Research Statement

Judy Hoffman

Demand for automated understanding of visual data is higher than ever before. Autonomous vehicles need to recognize pedestrians, traffic signs, and other cars. Companies and consumers need to automatically organize the ever increasing volume of visual media. Next generation personal robotics need to recognize user specific objects in unconstrained consumer environments. Despite substantial progress in recent years, our algorithms still fail to generalize to the variety of visual environments encountered by real-world applications.

Typical state-of-the-art learning approaches, especially deep learning, rely on supervision in the form of thousands to millions of human annotated data-label pairs to learn a competent visual recognition model. Consider understanding a road scene for autonomous driving. Current solutions need examples that span all weather conditions, times of day, cities of operation, and camera positions, to produce a model robust to these variations. This quickly becomes expensive for a particular application and infeasible when considering the breadth of applications for which visual recognition is needed. The fundamental limitation of this approach is that it does not effectively reuse information and thus needs new human labeled examples for each new visual task and for each variation within tasks. In order to scale our models to tackle the vast space of possible visual tasks under real-world diverse settings, we need to develop learning solutions which are not bottlenecked by direct human supervision.

My work effectively reuses and shares information across environments and tasks, enabling learning systems to tackle real-world variation and scale while minimizing human supervision. I develop learning algorithms which facilitate this transfer of information through unsupervised or semi-supervised model adaptation and generalization. While I have thus far focused on understanding visual data, many of my algorithms are applicable to other learning paradigms, such as policy transfer for robotics, and sentiment analysis in natural language processing. My previous work has addressed transfer of information under three main variation assumptions: 1) transfer across different imaging environments, such as from simulated to real driving scenes, 2) transfer across different visual processing tasks, such as between holistic scene understanding and per pixel object categorization, and 3) transfer across different imaging modalities, such as RGB and depth.

Learning to Adapt Across Domains

Consider potential imaging variations (*domains*) within driving scenes. This may include changing from day to night, varying resolutions, or transitioning from rural to urban scenes. A static collection of visual data only addresses a subset of these variations and suffers from reduced performance when evaluated under different variation conditions (*domain shift*). Given enough additional supervised data from the new environment, the desired performance could likely be recovered. However, this solution is often prohibitively expensive and intellectually dissatisfying as only the imaging conditions are changing, not the concepts we need to recognize. Thus, imaging conditions are not fundamental to the definition of concepts, and can rather be viewed as distractors to which our models should be invariant.

Optimizing for Domain Invariance. Most state-of-the-art visual recognition systems use end-to-end convolutional networks (ConvNets) which take as input pixel values and produce as output semantic predictions, such as the presence of a car in an image. My work was the first to point out that this ability to directly learn an image representation offers a unique opportunity for solving the domain shift problem through optimizing for domain invariance [1]. Theoretical analysis of domain shift defines a domain as a distribution over inputs and outputs, and the shift between them as the distance between those distributions.

The insight of my approach was to adapt by directly minimizing the distance between the domain distributions viewed under the current representation. Thus, we find a high level representation for our data under which the data sources are rendered indistinguishable. Critically, this objective does not require any semantic labels in the new imaging condition and thus can be applied without human supervision. I proposed multiple algorithms based on this principle of adaptation through distribution distance minimization using various distribution distance metrics including, max mean discrepancy [1], domain confusion [2], and

domain adversarial learning [3]. In addition, I recently introduced a hierarchical learning and alignment objective which is more effective when optimizing for invariance across large scale, fine-grained tasks, such as distinguishing between all vehicle types in Google Street View [4].

Transforming One Domain into Another. When data itself is scarse, modifying the full representation based on a few examples may be ill-conditioned. Thus, some of my earlier work focused on aligning domains with fixed representations by learning a transformation which transformed one domain into another by jointly optimizing the mapping and a max-margin classification objective [5]. Along with the novel objective, I proposed an optimization solution and proved that the optimization converged to a local optimum. Compared to prior work, this algorithm was much faster (constraints were linear in terms of training points while prior work was quadratic), and produced stronger classifiers. Importantly, the algorithm produced a category invariant transformation and so could be applied to test data from any category, even those for which no labeled data is available.

Recently, I have been exploring the use of content preserving image space transformations to map a labeled training image into the style of the the test domain, effectively generating the diverse labeled examples needed for supervised learning. This approach offers promising results on challenging problems such as transferring a model learned using simulated driving imagery to adapt for use in the real-world. Interestingly, I have also shown that image space and representation space adaptive approaches are complementary and may be combined for further overcoming the domain shift problem [6].

Joint Transfer and Adaptation. There has been a large push in the research community to address generalizing and adapting deep models across different domains and separately to learn tasks in a data efficient way through few-shot learning. In fact, often these situations appear simultaneously, e.g., driving imagery collected in San Francisco will look different than that in rural Idaho, but there may also be new categories appearing in Idaho, such as deer and horses. Therefore, I recently developed an algorithm to jointly adapt a representation using a new multilayer unsupervised domain adversarial formulation while introducing a cross-domain and within domain class similarity objective which relaxes the assumption of overlapping classes used in most adaptation frameworks [7]. This approach was used to dramatically improve the performance of a video action recognizer trained with sparse supervision and a representation transferred from a standard image classification data source (ImageNet).

