Intelligence Artificielle Avancée (RCP211)
Robustesse décisionnelle
Explicabilité

Nicolas Thome
Conservatoire National des Arts et Métiers (Cnam)
Laboratoire CEDRIC - équipe Vertigo

Outline

Context

Visualization Methods

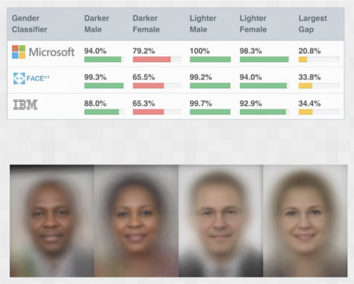
Distillation Methods

Intrinsic Methods

Deep Learning (DL) & Explainability

Need for explainability in machine learning

- Essential for critical systems, *e.g.* autonomous steering, healthcare...
- Legal reasons: responsibility, confidentiality, discriminability of ML systems
- For help to debug /improve algorithms



Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

Joy Buolamwini, Timnit Gebru: Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. FAT 2018: 77-91

Explainability and Interpretability

Oxford Dictionary of English

explanation | ɛksplə'neɪʃ(ə)n |

noun

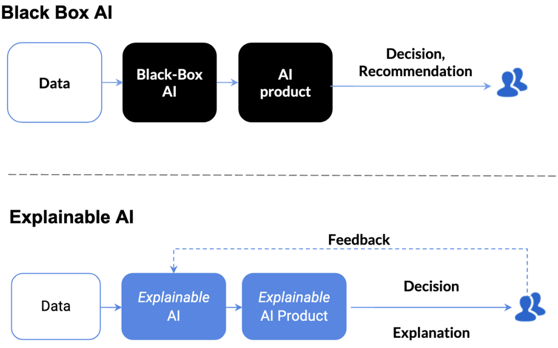
a statement or account that makes something clear: *the birth rate is central to any explanation of population trends*.

interpret | ɪn'təːprɪt |

verb **(interprets, interpreting, interpreted)** *[with object]*

1 explain the meaning of (information or actions): *the evidence is difficult to interpret*.

Black-box vs explainable models



Credit: Lecue et al., Tutorial on XAI. AAAI 2020. https://xaitutorial2020.github.io/

Confusion with Today's AI Black Box

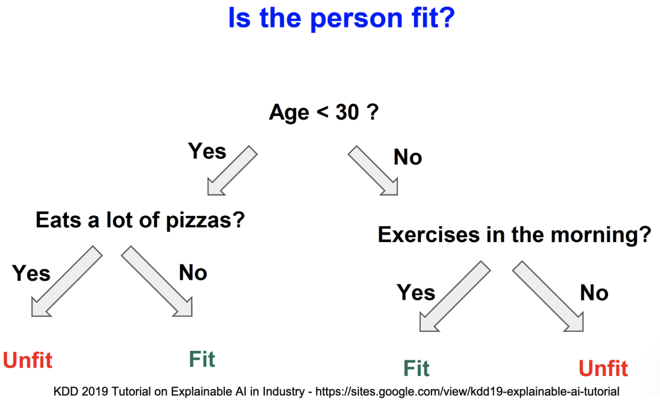
- Why did you do that?
- Why did you not do that?
- When do you succeed or fail?
- How do I correct an error?

Clear & Transparent Predictions

- I understand why
- I understand why not
- I know why you succeed or fail
- I understand, so I trust you

Some ML models naturally explainable

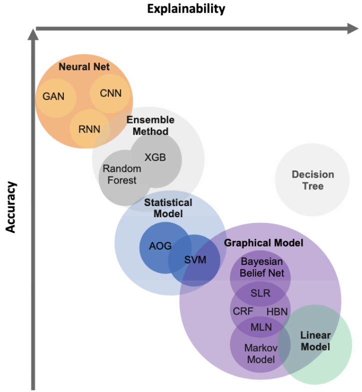
- Decision trees, Lists and Sets and rules
- (Generalized) Linear models, (generalized) additives models, k-NN



Explainability vs accuracy

Learning

- Challenges:
  - Supervised
  - Unsupervised learning
- Approach:
  - Representation Learning
  - Stochastic selection
- Output:
  - **Correlation**
  - **No causation**



