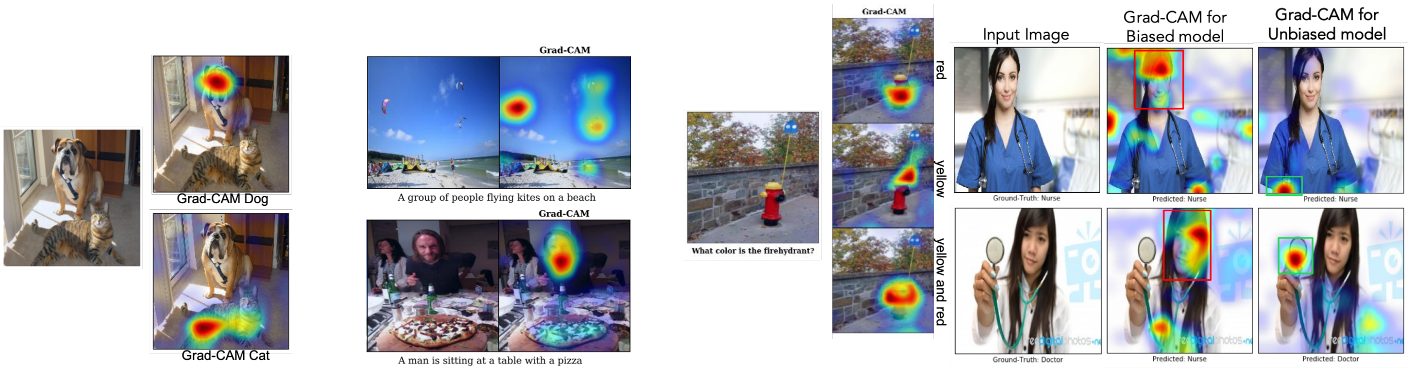
Towards Interpretable, Transparent And Unbiased AI

Convolutional Neural Networks (CNNs) and other deep networks have enabled unprecedented breakthroughs in a variety of computer vision tasks, from image classification to object detection, semantic segmentation, image captioning, visual question answering, and visual dialog. My thesis goal is to build algorithms that provide explanations for decisions emanating from any deep network– in order to debug and diagnose network errors, enable knowledge transfer between humans and AI, correct unwanted biases that may be learned by a network, and make models right for the right reasons.

In order to build trust in intelligent systems and move towards their meaningful integration into our everyday lives, it is clear that we must build 'transparent' models that are capable of explaining their decisions. The core of my research has been to develop explainable algorithms to make deep networks interpretable/transparent and to demonstrate the usefulness of this transparency at three different stages of AI evolution. First, when AI is significantly weaker than humans and not yet reliably 'deployable', the goal of transparency and explanations is to identify the failure modes or biases of models, thereby identifying avenues to make the models more accurate (say by providing focused feedback). Second, when AI is on par with humans and reliably 'deployable' (*e.g.*, image classification trained on sufficient data), the goal is to establish appropriate trust and confidence in users. Third, when AI is significantly stronger than humans (*e.g.*, AlphaGo in the game of Go), the goal of explanations is machine teaching – i.e., a machine teaching a human about how to make better decisions.



(a) Grad-CAM – Image classification (b) Grad-CAM – Image captioning (c) Grad-CAM – VQA (d) Grad-CAM – Uncovering & fixing dataset bias

Figure 1: **Grad-CAM applications:** (a,b,c) show Grad-CAM visual explanations for image classification, captioning and VQA. We see that Grad-CAM provides class-discriminative visualizations for variety of tasks/models, helping users place appropriate trust in model decisions. (d) shows an application of Grad-CAM for identifying biases in datasets. Grad-CAM explanations for 2 models – one trained on biased and other on unbiased dataset. We find that the biased model (middle column) looks at incorrect regions such as face/hairstyle (shown in red bounding boxes) and generalizes poorly compared to the unbiased model that looks at appropriate regions such as half sleeves for nurses, and stethoscopes for doctors (shown in green bounding boxes).

