CSCE 421: Machine Learning

Deep Learning New Trends

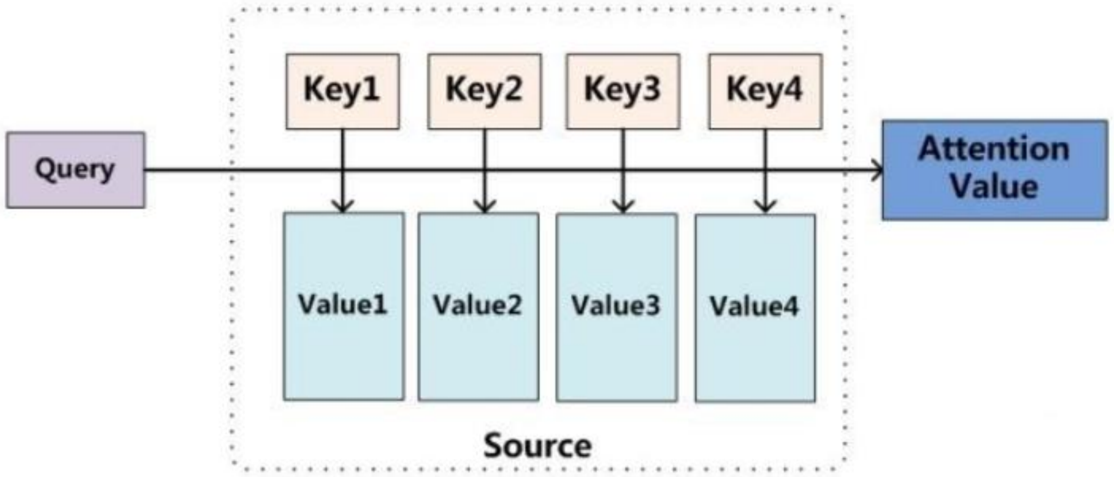
Texas A&M University

Outline

- Attention, Transformer, Feedback
- Deep Generative Models
- Automatic Deep Models
- Explainable AI
- Conversational AI

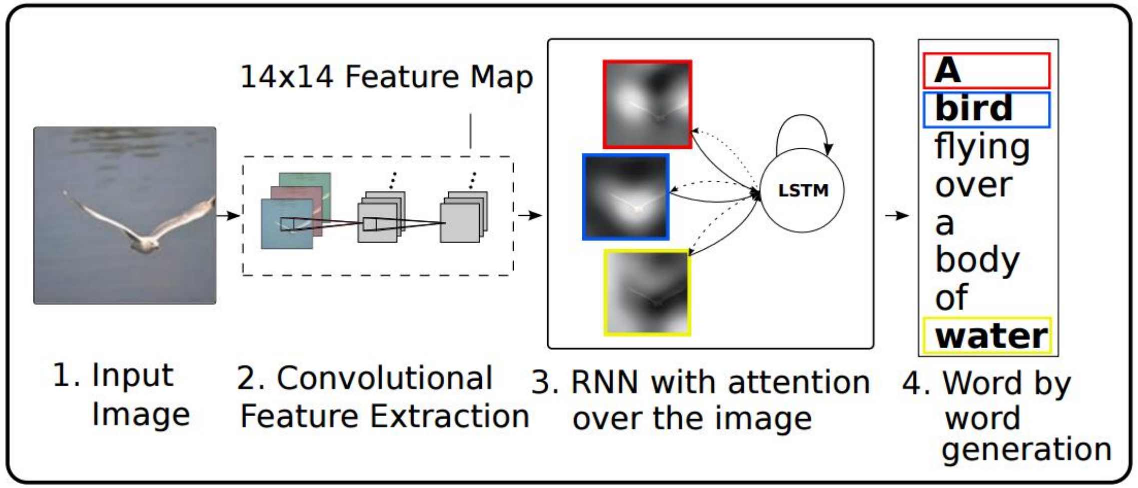
Attention

- Attention mechanism: addressing



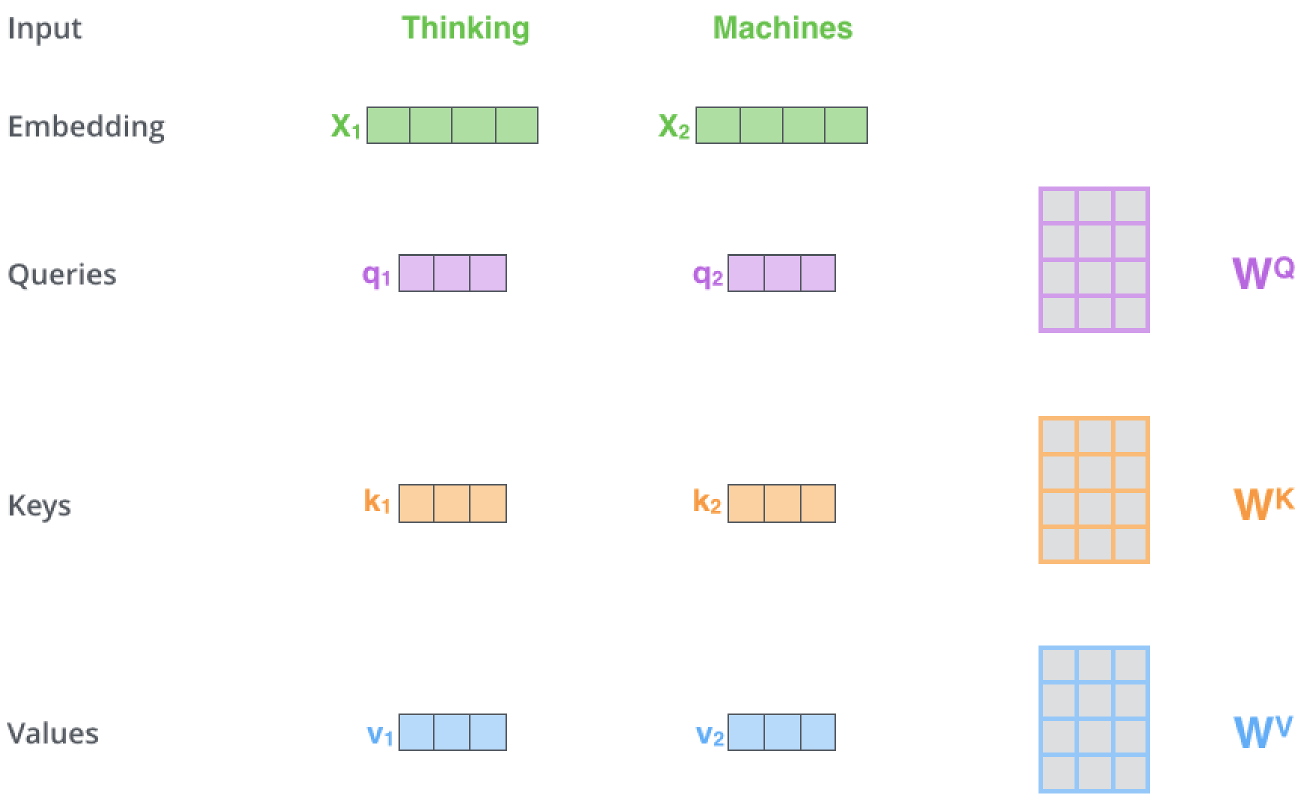
Attention

- Visual Attention Model
  - Extract feature vectors with a feed-forward network
  - Use a recurrent network to iteratively update the attention for each output word (the bright regions)
  - Obtain meaningful correspondences between words and attentions



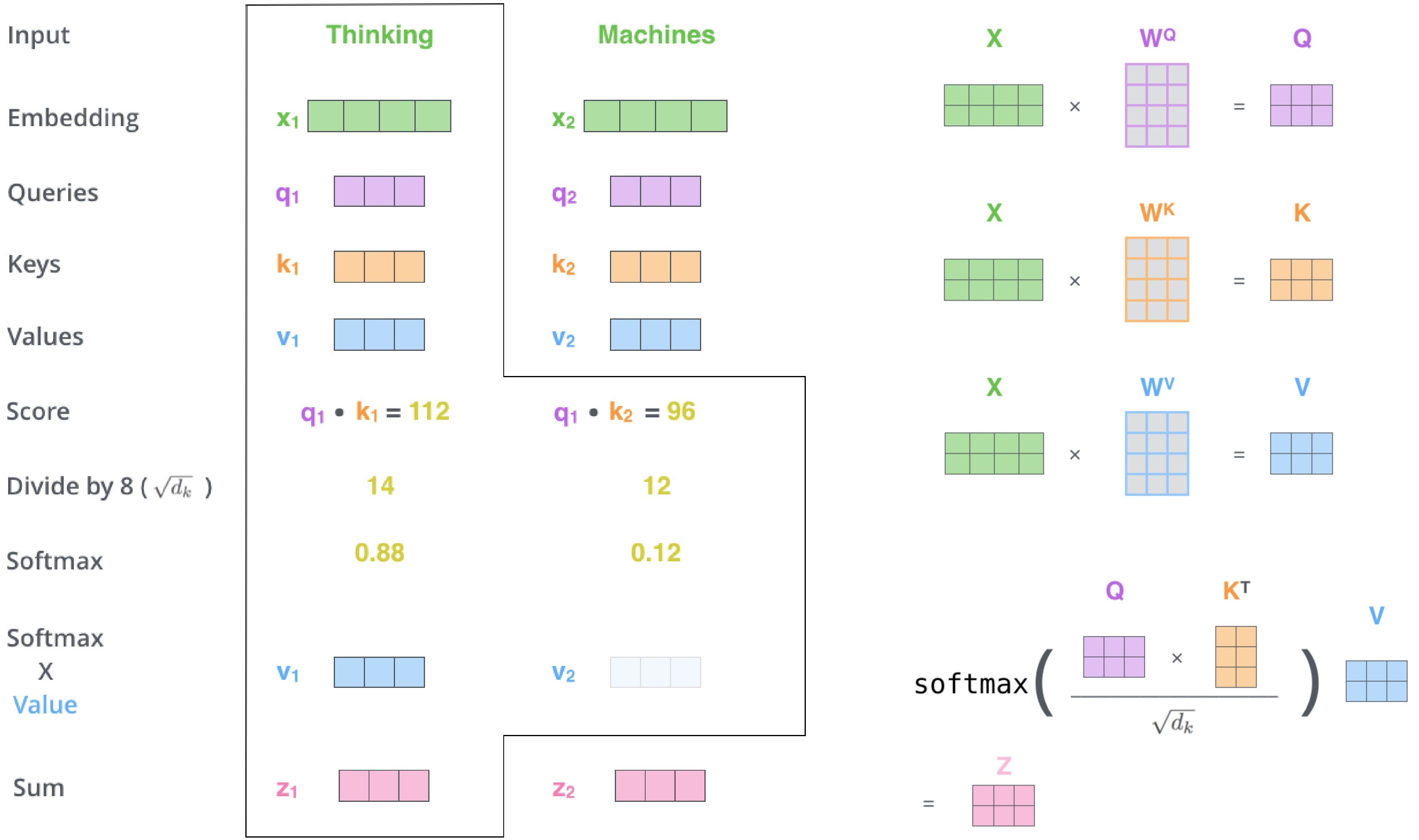
Xu, Kelvin, et al. "Show, attend and tell: Neural image caption generation with visual attention." *International conference on machine learning*. 2015.

