Zeynep Akata

Representing and Explaining Novel Concepts with Minimal Supervision

Introduction

Clearly explaining a rationale for a classification decision to an end-user can be as important as the decision itself. As decision makers, humans can justify their decisions with natural language and point to the evidence in the visual world which led to their decisions. In contrast, artificially intelligent systems are frequently seen as opaque and are unable to explain their decisions. This is particularly concerning as ultimately such systems fail in building trust with human users.

Explanations are valuable because they enable users to adapt themselves to the situations that are about to arise while allowing users to attain a stable environment and have the possibility to control it. Explanations in the medical domain can help patients identify and monitor the abnormal behaviour of their ailment. In the domain of self-driving vehicles they can warn the user of some critical state and collaborate with her to prevent a wrong decision. In the domain of satellite imagery, an explanatory monitoring system justifying the evidence of a future hurricane can save millions of lives. Hence, a learning machine that a user can trust and easily operate needs to be fashioned with the ability of explanation.

While deep neural networks lead to impressive successes, e.g. they can now reliably identify 1000 object classes [11], argue about their interactions through natural language [3], answer questions about their attributes [7] through interactive dialogues, integrated interpretability is still in its early stages. In other words, we do not know why these deep learning based visual classification systems work when they are accurate and why they do not work when they make mistakes. Enabling such transparency requires the interplay of different modalities such as images and text, whereas current deep networks are designed as a combination of different tools each optimising a different learning objective with

Resume:

In Explainable Machine Learning, we study the challenging problem of large-scale learning with vision and language. We use image labels, i.e. zebra, horse, etc. when they are available or side information in the form of attributes, i.e furry, striped, etc. when image labels are not provided. Combining vision and language in a single framework we aim to learn robust representations generalizable across different tasks. As user acceptance is likely to benefit from easy-to-interpret visual and textual rationales allowing them to understand what triggered a particular behavior, we aim to generate textual justifications for model decisions in a two-step framework. In the first stage, we use visual (spatial) attention to train a convolutional network to decisions, e.g. steering angle while driving. The attention model identifies image regions that potentially influence the network's output. In the second stage, we use a video-to-text language model to produce textual rationales that justify the model's decision. The explanation generator uses a spatiotemporal attention mechanism encouraged to match the attention of the decision maker.

”A learning machine that a user can trust and easily operate needs to be fashioned with the ability of explanation.“

Zeynep Akata

extremely weak and uninterpretable communication channels. However, deep neural networks draw their power from their ability to process large amounts of data in an end-to-end manner through a feedback loop with forward and backward processing. Although interventions on the feedback loop have been implemented by removing neurons and back propagating gradients, a generalizable multi-purpose interpretability is still far from reach.

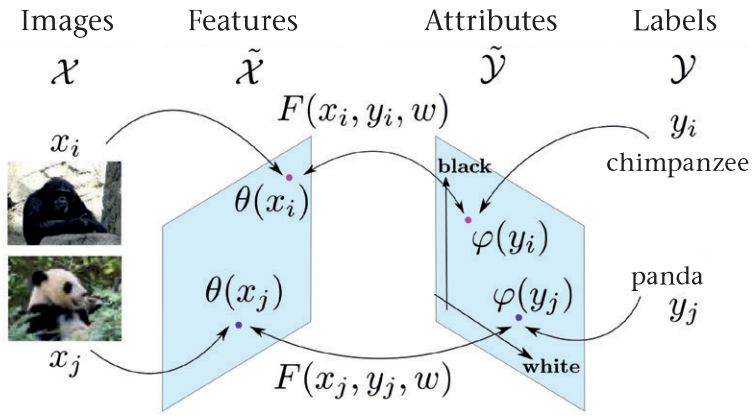
Apart from the lack of an integrated interpretability module, deep neural networks require a large amount of labeled training data to reach reliable conclusions. In particular, they need to be trained for every possible situation using labeled data. For instance, the system needs to observe the drivers behavior at the red light to be able to learn to stop at red light both in a sunny and rainy weather, both in daylight and in night, both in fog and in snow, and so on. This causes a signicant overhead in labelling every possible situation. Hence, our aim is to build an explainable machine learning system that can learn the meaning of red light and use this knowledge to identify many other related situations, e.g. although red light may look different in darkness vs daylight, the most important aspect in such a situation is to identify that the vehicle needs to stop. In other words, we would like to transfer the explainable behaviour of a decision maker to novel situations.

In summary, we propose an end-to-end trainable decision maker operating in sparse data regime with an integrated interpretability module. Our main research direction to build such a system is two folds: learning representations with weak supervision and generating multimodal explanations of classification decisions.