Visual Explanations. In my earlier work towards understanding why deep models predict what they predict, we developed Grad-CAM [1], Gradient-based Class Activation Mapping (presented at ICCV'17), a general technique for providing visual explanations from any deep network without requiring any change in architecture or retraining. Grad-CAM uses the gradients of any target decision flowing into the final convolutional layer to produce a coarse localization map highlighting the important regions in the image for making the decision. Grad-CAM is widely applicable to a variety of architectures and tasks such as image classification, image captioning and VQA. Grad-CAM for the classification example in Fig. 1 (a) and VQA example in Fig. 1 (c) shows its class-discriminative property. For *e.g.*, in Fig. 1 a (top), for explaining "dog", only the dog is highlighted and not the cat. For image captioning, the Grad-CAM for the generated caption in Fig. 1 b (bottom) correctly highlights the pizza and the man, but ignores the woman nearby, since "woman" is not mentioned in the caption.

The invention of Grad-CAM and our analysis revealed a number of interesting and novel findings – We found that even simple non-attention based captioning and VQA models learn to look at relevant image regions when making decisions, questioning the need for attention models. We showed how explanations can not only help establish trust with humans, but also help untrained users successfully discern a stronger network from a weaker one, *even when both make identical predictions.* We also made a first attempt at showing how interpretability helps in diagnosing failure modes of current deep models, and in uncovering biases in datasets. We created a proof-of-concept experiment – a simple classification problem of distinguishing *Doctors* from *Nurses*, curating a dataset from a popular search engine. Grad-CAM visualizations for the model predictions (see the red boxes in middle column of Fig. 1 (d)) revealed that models trained with this dataset learned to look at the person's face / hairstyle to distinguish nurses from doctors, thus learning a gender stereotype. Turns out the image search results were gender-biased. Through these intuitions gained from Grad-CAM, we reduced bias in the training set and retrained models which not only generalize better, but also look at the right regions (see green boxes in Fig. 1 (d)).

Impact: Grad-CAM has been cited by over 1300 papers, with over 15 re-implementations in various deep learning frameworks, and is also currently included as an official library in Pytorch and Tensorflow. Our interactive online demo (`http://gradcam.cloudcv.org`) has been accessed over 25000 times. There have been extensions of Grad-CAM for videos, 3D models and text. I was recruited by Facebook's Applied Machine Learning team to implement Grad-CAM for their vision models. As a result, Grad-CAM now runs on *all* the hundreds of millions of images uploaded to Facebook everyday. This tool is being used by researchers and model builders to better understand decisions made by their models.

Facilitating Knowledge Transfer between Humans and AI. Current deep models are extremely data hungry, so the most effective approach involves feeding a lot of labeled data. Instance-level annotations is expensive, private, or scarce in many applications. Hence it is important to find cheap and efficient ways to supervise networks. One such highly informative (and relatively cheap) form of supervision is domain knowledge from humans. Can we bake this knowledge into our models?

In my ECCV'18 paper [2], we proposed an approach NIWT (Neuron Importance-aware Weight Transfer) that allows humans to incorporate their domain knowledge into deep networks. Specifically, we focused on extending a pretrained network to build classifiers for novel classes (such as "Red bellied Woodpecker") given only descriptions from a domain expert (like *"A Red Bellied Woodpecker is a small, round bird with a white breast, red crown, and spotted wings"*) without any images of these classes. More specifically, NIWT learns a mapping between the class-discriminative neuron importance from the network and the domain-specific descriptions from an expert, allowing us to ground domain knowledge into network's neurons – *e.g.* neuron ID 32 looks for "red crown", and neuron ID 47 looks for "spotted wings". Given expert descriptions for novel classes (such as "Red Bellied Woodpecker"), we can optimize for the new classifier weights such that the neurons that are deemed important by the mapping (say neuron IDs 32 and 47) end up being important while making the decision.

In addition to outperforming the state of the art in zero-shot-learning on two datasets, by grounding neuron-importances in semantic human interpretable domains, NIWT is automatically able to explain network decisions in the form of text.

Leveraging Explanations to Teach AI. Today's state-of-the-art deep models, especially for vision and language tasks are known to rely heavily on superficial correlations in training data. As a result, these models are often biased by language priors, and do not make predictions sufficiently grounded in the image. Image captioning models often generate phrases like "standing next to a tree" when talking about giraffes because trees tend to co-occur in images of giraffes in the COCO train set, and VQA models blindly answer "yellow" when asked, "What color are the bananas?". This becomes apparent when explanation modalities such as Grad-CAM are employed to assess the evidence that the models are basing their decisions on. Using context or overly relying on priors while making decisions makes systems develop internal (incorrect) biases that don't help generalize to new data distributions. For eg., learning that boats are always in water or that traffic cones are always orange, will prevent the model from recognizing boats outside of water, and identifying traffic cones of different color.