Interpretability

- Non-Linear functions
- Polynomial functions
- Quasi-Linear functions

Explainability in different data types

Table of baby-name data
(baby-2010.csv)

name	rank	gender	year
Jacob	1	boy	2010
Isabella	1	girl	2010
Ethan	2	boy	2010
Sophia	2	girl	2010
Michael	3	boy	2010

Field names

One row (4 fields)

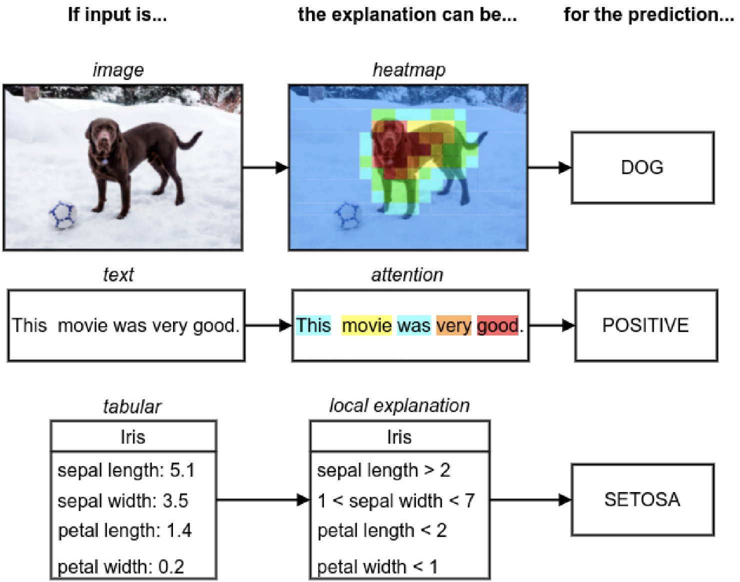
2000 rows all told

Tabular

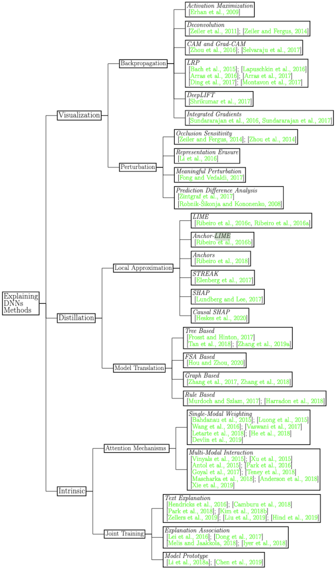
Images

Text

Explainability in different data types



- **Visualization:** for data with local info (text, audio, images)
- **Distillation:** approximate non X-AI models with explainable one
- **Intrinsic:** make the model explicitly explainable



[Xie et al., 2020]

Outline

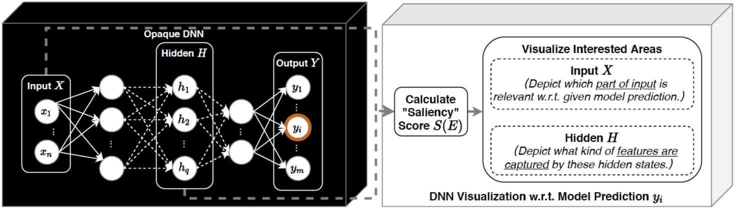
Context

Visualization Methods

Distillation Methods

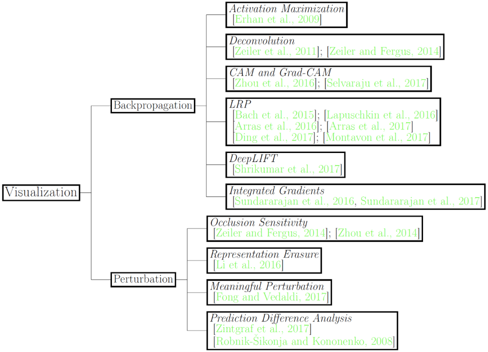
Intrinsic Methods

Explainability: visualization methods



Idea: Saliency $S(E)$ of prediction (output) wrt intermediate features

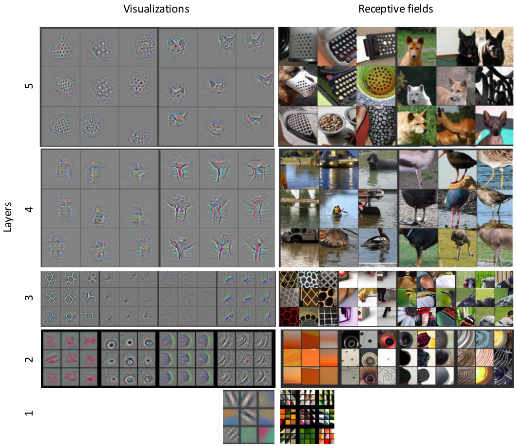
- Importance of input X wrt output Y
- Importance of latent representation H wrt output Y