Self-Attention

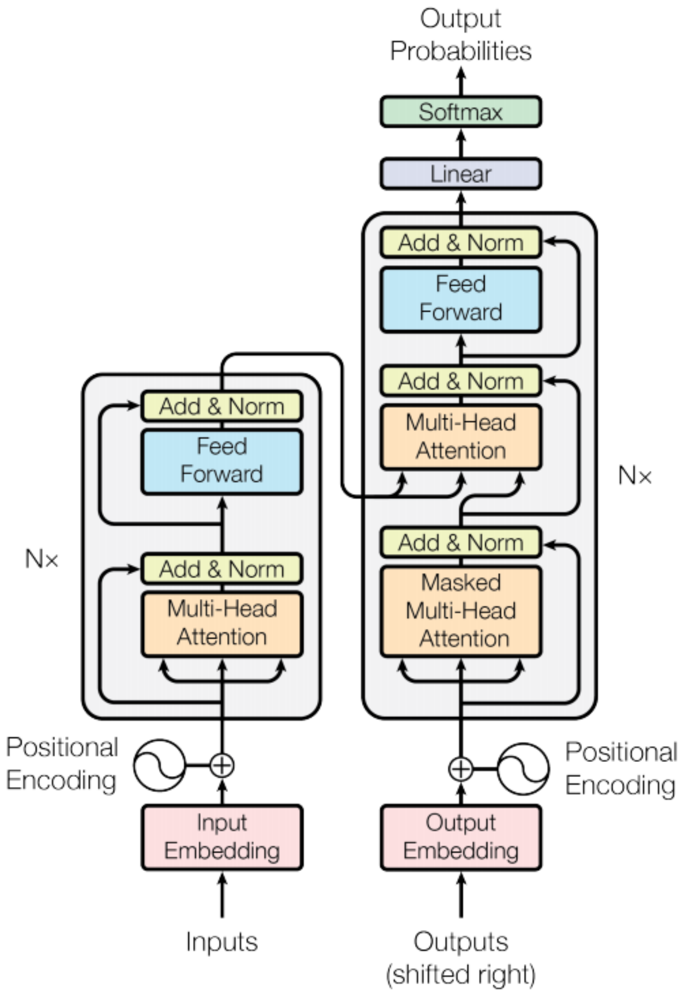


Source: https://jalammar.github.io/illustrated-transformer/

Self-Attention

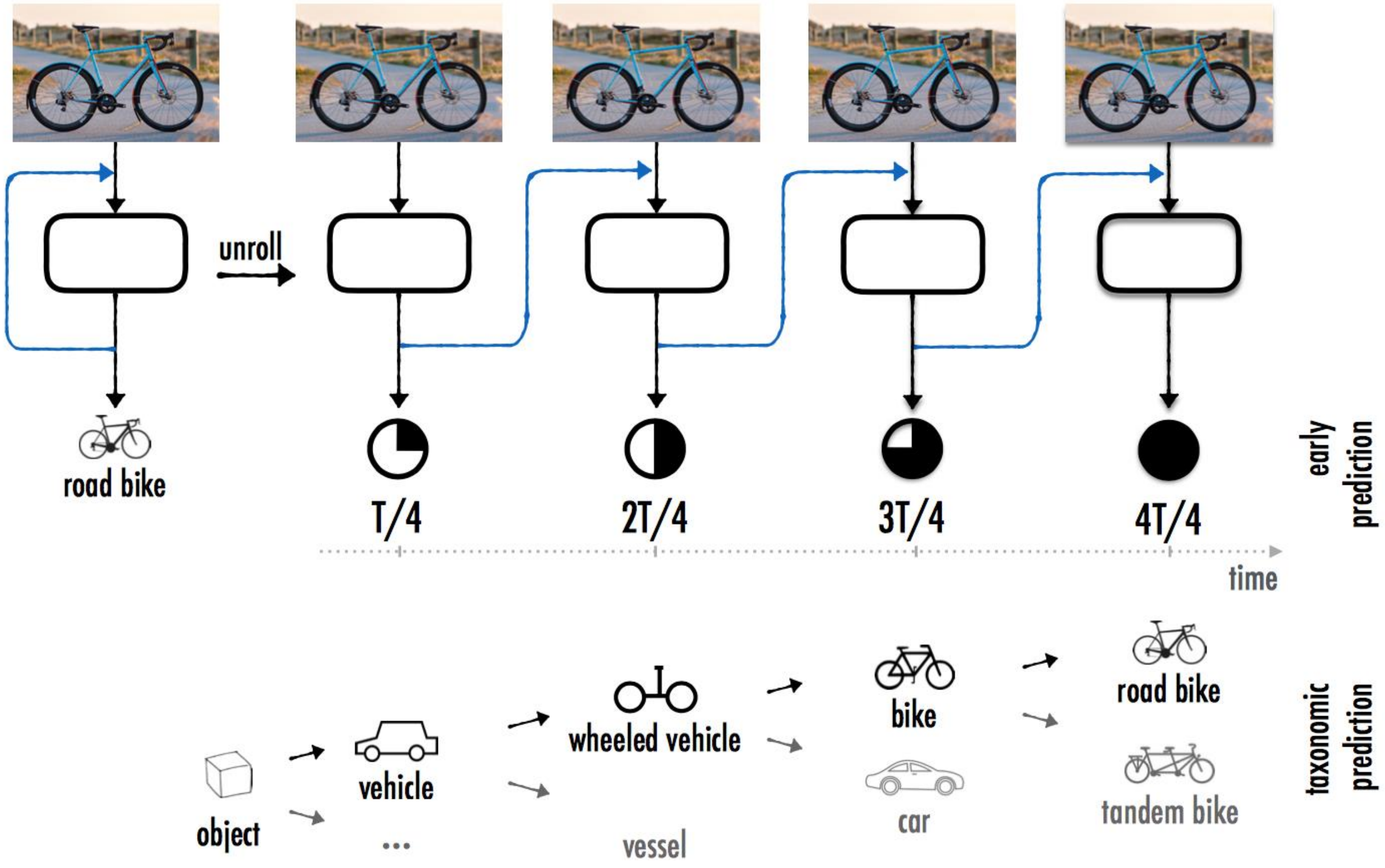


Transformer



Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems*. 2017.

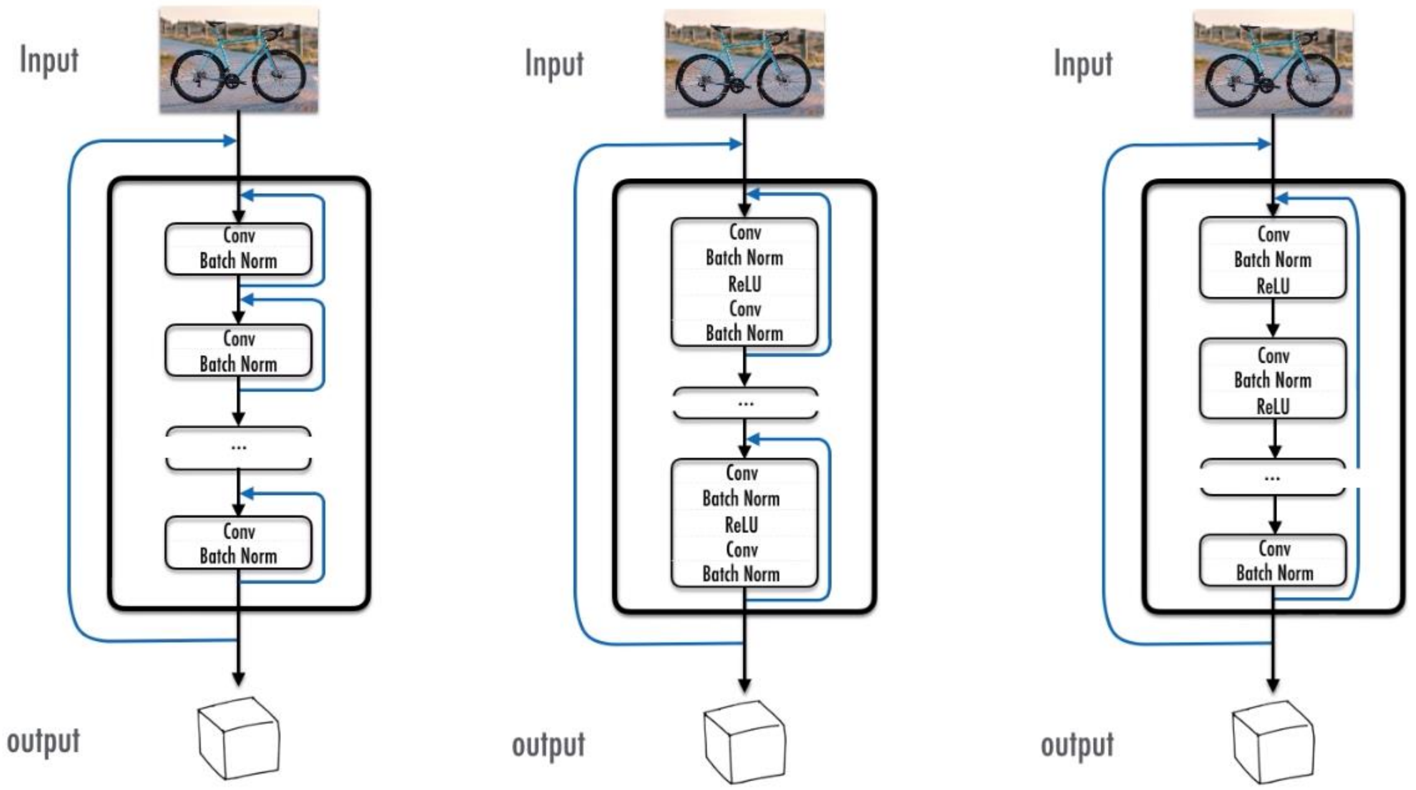
Feedback Networks



Zamir, Amir R., et al. "Feedback networks." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017.