Transferring Information Across Tasks and Modalities

To generate truly large scale models, those that approach or exceed human recognition capabilities, we need to develop models that can effectively learn from all related forms of supervision. For situations which lack the annotated data needed for supervised learning, there is often plentiful weakly labeled visual data (e.g. unverified user supplied tags) or strongly labeled related data (e.g. RGB images when learning a depth modality). My work effectively used supervision from auxiliary tasks and modalities to produce a large scale and multimodal visual recognition and localization model.

Large scale detection through adaptation. For many applications which need visual recognition, such as robotic manipulation, recognizing the presence of an object is not sufficient, one must also locate the object. Unfortunately, to supervise this task humans must draw a bounding box around and indicate a label for each object of interest. This process is tedious and an unnatural way for humans to interact with visual data. In contrast, images tagged with present objects are relatively easy to acquire or even freely available. While some algorithms have been introduced to use the large amounts of this weaker form of supervision, such systems still dramatically underperform their strongly supervised counterparts.

I introduced a hybrid approach which jointly learns from fully annotated (bounding box) data for a small subset of object categories and weakly annotated (full image tags) for the remaining categories [8]. Following my background in adaptation, I considered classification (full image recognition) and detection (localized recognition) tasks as two domains and cast the task of learning to transfer between them as

a domain adaptation problem. I further specialized my transfer to be category specific using a multiple instance learning to mine additional localized supervision from the image annotated data [9]. At the time of publishing, the largest set of object-centric classification labeled data had examples for around ten thousand categories. Using my approach I was able to transform these classifiers and release a 10K category detector, a substantially larger scale than the 200 category model from supervised systems. This offered a significant step towards having a generic large-scale visual recognition model and our released detector was adopted by colleagues for general purpose robotics applications.

Cross-Modal Transfer for RGB-d Detection Color (RGB) images are the predominant source of visual imagery freely available. However, for many tasks other sensory information is crucial to solving the desired recognition problem. In robotic manipulation depth sensors are often used to help determine the desired robot 3D grip position. LIDAR sensors are added to autonomous vehicles to aid in detecting and understanding distance to pedestrians. While different from RGB, these modalities also share many common components. I proposed a sequence of algorithms for transferring information across modalities to improve learning with limited depth training data [10, 11] and to improve performance in modality limited test environments [12]. The core principal in each of these approaches is to use paired sensory inputs, specifically RGB-d images, and supervise the learning of a new model under the assumption that the information extracted from each modality should at some intermediate representation match.

Beyond a Dataset as a Domain

Thus far I have discussed adapting across different static collections of data. Though common, this paradigm is overly limiting for many applications. Consider two examples below:

Discovering Latent Domains. Consider a robot recognizing a child's toy elephant. The data available to learn an elephant model may contain line drawings, cartoons, and images taken in a zoo or in the wild. Adapting directly from that diverse distribution may prove unnecessarily difficult when the toy elephant is most similar to the cartoon images. Thus, I introduced the concept of multiple latent domains within a single dataset and proposed a hierarchical constrained clustering algorithm for discovering these sub-domains [13]. My work inspired new research in multiple source visual domain adaptation and domain discovery.

Continuous Adaptation. Not only is each static training dataset often comprised of multiple domains, but for many real world applications, the test data may evolve over time. For example, consider an autonomous vehicle navigating around a city. The images will change throughout the day as shadows are cast across the scene, glare is introduced by sun positioning, weather changes cause the ground and lens to be covered in snow/rain, and at nighttime the sensor modality may change. A single transformation cannot adequately represent the difference between the training data and all variations of the test data.

I proposed an unsupervised algorithm to continuously adapt to an evolving visual domain [14], by representing an instantaneous domain distribution as a point on the Grassmannian manifold and predicting the current domain in time after receiving a new unlabeled data point. With a predicted domain distribution the algorithm then produces a transformation that best adapts the current domain back to the source labeled domain for classification.

Future Work

My goal is to produce lifelong learning systems which reliably and efficiently adapt and expand their knowledge in response to an ever changing world. I anticipate three main future directions towards this goal: 1) advancing self-supervised techniques for autonomous learning, 2) developing uncertainty awareness for selectively seeking supervision, and 3) formalizing our adaptation techniques and providing theoretical guarantees to avoid catastrophic failure.

Autonomous Learning through Self-supervision and Interaction. As the world continues to evolve, it is critical for sustainability that a learning system be capable of advancing without constant human

supervision. In the short term, I aim to advance progress in unsupervised and self-supervised learning techniques, making use both of intrinsic signals within data and across prior knowledge. The techniques I have developed for adaptation offer a strong starting point for developing general purpose unsupervised learning approaches within passively collected data. As a longer term goal, I aim to develop learning solutions that have control over the data acquisition process as well as the data processing pipeline. We know that children use play as a form of physical self-supervision, so to as our learning systems progress we need to move beyond learning from passive data into interactive learning with embodied systems. Towards this goal, I am excited to collaborate with roboticists, psychologists, and neuroscientists to both integrate algorithms within embodied systems and guide algorithmic development using our understanding of biological learning.

