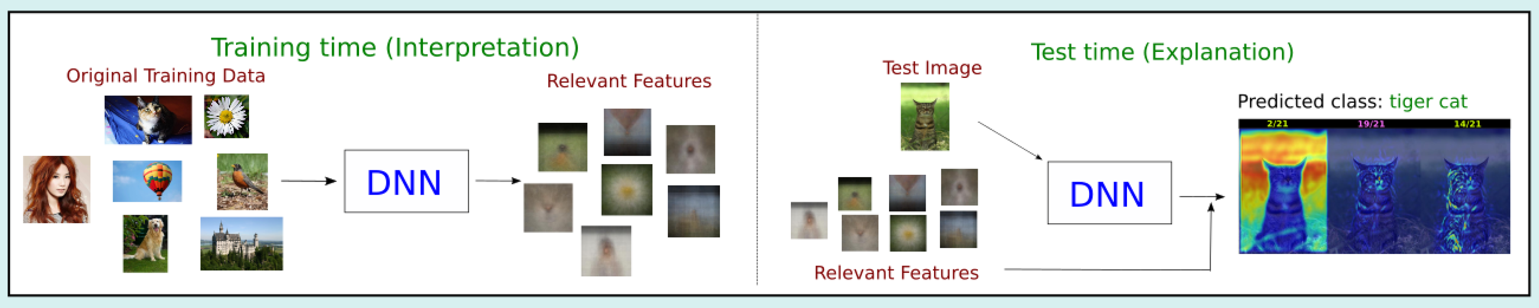
What if A.I. systems could explain themselves better?

(Improving feedback capabilities of deep models)

José Oramas M. , Kaili Wang , Tinne Tuytelaars

Abstract

- Focus on provinding richer visually-descriptive predictions.
- Interpretation: visualize a small set of internal network features relevant for the classes of interest.
- Explanation: extend the model prediction with visualizations highlighting the response of the identified relevant features.
- Design an objective evaluation protocol for visual explanations through a controlled dataset.



Motivation

- Methods that provide their train of thought as part of their output are more likely to be trusted and adopted by end users.
- Current methods for model interpretation are exhaustive or prone to subjectivity and noise. [1]
- Current evaluation protocols for visual explanation rely on user studies or proxy tasks. [5]

Algorithm

1) Identify relevant features for the classes of interest.

2) Generate model interpretation visualizations

3) Enrich model prediction

1) Identify relevant features for the classes of interest

Given a pre-trained model F for C classes of interest.
- Pass every image through the model.
- Collect internal activation response $x_i \in \mathbb{R}^m$ for every image i.
- Define the data matrix $X \in \mathbb{R}^{m \times N}$
- Define the binary label matrix $L = [l_1, l_2, ..., l_N]$ with $L \in \mathbb{R}^{C \times N}$
- Identify the subset W^* of relevant features for every class j

$$W^* = argmin_W \; ||X^T W - L^T||_F^2$$

$$subject\ to: \; ||w_j||_1 \leq \mu \;, \; \forall_j = 1, ..., C$$

(μ-LASSO problem)

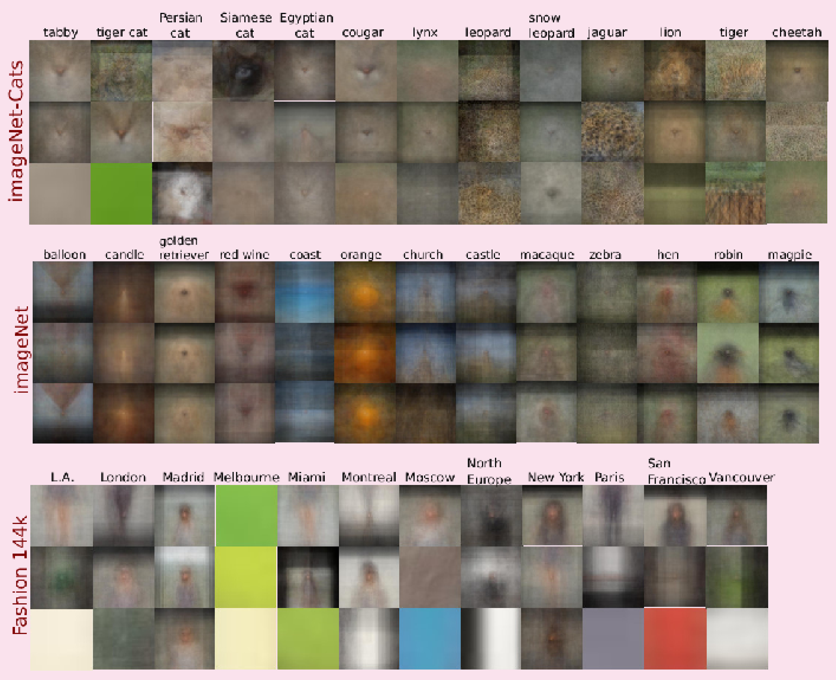
where:

$W = [w_1, w_2, ..., w_C]$ and μ : sparsity parameter

2) Generate model interpretation visualizations

For every identified relevant feature
- Select the top (100) images with highest response.
- Crop each selected image using the receptive field of the feature.
  (centered on the pixel of highest response)
- Scale all the image crops to a common size.
- visual interpretation image --> crop the pixel-wise average.

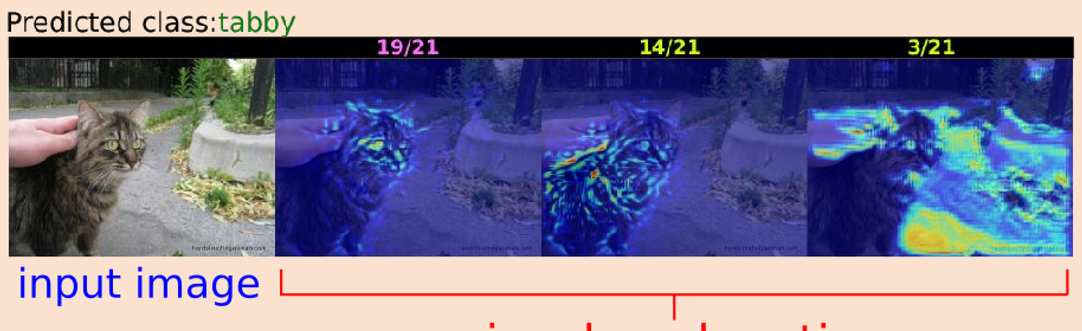
Some examples



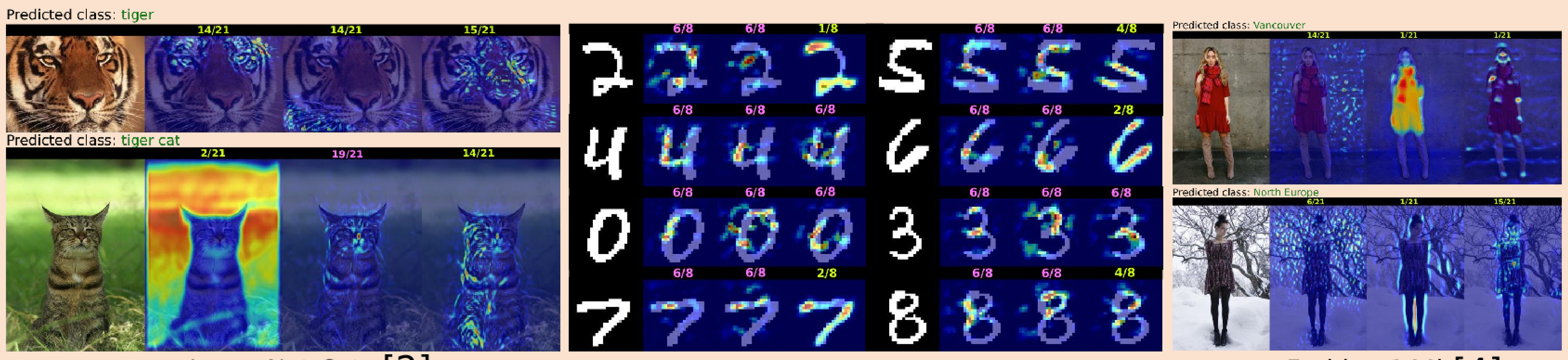
Observations
- Descriptive features of the classes of interest are identified.
- Identified features either focus on object features, e.g. color, shape, etc., or on the context, e.g. vegetation, buildings, etc.