Feedback Networks

- Design



Resources: http://feedbacknet.stanford.edu

Outline

- Attention, Transformer, Feedback
- Deep Generative Models
- Automatic Deep Models
- Explainable AI
- Conversational AI

Deep Generative Models

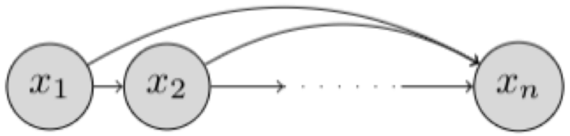
- Main idea: Learn to understand data through generation

- Setup:
  - Generative models:
    - Generate new data instances.
    - Recover the joint probability $p(X,Y)$, or $p(X)$ from given n examples X.
  - Discriminative models:
    - Discriminate between different kinds of data instances.
    - capture the conditional probability $p(Y \mid X)$.
  - Maximum-likelihood objective: $\prod_i p_\theta(x) = \sum_i \log p_\theta(x)$
  - Generation: sampling from $p_\theta(x)$.

Deep Generative Models

Autoregressive Models

- Generate: sample one step at a time, conditioned all the previous steps



- Factorize the joint distribution over the n-dimensions:

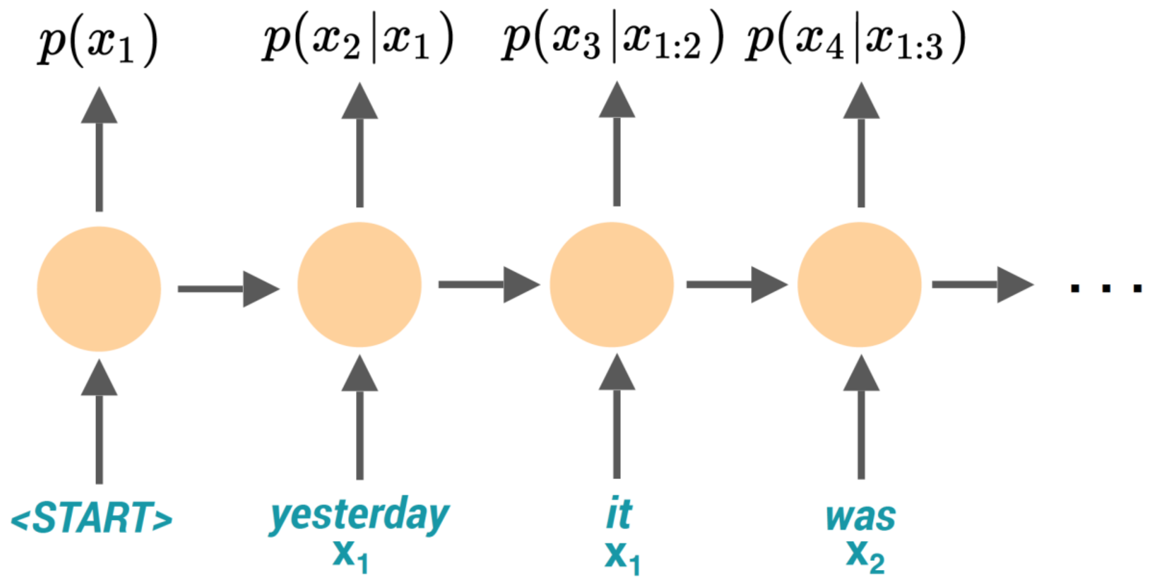
$$p(x) = \prod_{i=1}^{n} p(x_i | x_1, x_2, \dots, x_{i-1}) = \prod_{i=1}^{n} p(x_i | x_{<i})$$

- Discrete x: produce a probability for each possible value.
- Continuous x: produce parameters of a simple distribution.

Deep Generative Models

Autoregressive Models

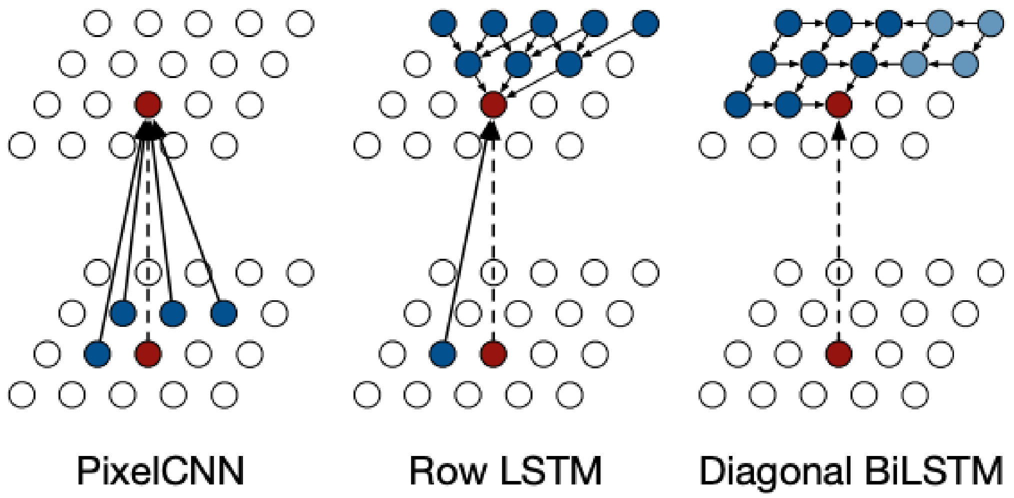
- Example: RNNs for Autoregressive Language Modeling



Deep Generative Models

Autoregressive Models

- PixelRNN
  - Apply language modeling on images.
  - 2-d images: grid LSTM



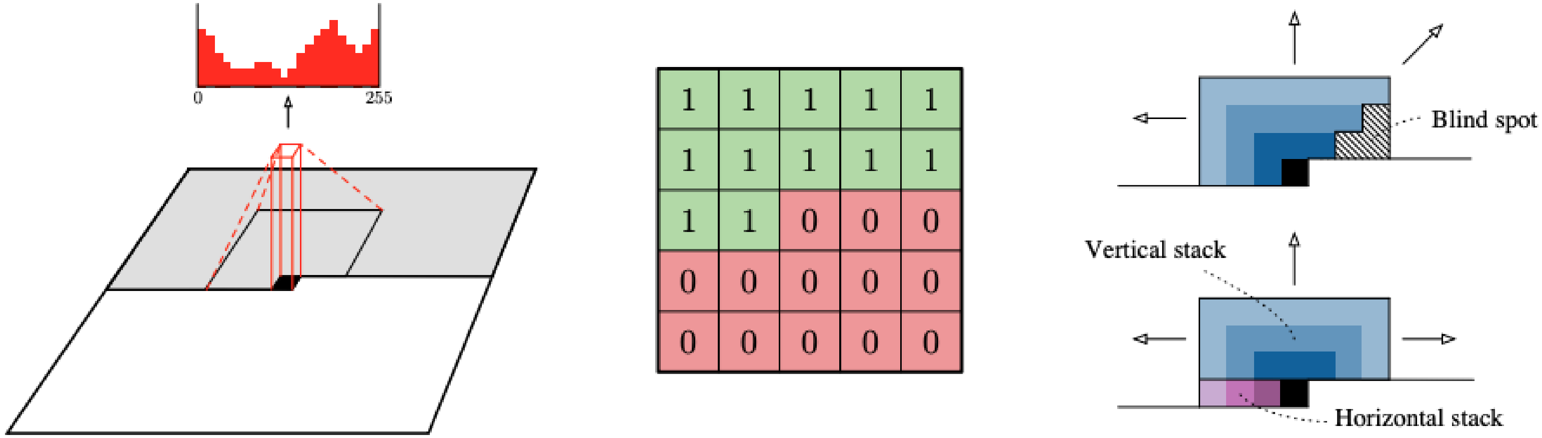
  - PixelCNN (old): has bland spot

Oord, Aaron van den, Nal Kalchbrenner, and Koray Kavukcuoglu. "Pixel recurrent neural networks." *arXiv preprint arXiv:1601.06759* (2016).

Deep Generative Models

Autoregressive Models

- PixelCNN
  - Overcome blind spot problem

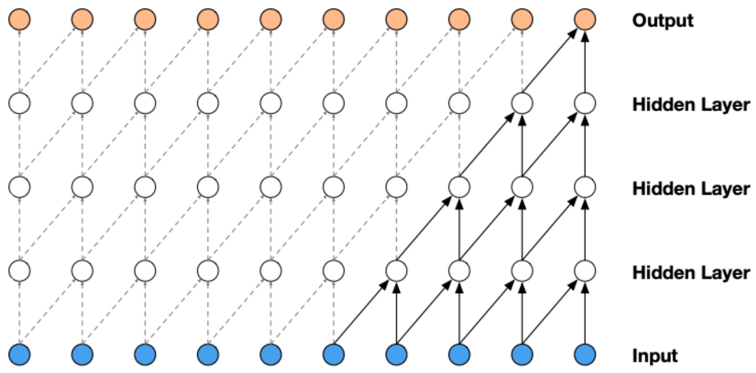


Van den Oord, Aaron, et al. "Conditional image generation with pixelcnn decoders." *Advances in neural information processing systems*. 2016.