- Activation maximization [Erhan et al., 2009]:

$$X^* = \arg\max_X h_{i,j}(X, \theta)$$

- **For each latent representation** $h_{i,j}$ (layer j, feaure i): explains which input X gives highest activation
- θ neural network parameters fixed

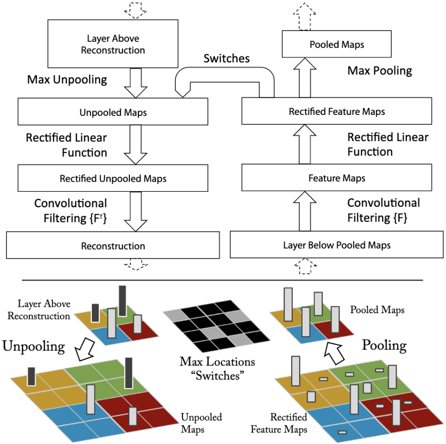


Explainability: Back-prop Visualization

- Deconvolution [Zeiler and Fergus, 2014]
  - From a ConvNet, build a DeconvNet : feature maps → input

$$\boldsymbol{A}^{\ell}, s^{\ell} = \texttt{maxpool}\left(\texttt{ReLU}\left(\boldsymbol{A}^{\ell-1} * \boldsymbol{K}^{\ell} + \boldsymbol{b}^{\ell}\right)\right)$$

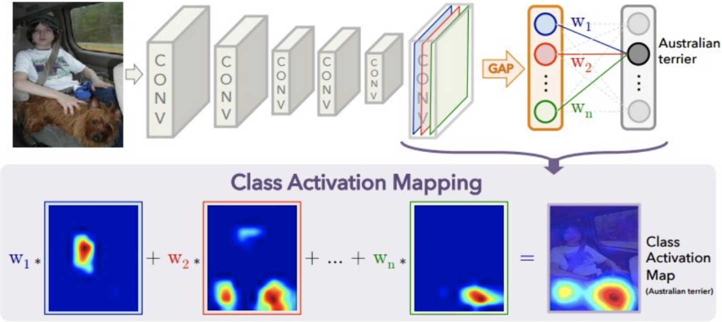
$$\boldsymbol{A}^{\ell-1} = \texttt{unpool}\left(\texttt{ReLU}\left(\left(\boldsymbol{A}^{\ell} - \boldsymbol{b}^{\ell}\right) * \boldsymbol{K}^{\ell T}\right), s^{\ell}\right)$$



- Unpool: keep the max switch, zeros other activations
- Deconv, aka "transposed convolution"

Explainability: Back-prop Visualization

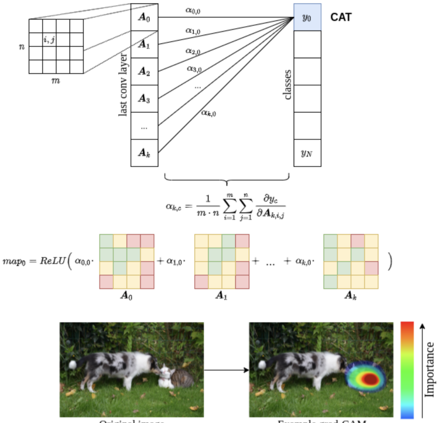
- Class Activation Maps (CAM) [Zhou et al., 2016]
  - ConvNet with Global Average pooling (GAP) layer
  - Revert linear GAP and class projection, upsampling ⇒ class importance in each input pixel