Explainability with Limited Supervision

The image classification problem has been redened by the emergence of large scale datasets such as ImageNet. Since deep learning reaches above human accuracy in such datasets, the attention of the computer vision community has been drawn to Convolutional Neural Networks (CNN). Training CNNs requires massive amounts of labeled data; but, in fine-grained image collections, where the categories are visually very similar, the data population decreases significantly. We are interested in the most extreme case of learning with a limited amount of labeled data, zero-shot learning, in which no labeled data is available for some classes.

When it comes to applying a previously trained deep learning model to a novel task, the main challenge is how to transfer knowledge from one task to the other without



having to label many instances of the novel situation and without having to retrain the model from scratch. Transferring knowledge to novel tasks requires learning with limited supervision. Achieving better-than-chance performance in these cases requires structure in the space of decisions, i.e. different decisions must contain similar characteristics such that they can be associated. We consider the image classification problem where the task is to annotate a given image with one (or multiple) class label(s) describing the visual properties of its most prominent object.

In this research direction, our aim is to build semantically interpretable features of objects that are shared among different categories and specific to a certain class. Semantic features such as attributes relate different situations through well-known and shared characteristics of objects, e.g. a red head is shared among "cardinals" and "red-headed woodpeckers". They also isolate the specific situation from the decision by providing a modular representation of the situation, e.g. a cardinal has a black patch on its face whereas a red-headed woodpecker does not.

Label Embeddings for Zero-Shot Learning. Much work in computer vision has been devoted to image embedding: how to extract suitable features from an image. We focus on label embedding: how to embed class labels in a Euclidean space. We use side information such as attributes for the label embedding and measure the compatibility

Figure 1:
Much work in computer vision has been devoted to image embedding (left): how to extract suitable features from an image. We focus on label embedding (right): how to embed class labels in a Euclidean space. We use side information such as attributes for the label embedding and measure the "compatibility" between the embedded inputs and outputs with a function F.

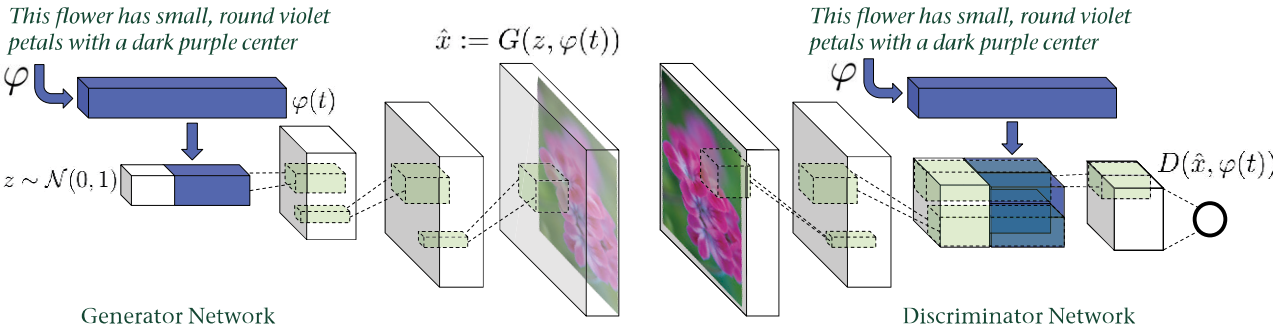


Figure 2: *Our text-conditional convolutional GAN architecture. Text encoding $\phi(t)$ is used by both generator and discriminator. It is projected to a lower-dimensional space and depth concatenated with image feature maps for further stages of convolutional processing.*

between the embedded inputs and outputs with a function F whose parameters are optimized with ranking loss [1, 2]. Attributes associate different images and classes for zero-shot learning. This means that attributes correspond to high-level properties of the objects which are shared across multiple classes, which can be detected by machines and which can be understood by humans. Each class can be represented as a vector of class-attribute associations according to the presence or absence of each attribute for that class. Such class attribute associations are often binary (in our work, we assume that the class-attribute association matrix is provided through annotations by an expert user). As an example, if the classes correspond to animals, possible attributes include has paws, has stripes or is black. For the class zebra, the has paws entry of the attribute vector is zero whereas the has stripes would be one.

To learn novel concepts, a function which measures the "compatibility" between an image x and its class attributes associated with its y can be used. The parameters of this function are learned on a training set of labeled samples to ensure that, given an image, the correct class(es) rank higher than the incorrect ones. Given a test image, labeling consists in searching for the class with the highest compatibility score.