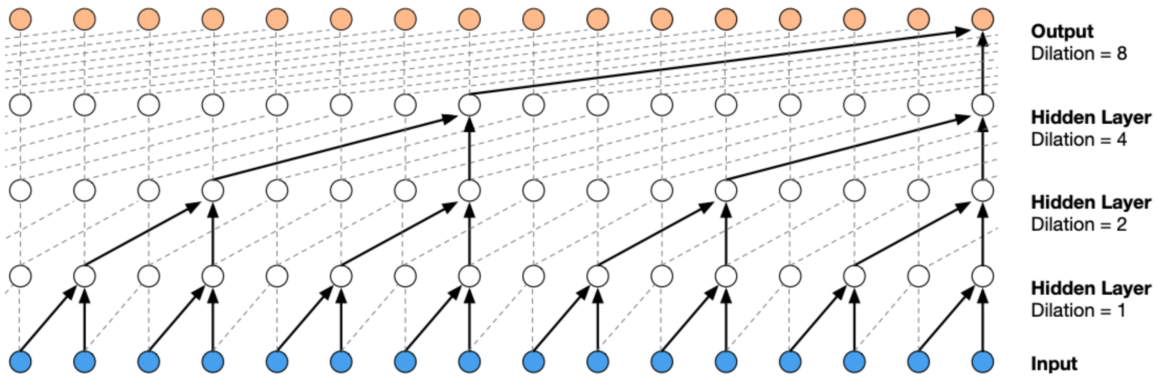
Deep Generative Models

Autoregressive Models

- WaveNet



Visualization of a stack of causal convolutional layers

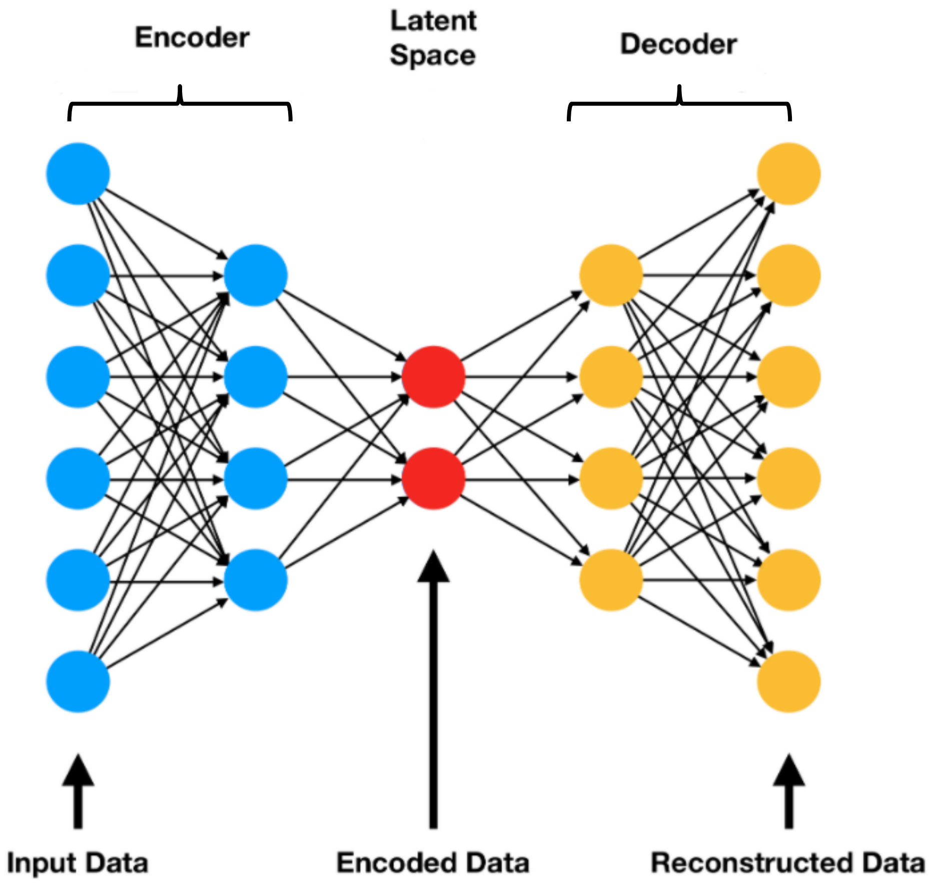


Visualization of a stack of dilated causal convolutional layers.

Oord, Aaron van den, et al. "Wavenet: A generative model for raw audio." *arXiv preprint arXiv:1609.03499* (2016).

Deep Generative Models

Autoencoder



- Idea: compression as implicit generative modeling
- Output: reconstructed data
- Label: input data
- Loss: $L = (x - \hat{x})^2$
- Latent space: encoded features

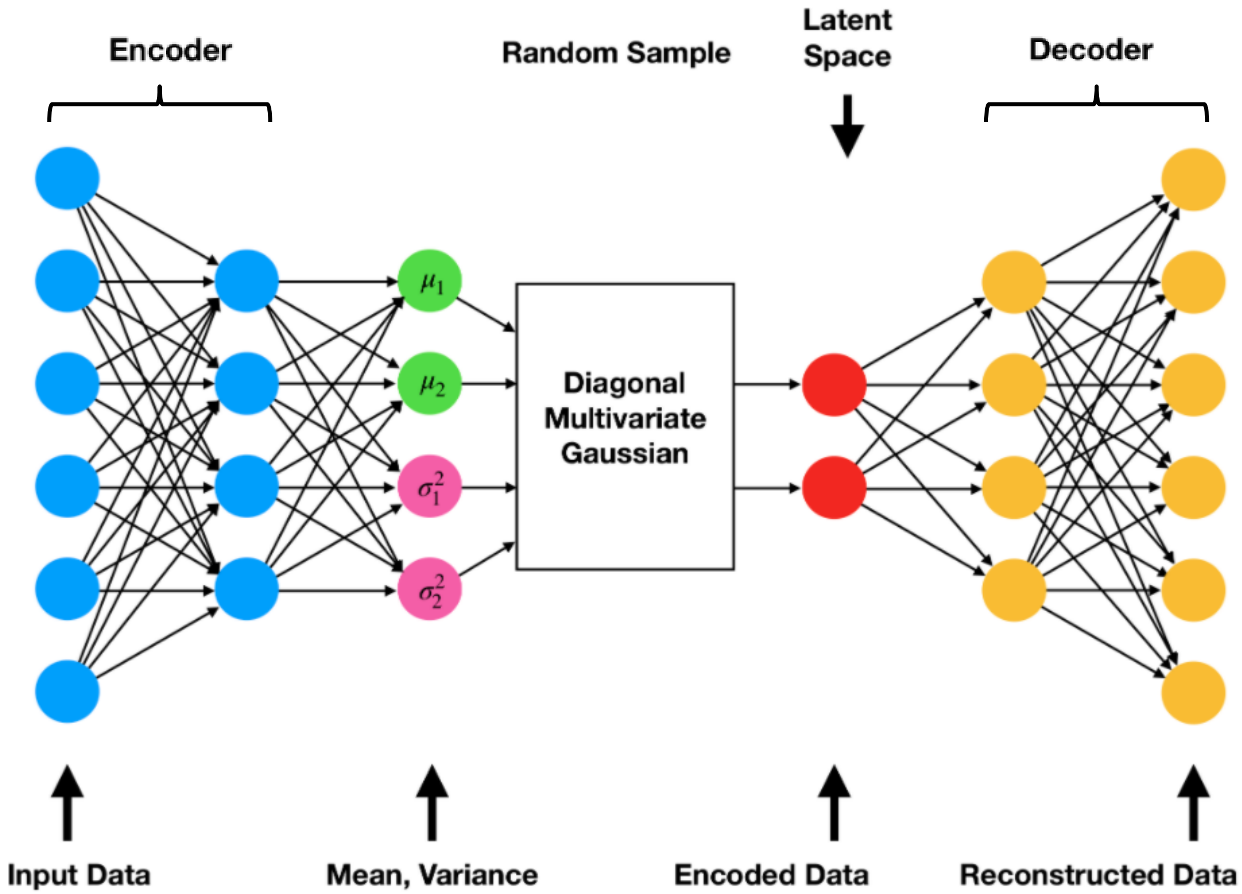
Deep Generative Models

Autoencoder

- The encoder-decoder architecture only ensures the main structured part of the information can go through and be reconstructed
- The dimension of the latent space and the depth of autoencoders need to be carefully controlled
  - Dimensionality reduction purpose: reduce this number of dimensions and keep the major data structure information in the reduced representations.
  - Need interpretable and exploitable structures in the latent space.

Deep Generative Models

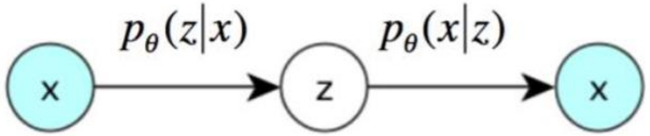
Variational Autoencoder (VAE)

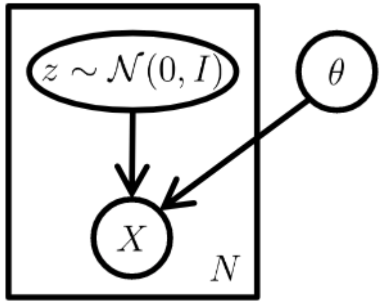


Deep Generative Models

Variational Autoencoder (VAE)

- Similar idea as variational Bayesian and graphical model





- Problem: $p_\theta(z|x)$ cannot be calculate
- Solution: approximate $p_\theta(z|x)$ with $q_\phi(z|x)$:

$$q_\phi(z|x) = \mathcal{N}(z; \mu_z(x), \sigma_z(x))$$

- Train: maximize a lower bound on log probabilities

$$\log p(x) \geq \mathbb{E}_{z \sim q(z|x)}[\log p(x|z) + \log p(z) - \log q(z)]$$

Deep Generative Models

Variational Autoencoder (VAE)

- Problems:
  - Encoder and decoder's output distributions are typically limited (diagonal-covariance Gaussian or similar)
  - This prevents the model from capturing fine details and leads to blurry generations

1st epoch

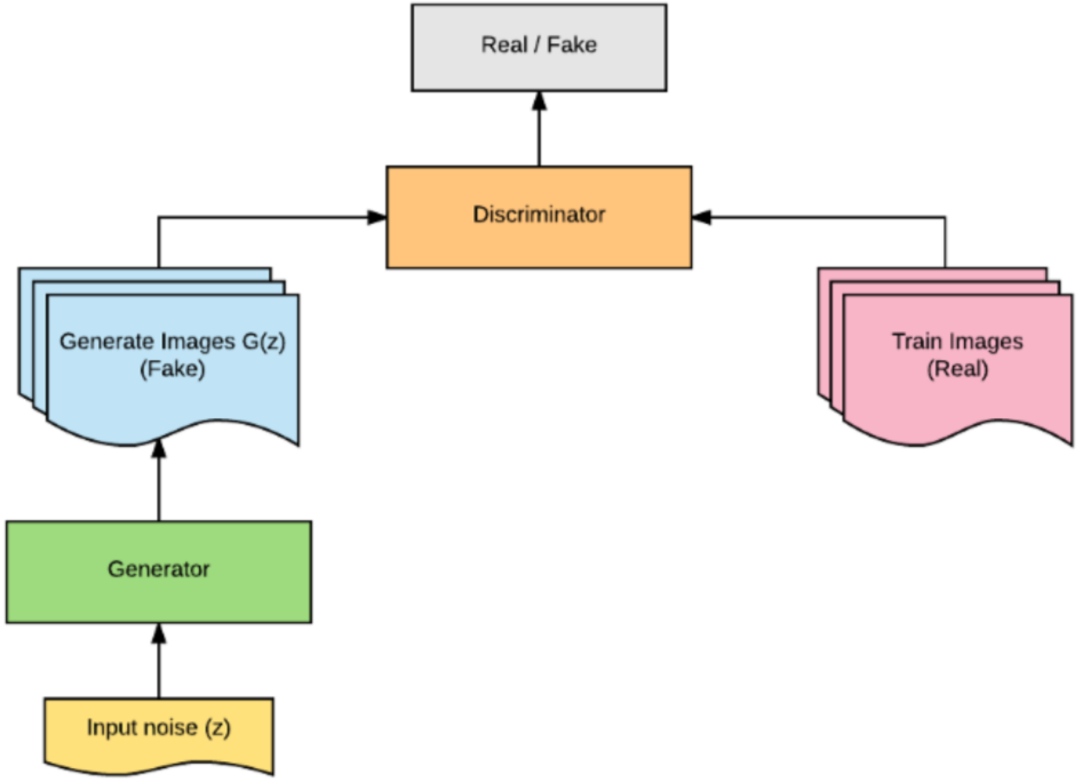
9th epoch

Original

Deep Generative Models

Generative Adversarial Networks (GAN)