Explainability: Back-prop Visualization

- Grad-CAM [Selvaraju et al., 2017]: extends CAM without GAP

- Each feature map A^k in final conv layer
- Compute gradient of each logit class y_c wrt $A^k_{i,j}$: $\alpha_{k,c} = \frac{\partial y_c}{\partial A^k_{i,j}}$
- Weighted avg of class for each map A^k (with Relu)
- Upsample (bilinear) to get initial image size

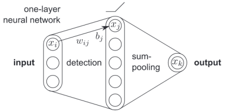


Original image
Example grad-CAM

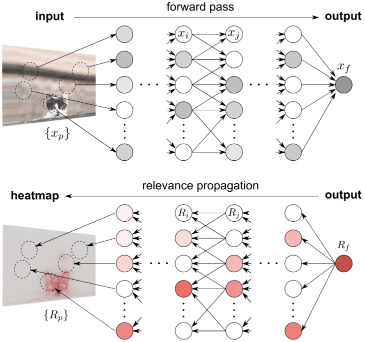
Explainability: Back-prop Visualization

- **Layer-Wise Relevance Propagation (LRP)** [Montavon et al., 2017]
- **Relevance rather than sensitivity**
- Deep Taylor expansion: back-propagate relevance output → inputs

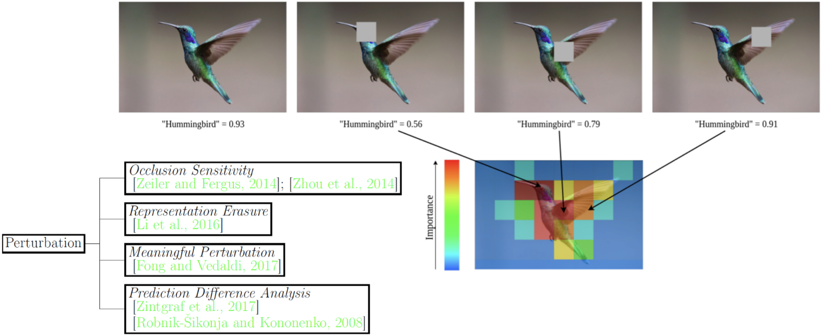
- $R_j = \frac{\partial R_k}{\partial x_j} \cdot (x_j - \tilde{x}_j)$, $\tilde{x}$ root



- ⊕ Higher-resolution visualization than CAM-like
- ⊖ needs the root $\tilde{x}$ definition



- **Altering / removing input feature** ⇒ difference in network output
- **Occlusion sensitivity**: gray patch + Deconv [Zeiler and Fergus, 2014]
- **Variations:** information removal strategy, the size of the patch, how the patches are sampled.
- **Representation erasure:** for NLP



Outline

Context

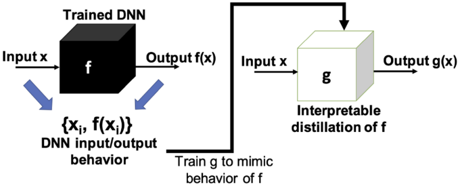
Visualization Methods

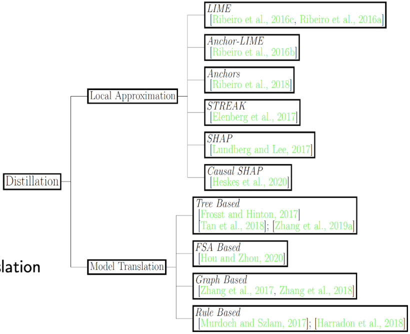
Distillation Methods

Intrinsic Methods

Distillation methods

- Complex black box model f, simpler explainable one: $g(x) \approx f(x)$
- Perfs of f not necessarily below g

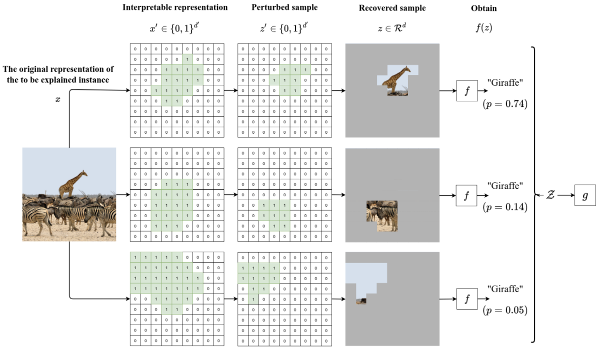




- **Global approximation:** for the whole distribution, *a.k.a.* translation
- **Local approximation:** for a subset of input data