3) Enrich model prediction

- Given a model prediction $\hat{j} = F(I)$ for the input image I
- Compute the filter-wise reponse x_i during the forward pas
- Compute the response $r_i^j = (w_j \circ x_i)$ using the Hadamard product $\circ$
- Select the features with strongest contribution to prediction
  (i.e. layer/filter pairs (p^*, q^*) with maximum response in $r_i^{\hat{j}}$)
- Generate a heatmap visualization of each feature.
  (e.g. via deconveNet+ guided backprogation[3])

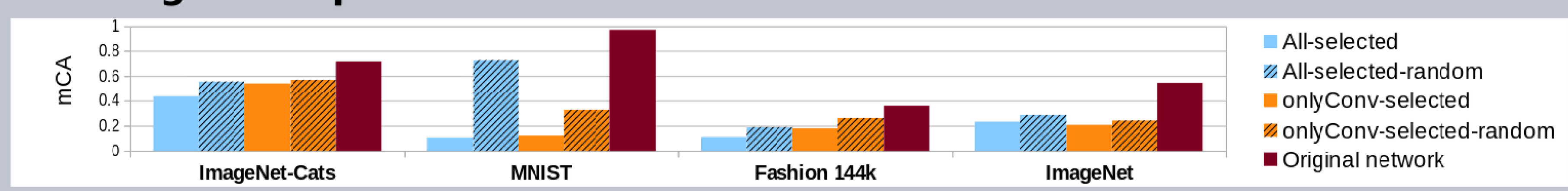


Some generated visual explanations



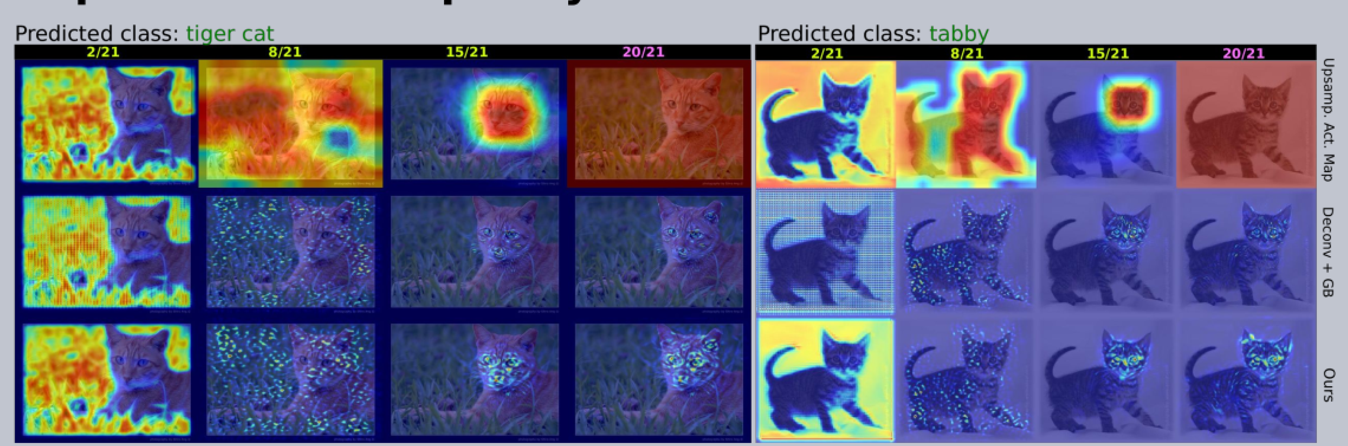
Evaluation

Measuring the importance of the identified relevant features



- Changes in mean classification accuracy (mCA) as the identified relevant filters are ablated
- Ablating filters with higher relevance produces a bigger drop in performance.

Impact on visual quality



Observations

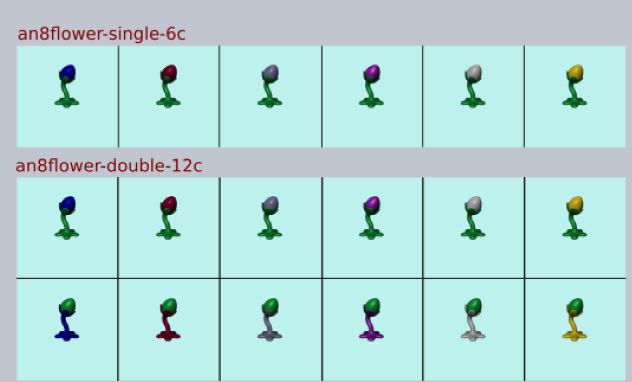
At lower layers: attenuates grid-like artifacts from deconvNet methods.

At higher layers: provides more precise visualizations than upsampled activation maps.