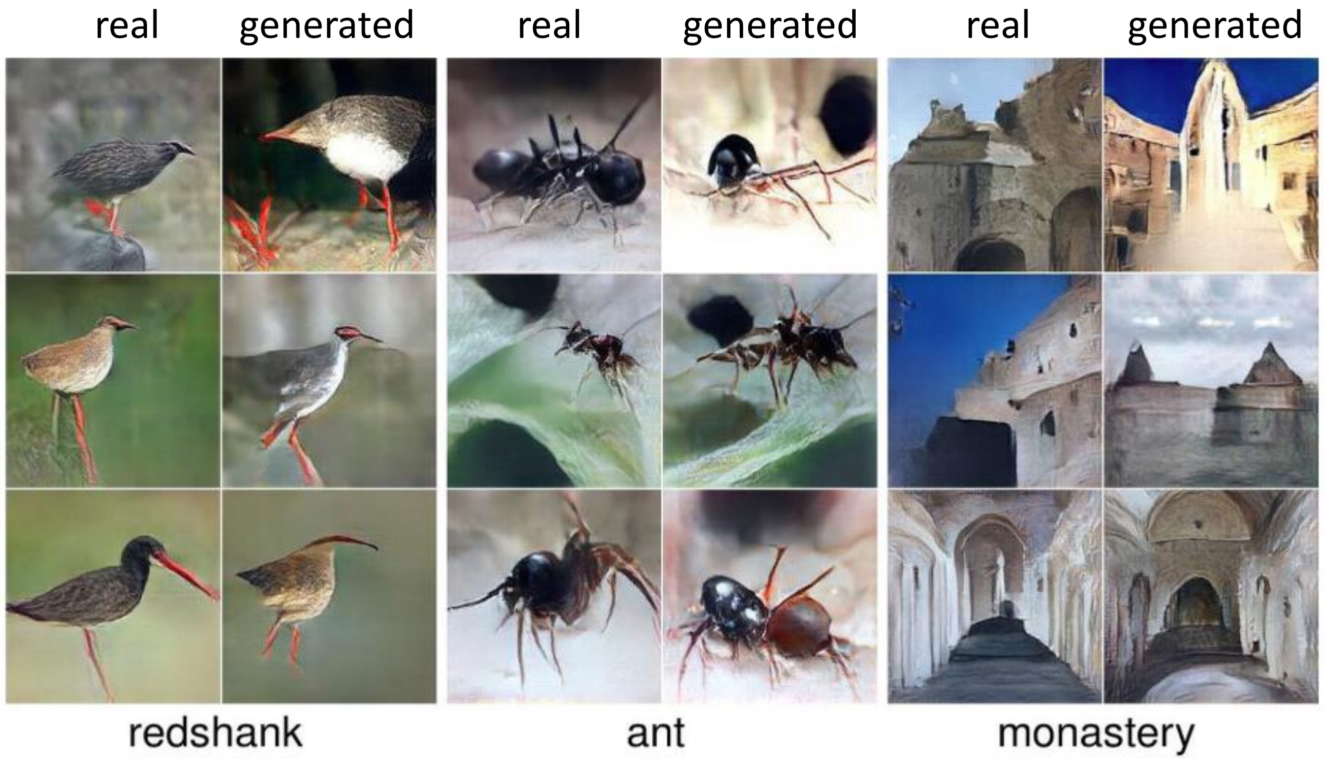
- Two networks competing
    - Discriminator: distinguishes real data and generated images from generator
    - Generator: turns random noise (input) to imitations of data



Deep Generative Models

Generative Adversarial Networks (GAN)

- Success: Small Size, Limited Scene, Simple Background …



Deep Generative Models

Generative Adversarial Networks (GAN)

- Challenges
  - Training is notoriously difficult and unstable
  - Easily biased towards either Generator or Discriminator
  - Few “decisive” success in generating “real-scale” complicated images

- GAN failure
  - The state-of-the-art GANs seem to learn “parts”, but not the correct combination way (anatomy).
  - Little success in training GANs, e.g. on ImageNet scale.

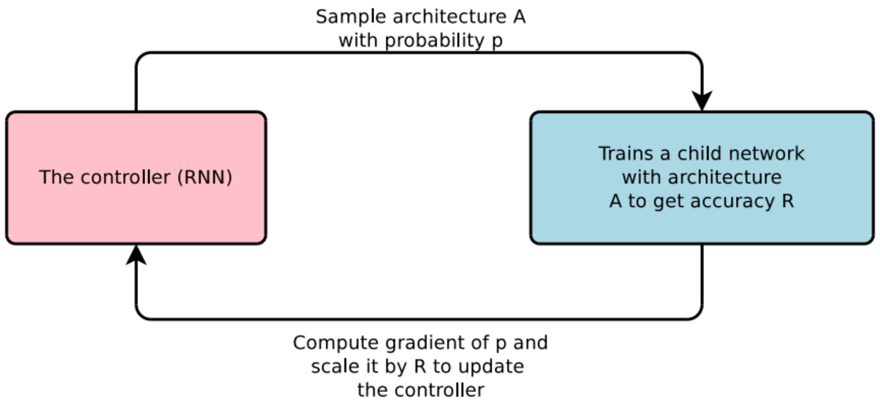
Outline

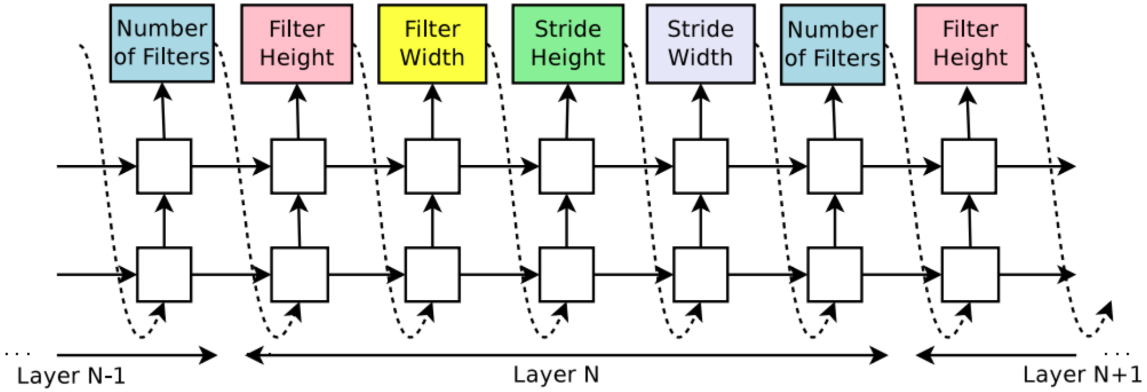
- Attention, Transformer, Feedback
- Deep Generative Models
- Automatic Deep Models
- Explainable AI
- Conversational AI

Automatic Deep Models

Neural Architecture Search (NAS)

- NAS with RL



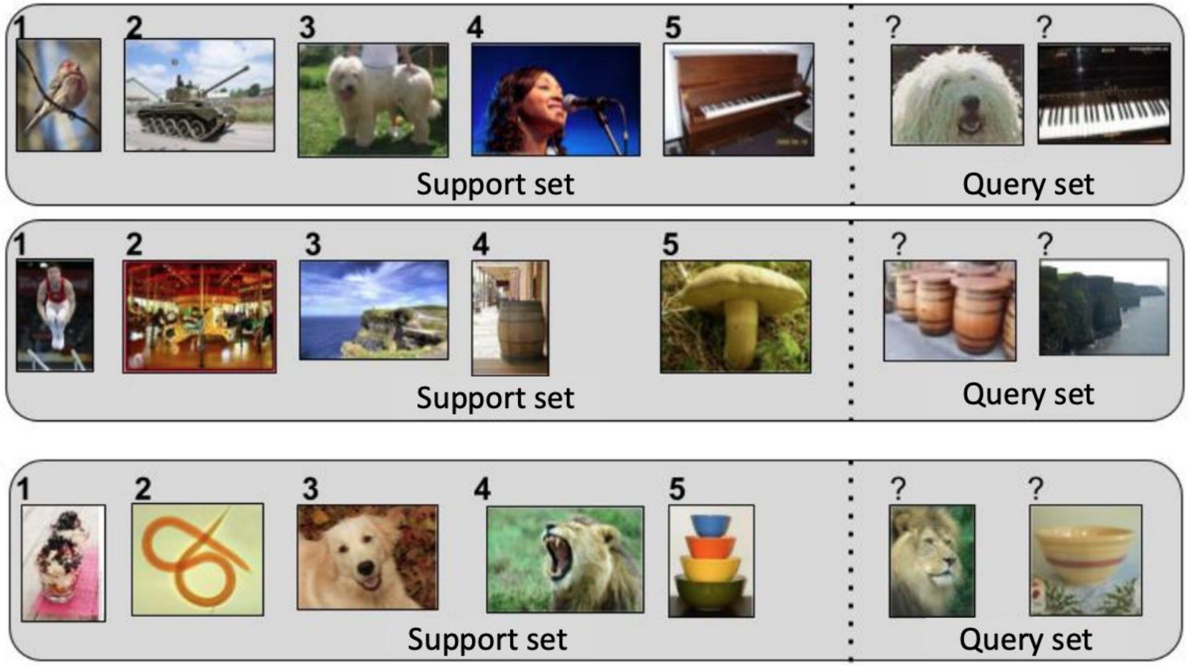


Zoph, Barret, and Quoc V. Le. "Neural architecture search with reinforcement learning." *arXiv preprint arXiv:1611.01578* (2016).

- Few-shot learning: n-shot, k-way
  - Discriminate between N classes with K examples of each.
  - Support set (known) v.s. query set (unseen)
  - Determine which of the support set classes the query sample belongs to.