Local approximation in distillation: LIME

- **Local Interpretable Model-Agnostic Explanations (LIME)** [Ribeiro et al., 2016]. For an image x:
  - **Decompose** x **into** d **super-pixels SP** (small, homogeneous patches)
  - **Generate** N **perturbed images** $(x_1, ..., x_N)$ by sampling SPs, Bernouilli ($p = 0.5$) $\Rightarrow Z := (z_1, ..., z_N)$, where $z_i \in \mathbb{R}^d$ - replace *e.g.* with mean for SPs turned off
  - **Compute** $y_i = f(x_i)$ $\forall i \Rightarrow Y$ (f black box) and $\pi_i = dist(x, x_i)$ (some distance)
  - **Fit a ridge regression model:** $Y \approx Z\beta$ with weights π_i



Local approximation in distillation: LIME

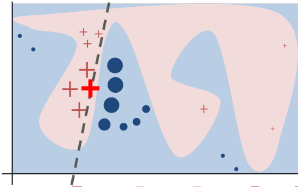
- **Fit a ridge regression model:** $Y \approx Z\boldsymbol{\beta}$ with weights π_i
- $\boldsymbol{\beta} = (\beta_0, ..., \beta_d)$: importance of each SP
  - Importance of SPs ~ perturbation approches **BUT:** interpretable model supposed to be valid locally (close to x)

Credit: M. Chaves, D. Garreau

- LIME not limited to linear regression model, *e.g.* generalized formulation:

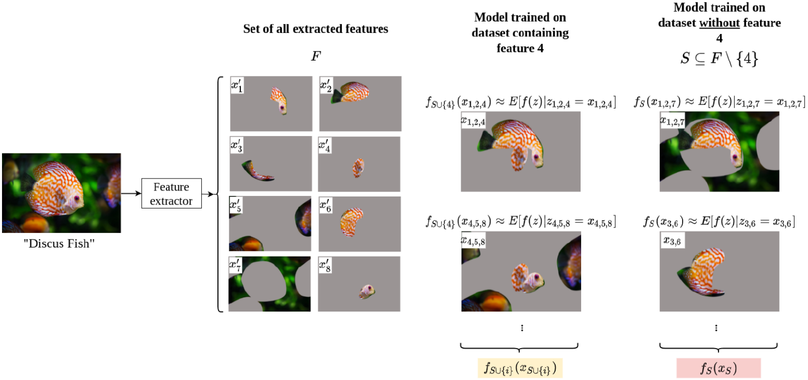
$$g^* = \arg\min_g \mathcal{L}(f, g, \pi_{x'}) + \Omega(g)$$

- $\Omega(g)$ controls models complexity ~ ℓ_2 in ridge



Local approximation in distillation: SHAP

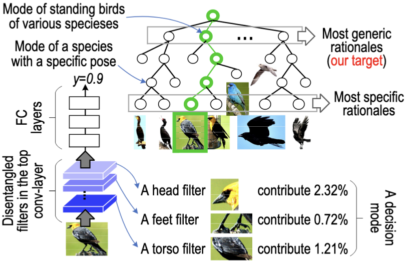
- Shapley Additive Explanations (SHAP) [Lundberg and Lee, 2017]
- Based on additive feature attribution methods ~ linear LIME
- Grounded in a game-theoretical perspective: contribution of adding a feature vs measuring its removal in perturbation approaches

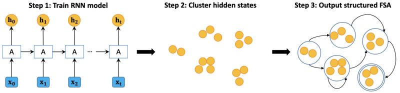


Translation methods in distillation

Use global approximation of a black box model by an explainable one

- **Decision tree/graphs** [Zhang et al., 2019]:
  1. Mine semantic patterns (objects, parts, "decision modes") as interpretable components for the tree
  2. Learn the decision tree to mimic a complex DNN model
- **Finite state automaton (FSA)** [Hou and Zhou, 2020] to approximate RNNs in binary classifications
- Causal classifiers