While somewhat dissatisfying, these findings are not entirely surprising – after all, standard training protocols do not provide any guidance for visual grounding. Instead, models are trained on i/o pairs and must resolve grounding from co-occurrences – a challenging task, especially in the presence of more direct and easier to learn correlations in language. Consider our previous example question – the words 'color', 'banana', and 'yellow' are given as discrete tokens that will trivially match in every occurrence when these concepts are referenced. In contrast, actually grounding this question requires dealing with all visual variations of bananas and learning the common feature of things described as 'yellow'. To address this, we explore if giving a *small hint* in the form of human attention can help improve grounding and reliability.

In my ICCV'19 paper [3], we introduced Human Importance-aware Network Tuning (HINT), which enforces a ranking loss between human annotations of input importance and Grad-CAM explanations from a deep network – updating model parameters via a gradient-of-gradient step. Importantly, this constrains models to not only look at the correct regions but to also be sensitive to the content present there when making predictions. We apply HINT to two tasks – Visual Question Answering (VQA) and image captioning – and find our approach that forces visual grounding also significantly improves task performance and human trustworthiness. While we experiment with HINT in the context of vision-and-language problems, the approach itself is general and can be applied to focus model decisions on specific inputs in any context.

Making AI Right for Right Reasons. While measuring progress on standard benchmarks has led to significant model innovation, in our effort to improve accuracy on a specific dataset, we are inadvertently building models that are able to exploit undesirable shortcuts in the data. In my recent work [4], we analysed the *reasoning abilities* of current Visual Question Answering (VQA) models. We noticed that current VQA models have consistency issues – they are able to answer seemingly harder reasoning questions right (e.g. "Is the banana ripe enough to eat?"), but fail on simpler, perception questions (e.g., "Are the bananas mostly green or yellow?") – indicating that the model possibly answered the original question for the wrong reasons, even if the answer was right. In order to quantify the extent to which this phenomenon occurs, we collected a new dataset of *perception* sub-questions for questions in the VQA dataset requiring reasoning abilities and observed that state-of-the-art models are inconsistent ∼30% of the time. We then proposed an approach which encourages the model to attend to the same parts of the image when answering the reasoning question and the perception sub-questions. We showed that our approach improves model consistency by 7.6%, also marginally improving its performance on the reasoning questions in VQA, while also displaying qualitatively better attention maps. The key takeaways of our work are: 1) it is important to use trust metrics (such as consistency & reliability) besides accuracy, and 2) it is important to understand our data well and develop models to use it efficiently, i.e., lesser data collected and used in a specific way can be more useful than large amounts of data collected and used in a generic way. This leads to better reasoning models which cannot easily rely on undesirable shortcuts, thereby making models *right for right reasons.*

Overall my research focuses on building algorithms that facilitate understanding decisions made by AI systems, allow ease of knowledge transfer between humans and AI, and teaching AI to be bias-free, which is important not just for generalization, but also for fair and ethical outcomes as more algorithmic decisions are made in society.

References

[1] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization,” in *ICCV*, 2017.

[2] R. R. Selvaraju, P. Chattopadhyay, M. Elhoseiny, T. Sharma, D. Batra, D. Parikh, and S. Lee, “Choose Your Neuron: Incorporating Domain Knowledge through Neuron-Importance,” in *ECCV*, 2018.

[3] R. R. Selvaraju, S. Lee, Y. Shen, H. Jin, S. Ghosh, L. Heck, D. Batra, and D. Parikh, “Taking a HINT: Leveraging Explanations to Make Vision and Language Models More Grounded,” in *ICCV*, 2019.

[4] R. R. Selvaraju, P. Tendulkar, D. Parikh, E. Horvitz, M. Ribeiro, B. Nushi, and E. Kamar, “SQuINTing at VQA Models: Interrogating VQA Models with Sub-Questions,” *Under review*, 2019.