Train

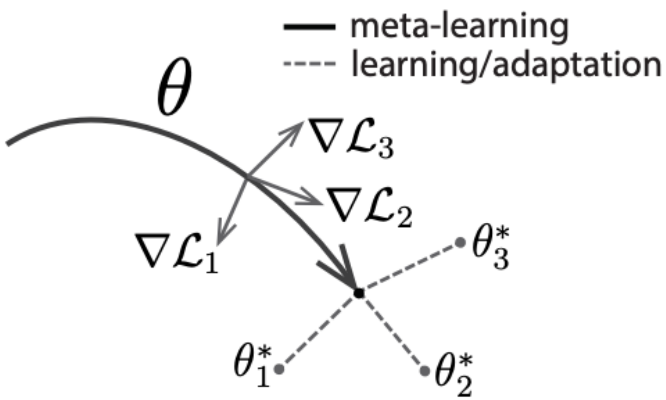
Test



Automatic Deep Models

Meta-learning

- MAML



Algorithm 1 Model-Agnostic Meta-Learning

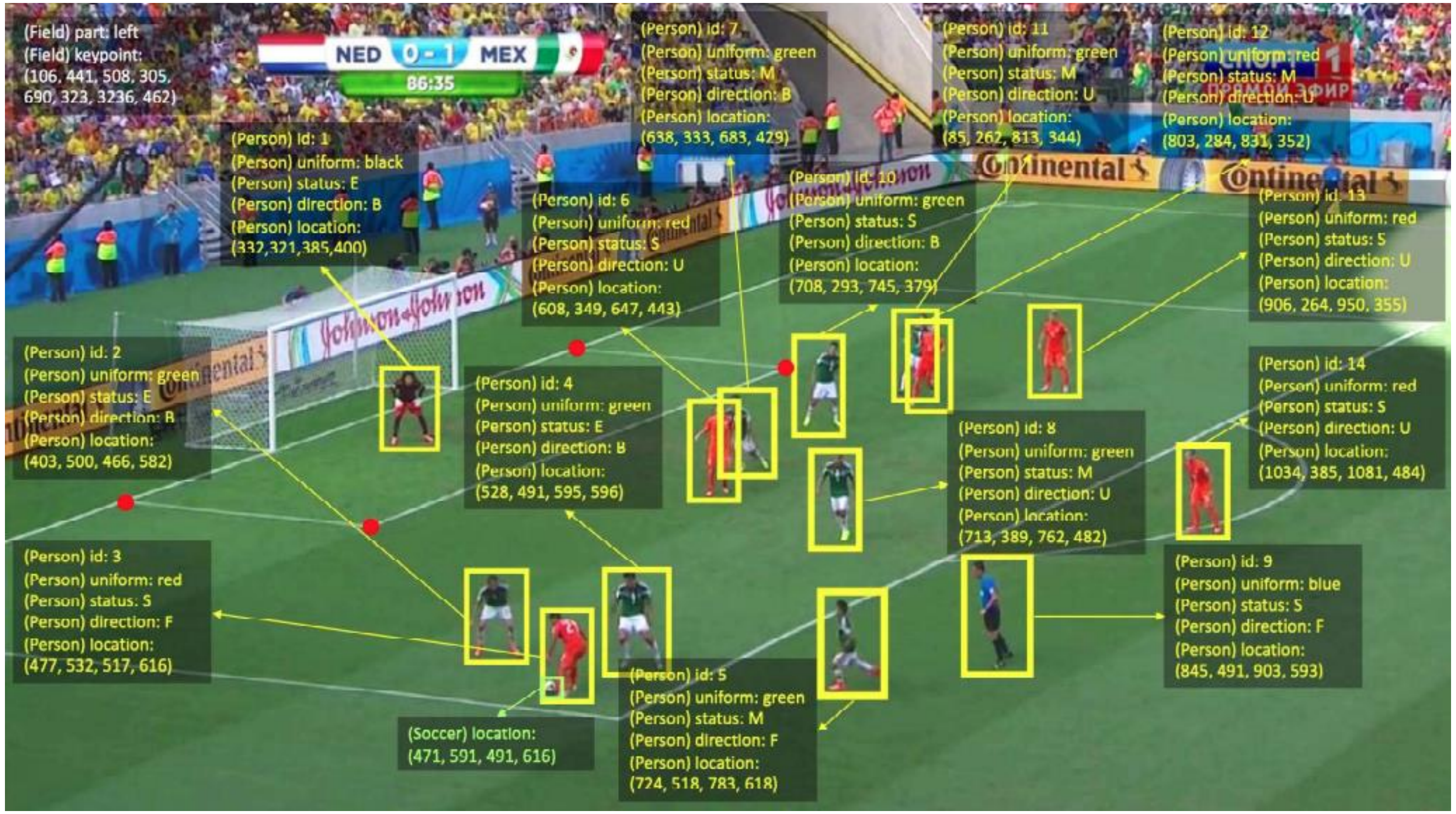
Require: $p(\mathcal{T})$: distribution over tasks
Require: α, β: step size hyperparameters

```
randomly initialize θ
while not done do
    Sample batch of tasks T_i ~ p(T)
    for all T_i do
        Evaluate ∇_θ L_{T_i}(f_θ) with respect to K examples
        Compute adapted parameters with gradient descent: θ'_i = θ − α∇_θ L_{T_i}(f_θ)
    end for
    Update θ ← θ − β∇_θ Σ_{T_i~p(T)} L_{T_i}(f_{θ'_i})
end while
```


Finn, Chelsea, Pieter Abbeel, and Sergey Levine. "Model-agnostic meta-learning for fast adaptation of deep networks." *arXiv preprint arXiv:1703.03400* (2017).

Visual Query Answering (VQA)

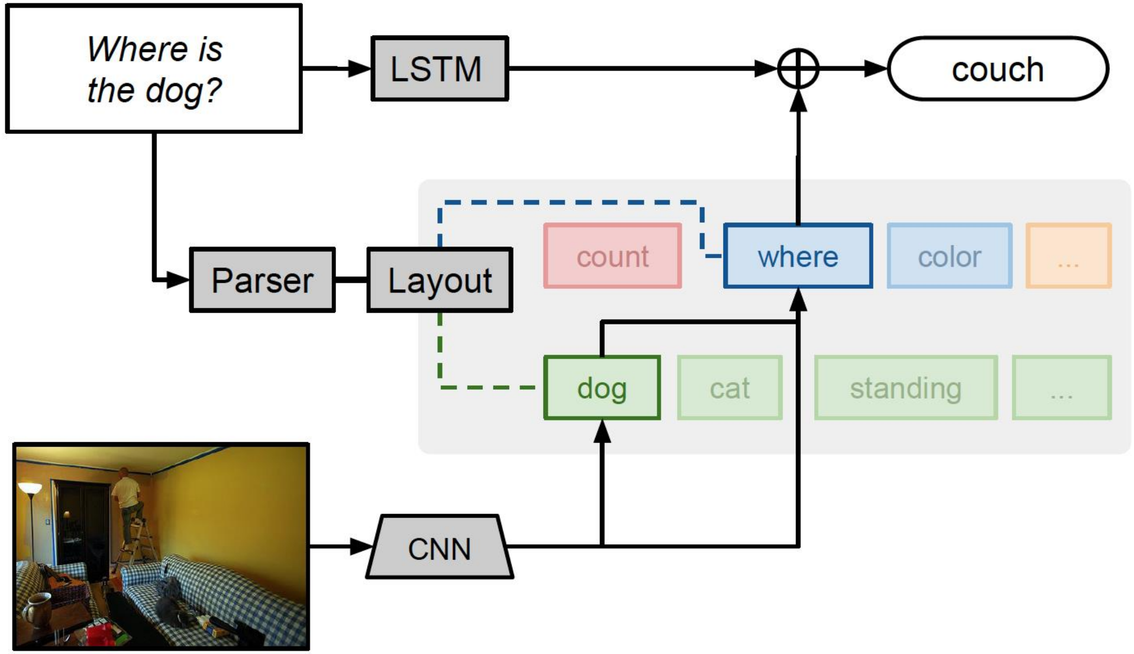
- VQA: raise a question for details of objects, or high-level understanding of the scene over images.



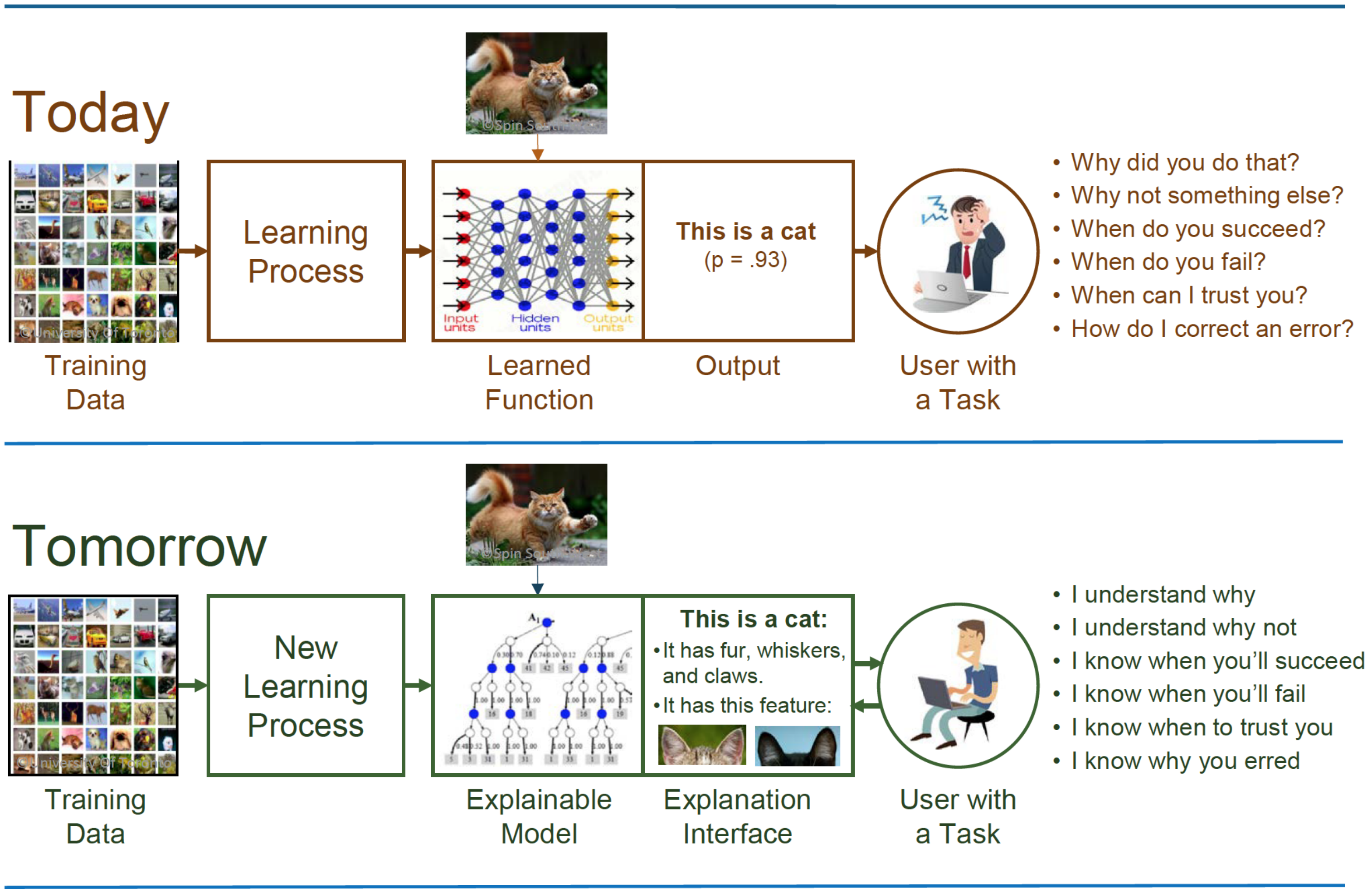
Xiong, Peixi, et al. "Visual query answering by entity-attribute graph matching and reasoning." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.