Outline

Context

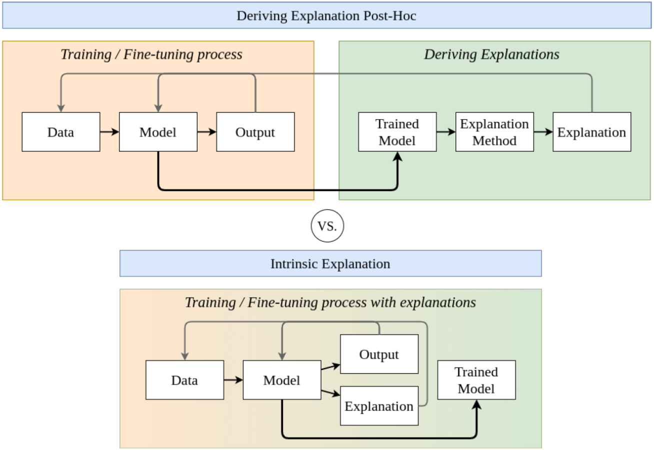
Visualization Methods

Distillation Methods

Intrinsic Methods

Intrinsic Methods

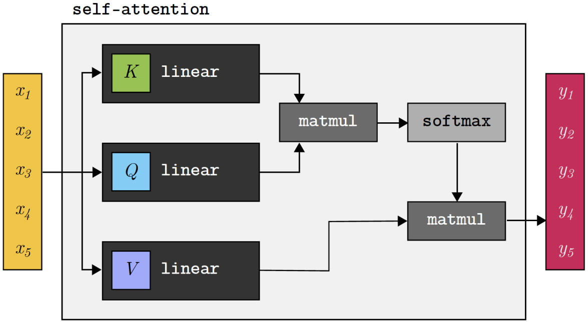
- **Previous methods: post-hoc explainability**: black-box → explainable model
- **Intrinsic methods: build models explainable by design**

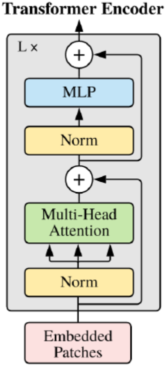


Credit: [Xie et al., 2020]

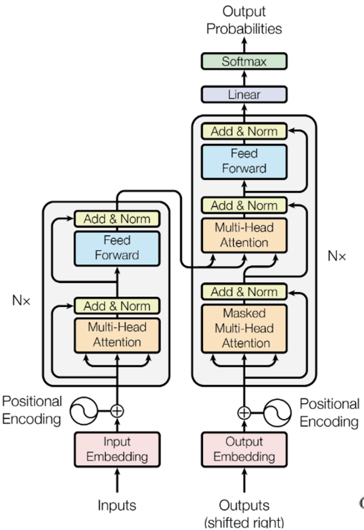
Intrinsic Methods: Attention models

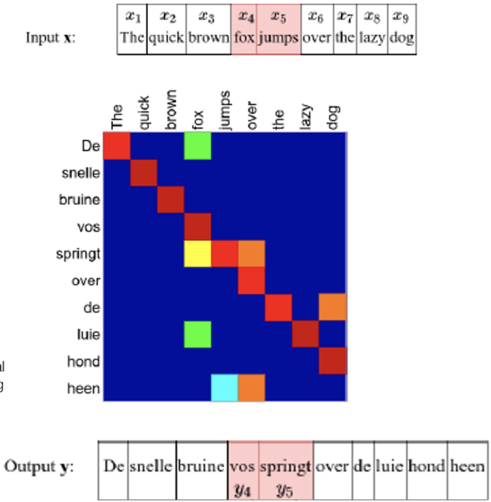
- **Self-attention**: compute similarity matrix between input "tokens"
  - Self-attention (transformers), "attention is all you need" [Vaswani et al., 2017]
- More details in RCP217





Attention for explainability: translation in NLP





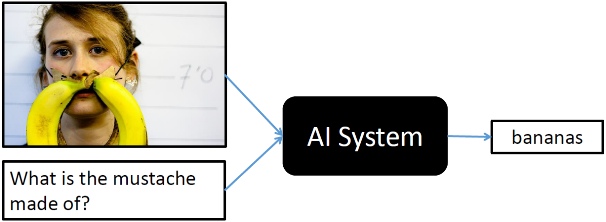
Multi-modal Attention

- Alignment/fusion across different feature space, *e.g.* text/image
  - Image captioning, Visual Question Answering (VQA)

How many slices of pizza are there?
Is this a vegetarian pizza?