Generative Models for Data Augmentation. An orthogonal approach to zero-shot learning is to augment data by generating artificial images using their textual descriptions. In [10] we develop a novel deep architecture and generative adversarial networks (GAN) formulation to effectively bridge recent advances in text and image modeling, translating visual concepts from characters to pixels. Our approach is to train a deep con-

volutional generative adversarial network (DC-GAN) conditioned on text features encoded by a hybrid character-level convolutional-recurrent neural network. Both the generator network G and the discriminator network D perform feed-forward inference conditioned on the text feature. As shown in the figure, our text-conditional convolutional GAN architecture is composed of a generater and the discriminator. The text encoding is used by both generator and discriminator. It is projected to a lower-dimensions and depth concatenated with image feature maps for further stages of convolutional processing. We demonstrate in [10] that this model can synthesize many plausible visual interpretations of a given text caption where the text caption talks about different visual properties of a bird or a ower image. We showed disentangling of style and content, and bird pose and background transfer from query images onto text descriptions. This work has been the first work on text to image generation which lead a large amount of further research in the recent machine learning literature. On the other hand, due to the level of detail missing in the synthetic images, image features extracted from them do not improve classification accuracy. Discriminative visual features can be extracted from: 1) real images, however in zero-shot learning we do not have access to any real images of unseen classes, 2) synthetic images, however they are not accurate enough to improve image classification performance. We tackle both of these problems and propose a novel attribute conditional feature generating adversarial network formulation in [14], to generate CNN features of unseen classes. This simplifies the task of the generative model and directly optimizes the loss on image features. The main insight of this model is that by feeding additional synthetic CNN features of unseen classes, the learned classifier will also explore the embedding space of unseen classes. Hence, the key to our approach is the ability to generate semantically rich CNN feature distributions conditioned on a class specific semantic vector e.g. attributes, without access to any images of that class. This alleviates the imbalance between seen and unseen classes, as there is no limit to the number of synthetic CNN features that our model can generate. It also allows to directly train an ad-hoc discriminative classifier even for unseen classes.

”An orthogonal approach to zero-shot learning is to augment data by generating articial images using their textual descriptions.“

Zeynep Akata

One disadvantage of the GAN-based loss functions is that they suffer from instability in training. As a methodological improvement, in [12], we train Variational Autoencoders (VAEs) to encode and decode features from different modalities, e.g. images and class attributes, and use the learned latent features to train a zero-shot learning classifier. By explicitly enforcing alignment both

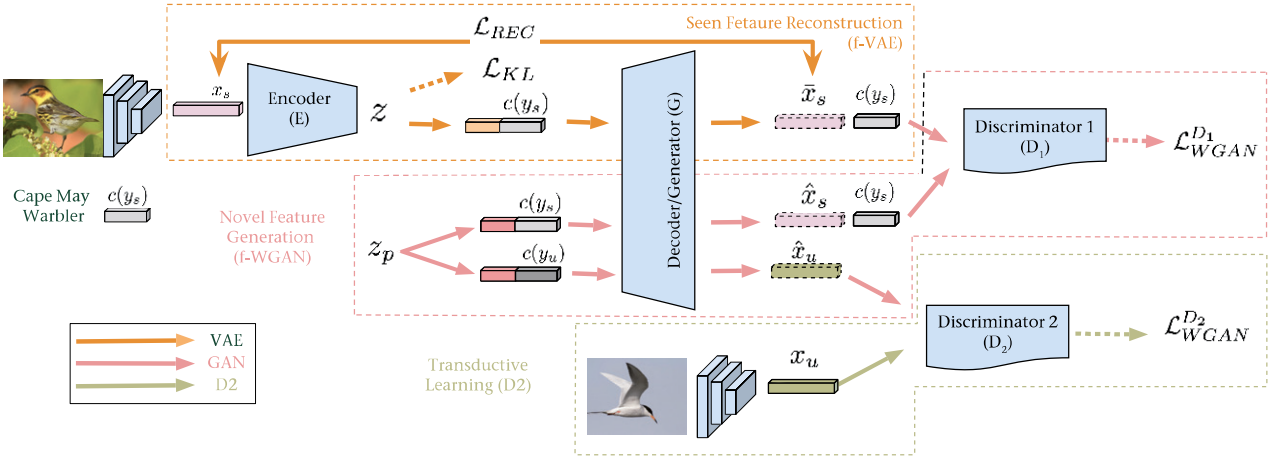


Figure 3: *Our any-shot feature generating network (f-VAEGAN-D2) consist of a feature generating Variational Auto Encoder (f-VAE), a feature generating Wasserstein Generative Adversarial Network (f-WGAN) with a conditional discriminator (D1) and a transductive feature generator with a non-conditional discriminator (D2) that learns from both labeled data of seen classes and unlabeled data of novel classes.*

in the latent features and in the distributions of latent features learned using different modalities, the VAEs enable knowledge transfer to unseen classes without forgetting the previously seen ones. As shown in the figure, our model learns a latent embedding (z) of image features (x) and class embedding (c(y) of labels y) via aligned VAEs optimized with cross-alignment (LCA) and distribution alignment (LDA) objectives, and subsequently trains a classifier on sampled latent features of seen and unseen classes. The main insight of our proposed model is that instead of generating images or image features, we generate low-dimensional latent features and achieve both stable training and state-of-the-art performance. Hence, the key to our approach is the choice of a VAE latent-space, a reconstruction and cross-reconstruction criterion to preserve class-discriminative information in lower dimensions, as well as explicit distribution alignment to encourage domain-agnostic representations.