Visual Query Answering (VQA)

- Analyze the question with the parser
- Determine the basic computational units which are needed to answer the question
- Determine the relationships between the modules
- Assemble the modules and train the network jointly



Visual Query Answering (VQA)

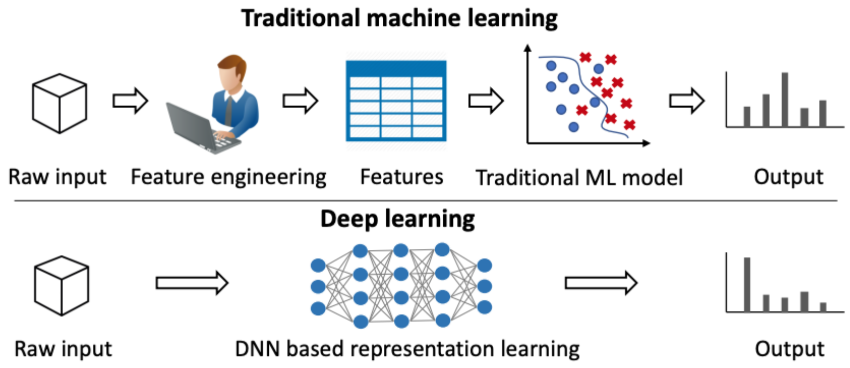


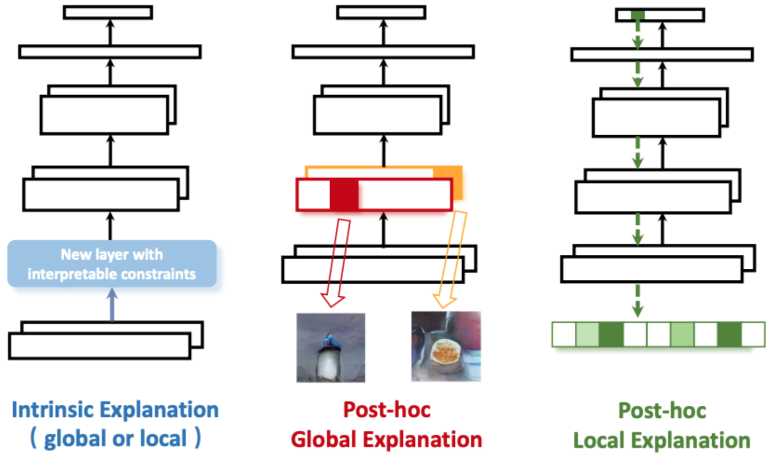
Outline

- Attention, Transformer, Feedback
- Deep Generative Models
- Automatic Deep Models
- Explainable AI
- Conversational AI

Explainable AI – Interpretability

- Goals:
  - Understand the behaviors of Machine Learning algorithms.
  - Model results trustworthiness.
  - Explain how a model arrives at a particular decision.
  - ……
- Tools:
  - Mimic cognitive science
  - Relate black box models with simpler models
  - Statistical analysis of results
  - Visualization

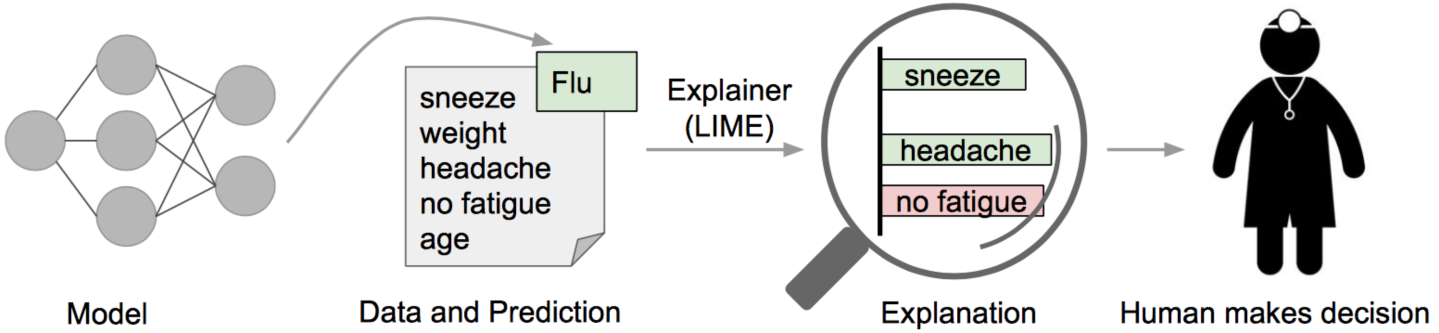


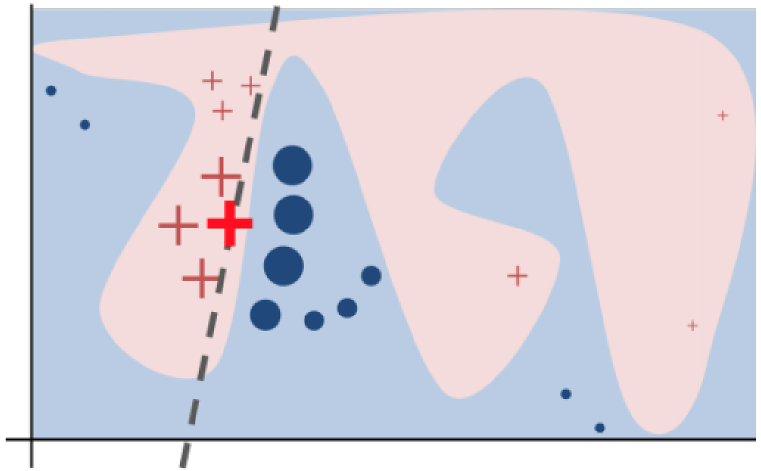


- Du, Mengnan, Ninghao Liu, and Xia Hu. "Techniques for interpretable machine learning." *Communications of the ACM* 63.1 (2019): 68-77.
- Lipton, Zachary C. "The mythos of model interpretability." *Queue* 16.3 (2018): 31-57.

Explainable AI – Interpretability

Explaining individual predictions





- Decision of the black-box model: cannot be approximated well by a compact functional.
- LIME: samples instances, gets predictions, and weighs them by the proximity to the instance being explained (by size).
- Red across: the instance being explained.
- The dashed line: the learned explanation that is locally (but not globally) faithful

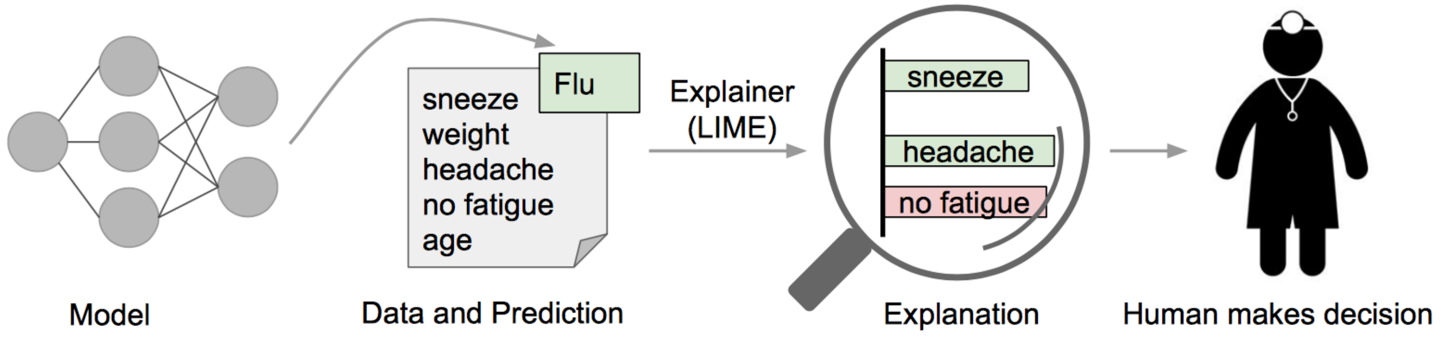
Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should I trust you?" Explaining the predictions of any classifier." *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016.

Explainable AI – Interpretability

LIME: Local Interpretable Model-agnostic Explanations

- Interpretable Data Representations
- Fidelity-Interpretability Trade-off
- Sampling for Local Exploration
- Sparse Linear Explanations

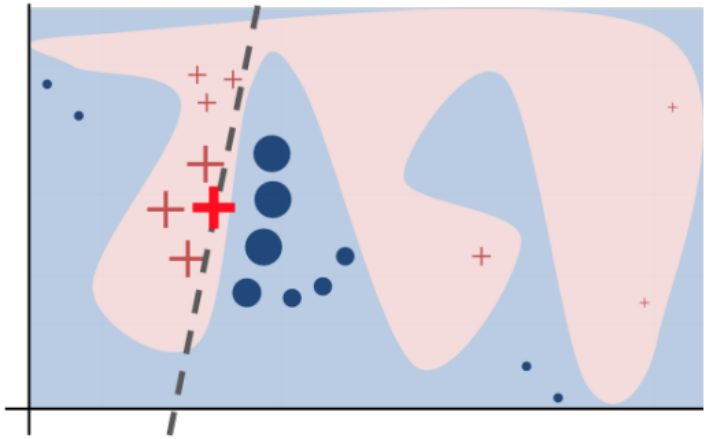
Explaining individual predictions



Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should I trust you?" Explaining the predictions of any classifier." *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016.

Explainable AI – Interpretability

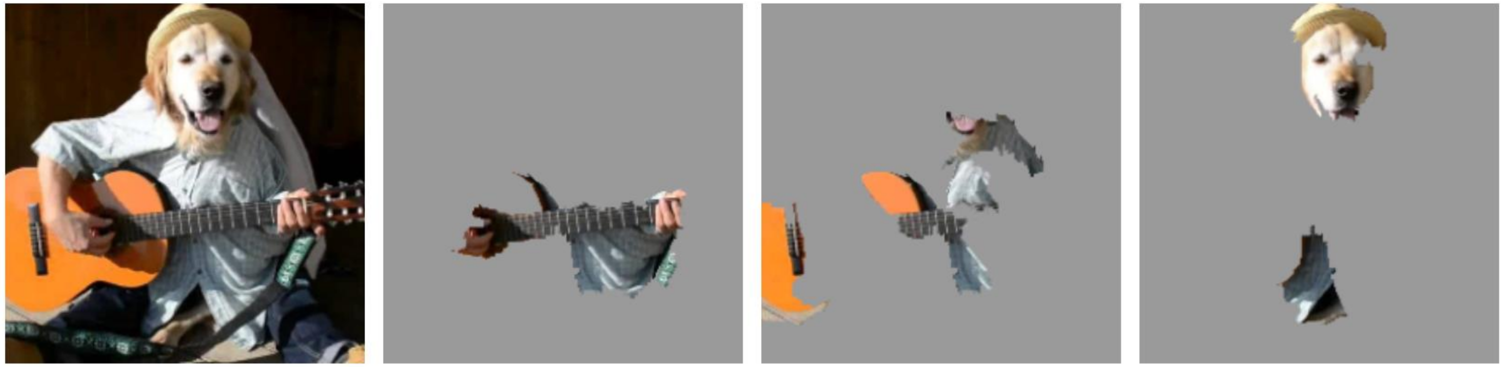
LIME: Local Interpretable Model-agnostic Explanations



- Decision of the black-box model: cannot be approximated well by a compact functional.
- LIME: samples instances, gets predictions, and weighs them by the proximity to the instance being explained (by size).
- Red across: the instance being explained.
- The dashed line: the learned explanation that is locally (but not globally) faithful.