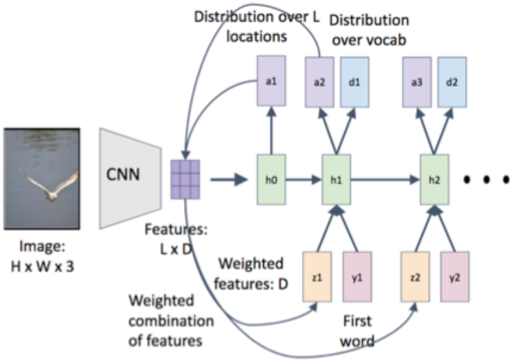
What color are her eyes?
What is the mustache made of?

- Multi-modal attention between words and image regions



Multi-modal attention

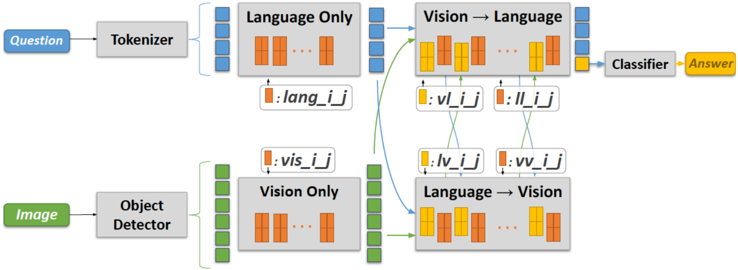
- **Image captioning**: Show, Attend and Tell (SAT) [Xu et al., 2015]
- a_i region feature $(L \times D)$, $(a_i, h_{t-1}) \Rightarrow$ MLP $e_{t,i} = f_{att}(a_i, h_{t-1})$
  - + soft-max : $\alpha_{t,i} = softmax(e_{t,i})$
- LSTM $\hat{z}_t$ representation: context vector: $\hat{z}_t = \phi(a_i, \alpha_{t,i}) = \sum_i \alpha_{t,i} a_i$



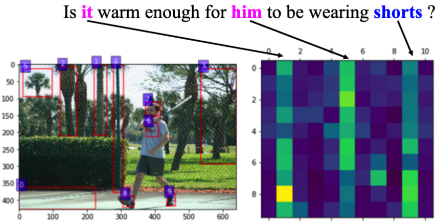
Based on CS231n by Fei-Fei Li, Justin Johnson & Serena Yeung

Multi-modal attention

- **VQA**: multi-modal transformers *e.g.* LX-MERT

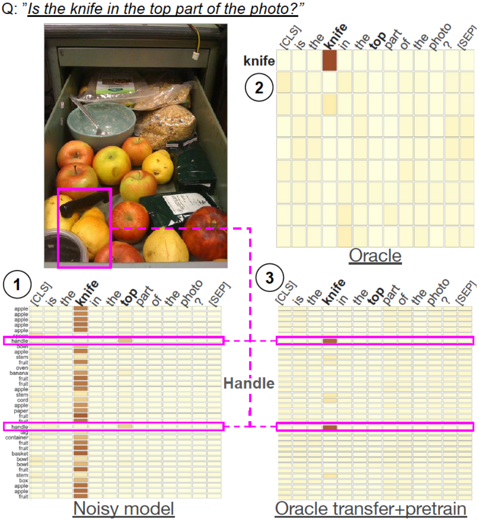


  - Cross-attention between words and image regions
  - Re-embedding of text inputs with visual content, and vice-versa



Cross attention for explainability [Jaunet et al., 2021]

- Automatically Select "peaky" attention maps, *i.e.* where the attention is sparse

