniques. Moreover, they derive the optimal strategy for the user assuring an upper bound on the recognition rate independent of the recogniser's counter measure.



Why is it interesting: Users in social media are interested in sharing personal photos, and at the same time make the automatic identification in their photos difficult. Classic obfuscation methods (e.g. blurring) are not only unpleasant but also not as effective. The authors have constructed a game theoretical framework to study a system with two players, user U and recogniser R, with antagonistic goals (dis-/enable recognition). They have examined existing and new adversarial image perturbation (AIP) techniques for U. In this work, the authors have constructed a game theoretical framework to study a system with two players, user U and recogniser R, with antagonistic goals (dis-/enable recognition). They also examine existing and new adversarial image perturbation (AIP) techniques for U. This paper serves as a first step towards the direction of analysing the user-recogniser dynamics in social media.

Learning with Noisy Labels

- Revisiting Unreasonable Effectiveness of Data in Deep Learning Era

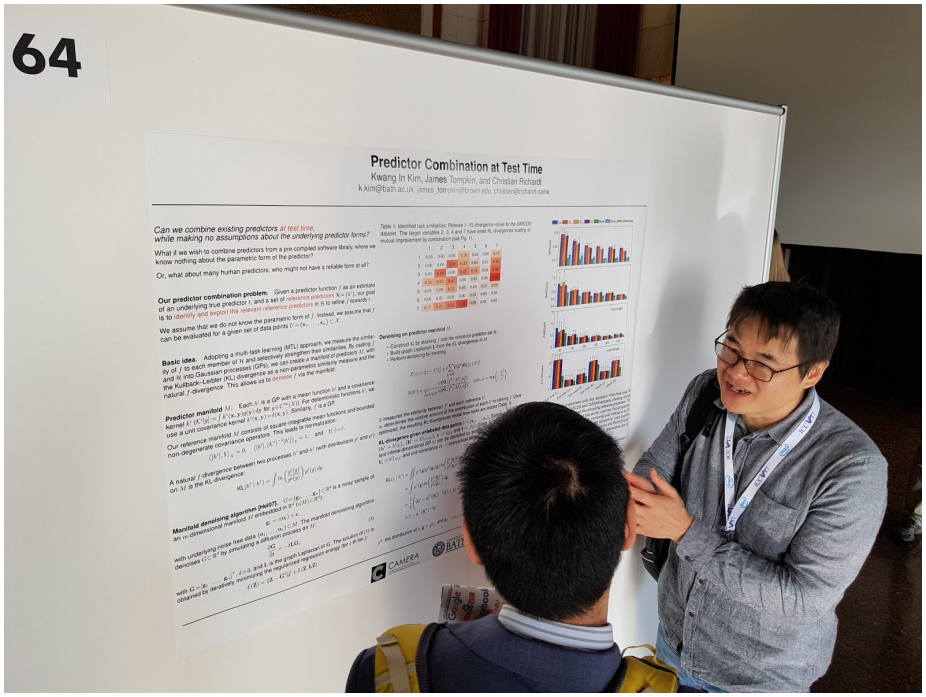
This work studies the effect of the dataset size on performance. In particular, it has two take-aways: (1) performance increases logarithmically with the size of the training data and (2) performance can be improved by simply training a better base model; in other words, learning a good representation is crucial to good performance. The authors also introduce a ultra large-scale dataset with 300M images and 375M noisy labels used in the study. By training a better base representation, they achieve state of the art results on many vision tasks, including classification, detection, segmentation and human pose estimation.

> **Why is it interesting:** The authors propose a very systematic study of the effect of dataset size on performance. This gives a feel of what improvement to expect by just increasing the size of the dataset, in general. It also shows the importance of learning a good base representation.

Other Interesting Topics

Predictor Combination at Test Time

The authors work on the problem of "predictor combination" and empirically test potentially unknown predictors on sampled data points to measure distances between different predictors (also human predictors). They obtain a redefined predictor, applied on multi-task and transfer learning algorithms.



Why is it interesting: The paper is a great application of denoising algorithm by Hein et Meier, NIPS 2017. It further allows the flexible use and combination of multi-source knowledge and pre-trained predictors.

Conclusion

ICCV remains a top conference for computer vision. Keeping very high quality standards in terms of papers, talks and workshops, the conference has attracted a growing international audience in Venice this year.

The SAP ML research team is working on a set of the introduced topics with our visiting students, such as the topic of ML under privacy, and engages in research collaborations with top-tier universities to drive progress in areas like generative few-shot learning.