Explainable AI – Interpretability

LIME: Local Interpretable Model-agnostic Explanations



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

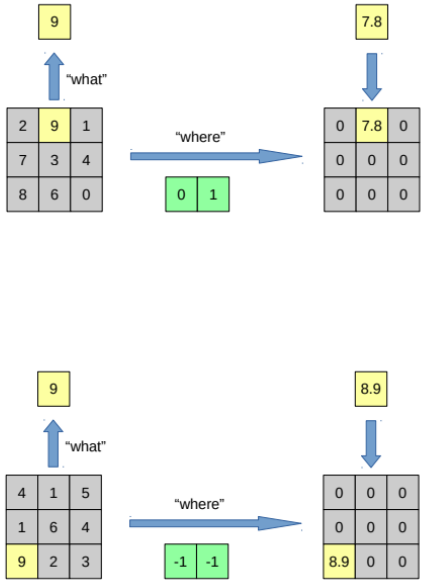
Figure 4: Explaining an image classification prediction made by Google's Inception neural network. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)

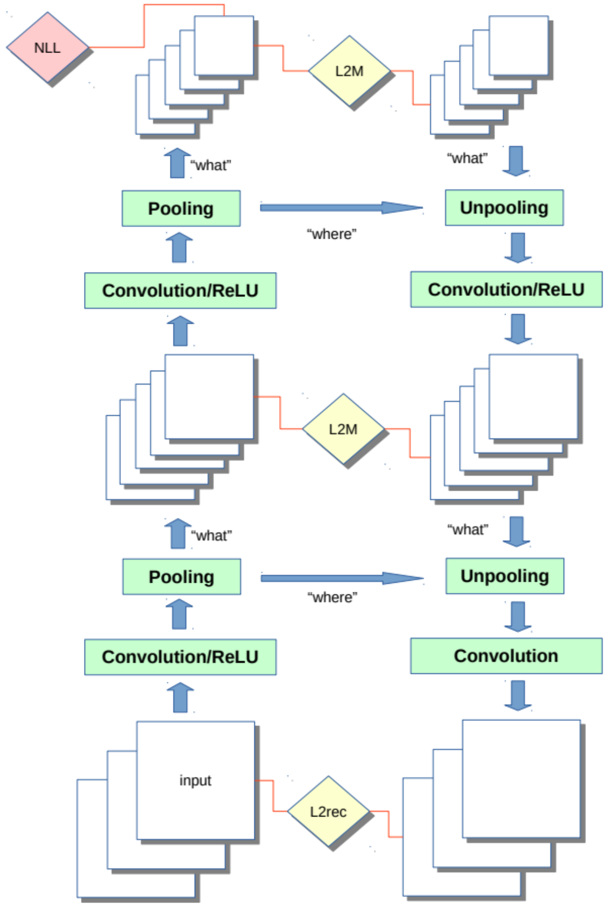
(a) Husky classified as wolf (b) Explanation

Explainable AI – Interpretability

Stacked What-Where Autoencoders

- "What":
  - fed to the next layer
  - the content with incomplete information about position.
- "Where":
  - fed to the corresponding layer
  - where interesting (dominant) features are located

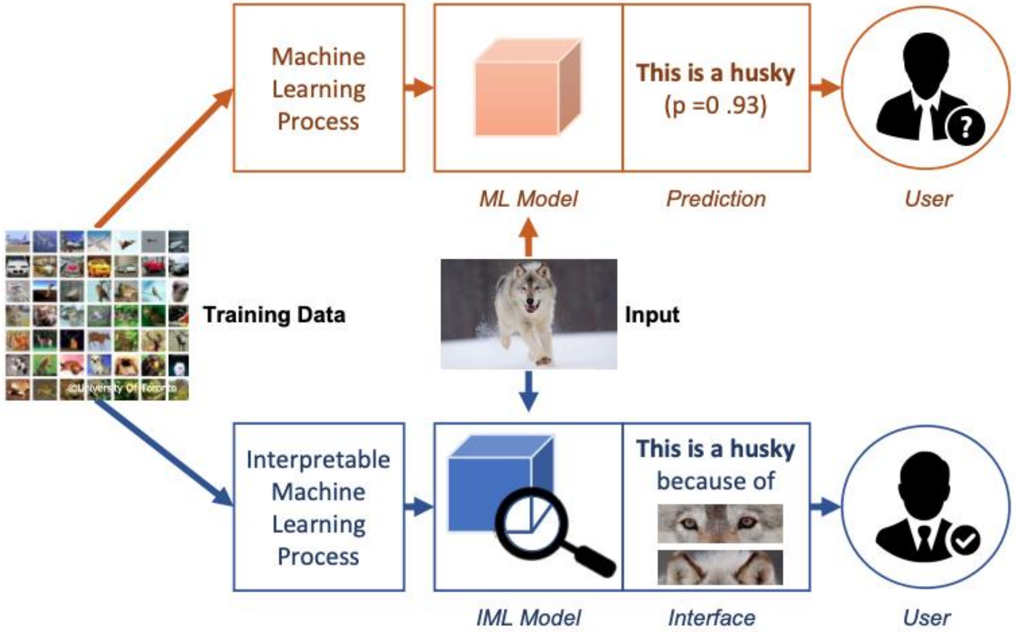


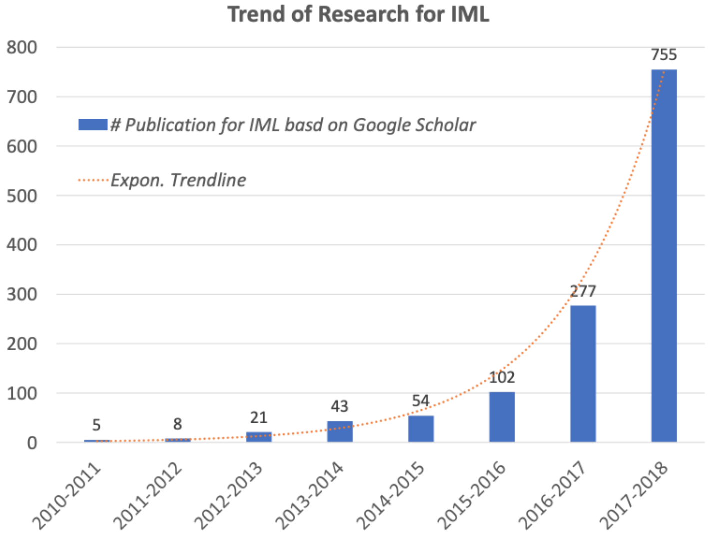


Zhao, Junbo, et al. "Stacked what-where auto-encoders." *arXiv preprint arXiv:1506.02351* (2015).

Explainable AI – Interpretability

Evaluating Explanation Without Ground Truth

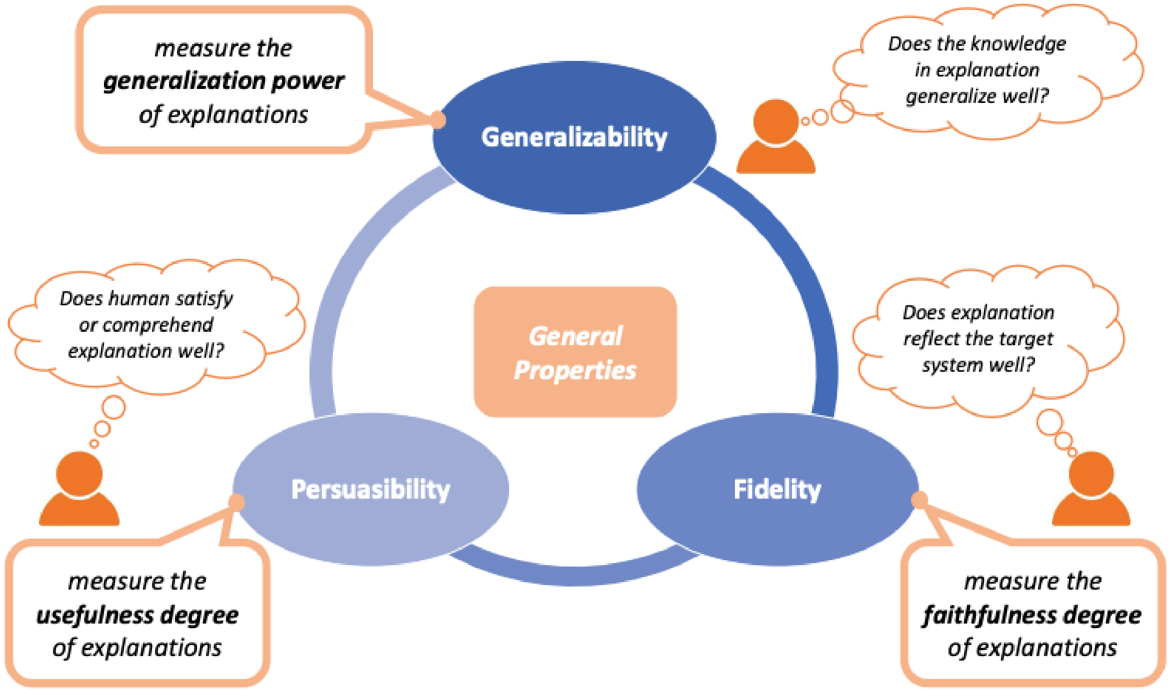




Yang, Fan, Mengnan Du, and Xia Hu. "Evaluating explanation without ground truth in interpretable machine learning." *arXiv preprint arXiv:1907.06831* (2019).

Explainable AI – Interpretability

Evaluating Explanation Without Ground Truth



- Evaluation on Generalizability
- Evaluation on Fidelity
- Evaluation on Persuasibility
- Evaluation on Other Properties

Conversational AI

- Recent conversational interfaces/assistants:
  – Amazon Alexa, Apple’s Siri, Google Assistant

- Goal:
  – Identify where you'll have the greatest conversational impact.
  – Understand your audience.
  – Build complete experiences.

- Data:
  – Content enables conversations.
  – Capture user context.