A novel protocol for the objective evaluation of visual explanations

- Focus on a problem where the discriminative feature between classes can be controlled.
- Design a dataset where the regions to be highlighted by the explanation are pre-defined.

Proposed dataset



Protocol
- Generate GT-masks for the discriminative regions.
- Threshold the visual explanation heatmaps.
- Measure pixel-level interesection over union (IoU).
- Compute mean performance over different heatmap threshold values.

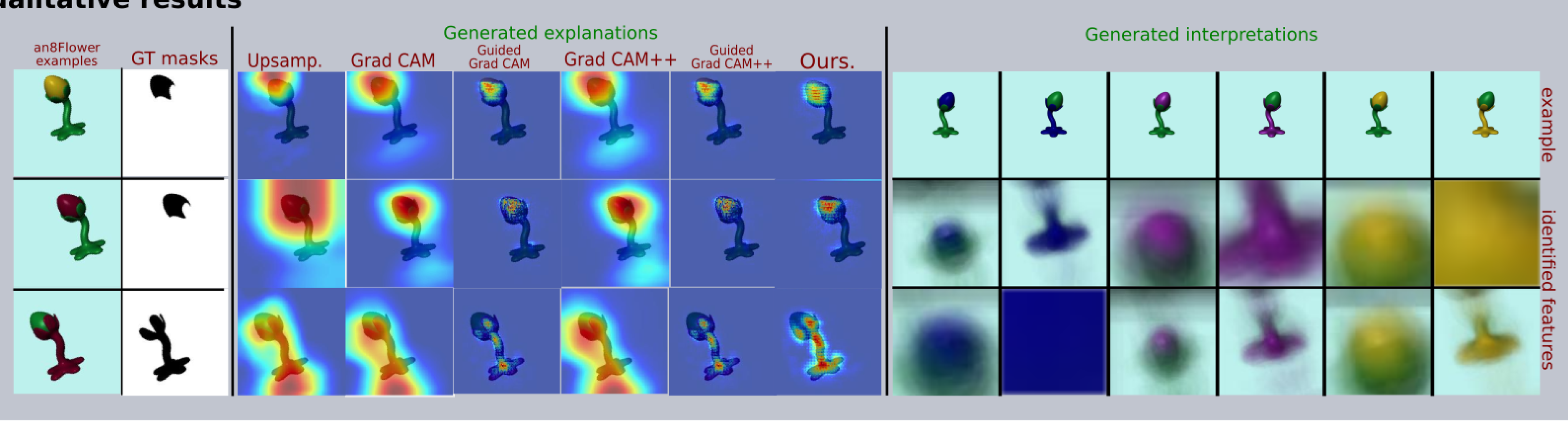
Observations
- Our method effectively identifies the pre-defined discriminative features.
- Our explanations highlight these features and have a better balance between coverage and level of detail

Quantitative results

Method	single-6c
Upsam. Act.	16.8±2.63
Deconv+GB, Springenberg et al. (2015)	21.3±0.77
Grad-CAM, Das et al. (2016)	17.5±0.25
Guided Grad-CAM, Das et al. (2016)	19.9±0.61
Grad-CAM++, Chattopadhyay et al. (2018)	15.6±0.57
Guided Grad-CAM++, Chattopadhyay et al. (2018)	19.6±0.65
Ours	**22.5±0.82**

Method	double-12c
Upsam. Act.	16.1±1.30
Deconv+GB, Springenberg et al. (2015)	21.9±0.72
Grad-CAM, Das et al. (2016)	14.8±0.16
Guided Grad-CAM, Das et al. (2016)	19.4±0.34
Grad-CAM++, Chattopadhyay et al. (2018)	14.6±0.12
Guided Grad-CAM++, Chattopadhyay et al. (2018)	19.7±0.27
Ours	**23.2±0.60**

Qualitative results



References

[1] Bau et al., CVPR 2017.
[2] Russakovsky et al., IJCV 2015.
[3] Springenberg et al., ICLR 2015.
[4] Simo-Serra et al., CVPR 2015.
[5] Zhou et al., ICLR 2015.

Conclusions

- The proposed method enriches the prediction of a deep neural network by indicating the visual features that contributed to such prediction.
- Our method effectively identifies relevant features encoded by the model and allows interpretation of such features through average feature visualizations.
- The proposed evaluation protocol allows for objective evaluation of methods for visual explanation of deep models.