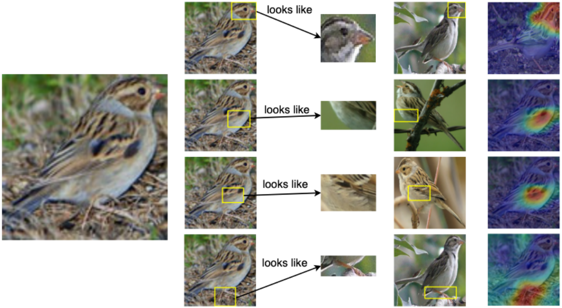


- Use for explainability
- Model failure detection

Intrinsic Methods

Prototypes

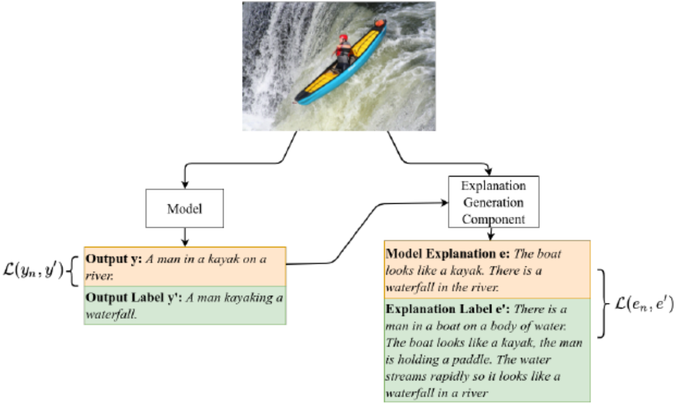
- Similarity between input and each prototype observation in the dataset
  - ProtoPNet [Chen et al., 2019]



Leftmost: a test image of a clay-colored sparrow
Second column: same test image, each with a bounding box generated by our model -- the content within the bounding box is considered by our model to look similar to the prototypical part (same row, third column) learned by our algorithm
Third column: prototypical parts learned by our algorithm
Fourth column: source images of the prototypical parts in the third column
Rightmost column: activation maps indicating how similar each prototypical part resembles part of the test bird

Adding more supervision to drive explainability

- Explanation Association: add human-understandable concepts to inputs
- Text Explanation



$$argmin_\theta \frac{1}{N} \sum_{n=1}^{N} \alpha \mathcal{L}(y_n, y') + \mathcal{L}(e_n, e')$$

References I

[Chen et al., 2019] Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., and Su, J. K. (2019).
This looks like that: Deep learning for interpretable image recognition.
In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

[Erhan et al., 2009] Erhan, D., Bengio, Y., Courville, A., and Vincent, P. (2009).
Visualizing higher-layer features of a deep network.
Technical Report, Univeristé de Montréal.

[Hou and Zhou, 2020] Hou, B. and Zhou, Z. (2020).
Learning with interpretable structure from gated RNN.
IEEE Trans. Neural Networks Learn. Syst., 31(7):2267–2279.

[Jaunet et al., 2021] Jaunet, T., Kervadec, C., Vuillemot, R., Antipov, G., Baccouche, M., and Wolf, C. (2021).
VisQA: X-raying Vision and Language Reasoning in Transformers.
IEEE Transactions on Visualization and Computer Graphics.

[Lundberg and Lee, 2017] Lundberg, S. M. and Lee, S.-I. (2017).
A unified approach to interpreting model predictions.
In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

[Montavon et al., 2017] Montavon, G., Lapuschkin, S., Binder, A., Samek, W., and Müller, K.-R. (2017).
Explaining nonlinear classification decisions with deep taylor decomposition.
Pattern Recognition, 65:211–222.

[Ribeiro et al., 2016] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016).
"why should I trust you?": Explaining the predictions of any classifier.
In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144.

[Selvaraju et al., 2017] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017).
Grad-cam: Visual explanations from deep networks via gradient-based localization.
In *ICCV*, pages 618–626. IEEE Computer Society.

References II

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017).
Attention is all you need.
In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc.

[Xie et al., 2020] Xie, N., Ras, G., van Gerven, M., and Doran, D. (2020).
Explainable deep learning: A field guide for the uninitiated.
CoRR, abs/2004.14545.

[Xu et al., 2015] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. (2015).
Show, attend and tell: Neural image caption generation with visual attention.
In Blei, D. and Bach, F., editors, *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 2048–2057. JMLR Workshop and Conference Proceedings.

[Zeiler and Fergus, 2014] Zeiler, M. D. and Fergus, R. (2014).
Visualizing and understanding convolutional networks.
In *ECCV*.

[Zhang et al., 2019] Zhang, Q., Yang, Y., Ma, H., and Wu, Y. N. (2019).
Interpreting cnns via decision trees.
In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6254–6263.

[Zhou et al., 2016] Zhou, B., Khosla, A., A., L., Oliva, A., and Torralba, A. (2016).
Learning Deep Features for Discriminative Localization.
CVPR.