Uncertainty Awareness and Active learning. In certain situations, the most efficient and safe decision is to ask for help. I plan to develop learning approaches with uncertainty awareness to equip our systems with the ability to seek guidance. Most modern learning systems possess a recognition score which may be loosely interpreted as a confidence measure. However, these scores are notoriously poorly scaled, leading to overly confident mistakes and an inability to combine predictions from multiple models. As we continue learning with limited human supervision and across multiple sources of data, it becomes even more pressing to develop solutions both for self-calibration and for predicting uncertainty. Enabling systems to seek human advice when confronted with truly novel and undetermined objects or environments, leads to faster and more accurate adaptation and expansion while still requiring minimal human supervision. As systems begin to rely on human in the loop learning, new human-computer interaction problems emerge, such as how to engage users and how to ask questions so as to receive proper supervision.

Avoiding Catastrophic Failure. When systems learn without direct supervision, there is always a risk of catastrophic failure. To prevent this, I seek to analyze our self-supervised and adaptive learning algorithms using a formal framework which may provide clear and concise guarantees for our solutions. I previously co-wrote an awarded NSF grant to merge theory and practice in the context of domain adaptation. In connection with this effort I initiated a collaboration with Professor Mehryar Mohri from NYU Courant Institute, and we have been developing algorithms applicable to the ConvNet learning systems used in computer vision which are accompanied by theoretical worst case guarantees. Guiding algorithm development with theoretical frameworks will be especially useful for uncertainty assessment and making active learning decisions.

In the long term, I aim to make progress towards intelligent systems which are capable of performing a variety of tasks. Though I have focused mainly on understanding visual data thus far, future learning and decision making will require seamless integration across multiple modalities. Active learning agents rely on communication with people and thus need to have a clear grasp on natural language. Interactive agents need to be able to safely navigate and explore new environments. Model expansion over time will require new learning tools which will benefit from both theoretical machine learning advances as well as understanding of learning in biological systems. My work on adaptive learning has already positioned me to initiate collaborations across robotics and machine learning, I look forward to continuing these collaborations and expanding to natural language processing and cognitive science.

References

[1] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *CoRR*, abs/1412.3474, 2014.

[2] Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. Simultaneous deep transfer across domains and tasks. In *International Conference on Computer Vision (ICCV)*, 2015.

[3] Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. Adversarial discriminative domain adaptation. In *Computer Vision and Pattern Recognition (CVPR)*, 2017.

[4] Timnit Gebru, Judy Hoffman, and Li Fei-Fei. Fine-grained recognition in the wild: A multi-task domain adaptation approach. In *International Conference on Computer Vision (ICCV)*, 2017.

[5] Judy Hoffman, Erik Rodner, Jeff Donahue, Kate Saenko, and Trevor Darrell. Efficient learning of domain-invariant image representations. In *International Conference on Learning Representations (ICLR)*, 2013.

[6] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle consistent adversarial domain adaptation. *CoRR*, abs/1711.03213, 2017.

[7] Zelun Luo, Yuliang Zou, Judy Hoffman, and Li Fei-Fei. Label efficient learning of transferable representations across domains and tasks. In *Neural Information Processing Systems (NIPS)*, 2017.

[8] Judy Hoffman, Sergio Guadarrama, Eric Tzeng, Jeff Donahue, Ross Girshick, Trevor Darrell, and Kate Saenko. LSDA: Large scale detection through adaptation. In *Neural Information Processing Systems (NIPS)*, 2014.

[9] Judy Hoffman, Deepak Pathak, Trevor Darrell, and Kate Saenko. Detector discovery in the wild: Joint multiple instance and representation learning. In *Computer Vision and Pattern Recognition (CVPR)*, 2015.

[10] Judy Hoffman, Saurabh Gupta, Jian Leong, Sergio Guadarrama, and Trevor Darrell. Cross-modal adaptation for rgb-d detection. In *International Conference in Robotics and Automation (ICRA)*, 2016.

[11] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In *Computer Vision and Pattern Recognition (CVPR)*, 2016.

[12] Judy Hoffman, Saurabh Gupta, and Trevor Darrell. Learning with side information through modality hallucination. In *Computer Vision and Pattern Recognition (CVPR)*, 2016.

[13] Judy Hoffman, Brian Kulis, Trevor Darrell, and Kate Saenko. Discovering latent domains for multisource domain adaptation. In *European Conference on Computer Vision (ECCV)*, 2012.

[14] Judy Hoffman, Trevor Darrell, and Kate Saenko. Continuous manifold based adaptation for evolving visual domains. In *Computer Vision and Pattern Recognition (CVPR)*, 2014.