Finally, we propose a hybrid model [15] combining the strengths of VAE and GANs by assembling them to a conditional feature generating model, that synthesizes CNN image features from class embeddings, i.e. class-level attributes. Thanks to its additional discriminator that distinguishes real and generated features, our model is able to use unlabeled data from previously unseen classes without any condition. The features learned by our model are discriminative in that they boost the performance of any-shot learning as well as being visually and textually interpretable.

The bird has a white underbelly, black fearthers in the wings, a large wingspan, and a white beak.

This bird has distinctive-looking brown and white stripes all over its body, and its brown tail sticks up

This swimming bird has a black crown with a large white strip on its head, and yellow eyes.

This flower has a central white blossom surrounded by large pointed red petals which are veined and leaflike.

Light purple petals with orange and black middle green leaves.

This flower is yellow and orange in color, with petals that are ruffled along the edges.

Datasets and Results. Fine-grained visual categorization is an interesting test-bed for evaluating the generalization ability of the learned representations as the objects are distinguishable only by field experts which increases the annotation effort, therefore attributes or other source of side information is required. We choose Caltech UCSD Birds dataset [13] with 200 classes of bird images annotated with attributes and collect additional text-based annotations [9] that describe the image content specific to the object. After extracting vectorial representations from these sentences, all the vectors that belong to the same class is averaged to build a class-based representation that are then used to learn a compatibility function for label-embedding based zero-shot learning and as a conditioning variable to generate image pixels or image features from scratch.

When we look at the the generated images using these text descriptions. Although the interpolation between sentences "blue bird with black beak" and "red bird with black beak" may not have a semantic meaning individually, the embeddings between these two sentences correspond to semantically meaningful transitions. In addition that using powerful generative models, by describing how the object should look like one can generate an unlimited number of new data points for underrepresented classes. This is remarkable as it is practically impossible to capture visual data of all the objects and situations. On the other hand, generative models provide a means to obtain the missing data instances with a much lower cost.

Figure 4: *Several representative examples of the results from our data collection. The descriptions almost always accurately describe the image, to varying degrees of comprehensiveness. Thus, in some cases multiple captions might be needed to fully disambiguate the species of bird category. However, as we show subsequently, the data is descriptive and large enough to support training high-capacity text models and greatly improve the performance of textbased embeddings for zero-shot learning.*

Coming back to our starting point of classification in low-data regimes, the table on the right demonstrates our per-class top-1 accuracy (the higher the better) on previously seen classes (**s**), previously unseen classes (**u**) and their harmonic mean (**H**) on the bird recognition dataset described above using (**1**) the label-embedding method which learns to associate images and classes with a compatibility function, (**2**) a classifier trained on generated images by text-conditional GAN, (**3**) a classifier trained on generated features by text-conditional WGAN, (**4**) a classifier trained on generated features by text-conditional VAE and (**5**) a classifier trained on generated features by text-conditional VAE-GAN. These results demonstrate that indeed generating images using a GAN may lead to images that miss certain details required for recognition. On the other hand, GAN or VAE-based generative models are able to generate strong and generalizable visual features of previously unseen classes.

Data	**u**	**s**	**H**
(1) Label-Embedding	*23.7*	*62.8*	*34.4*
(2) With generated images (GAN)	*23.8*	*48.5*	*31.9*
(3) With generated features (GAN)	*43.7*	*57.7*	*49.7*
(4) With generated features (VAE)	*63.6*	*51.6*	*52.4*
(5) With generated features (VAE-GAN)	*63.2*	*75.6*	*68.9*

Explaining Neural Network Decisions

We argue that visual explanations must satisfy two criteria: they must be class discriminative and accurately describe a specific image instance, explanations are distinct from descriptions, which provide a sentence based only on visual information, and definitions, which provide a sentence based only on class information. Unlike descriptions and definitions, visual explanations detail why a certain category is appropriate for a given image while only mentioning image relevant features. For example, consider a classification system that predicts a certain image belongs to the class "cardinal". A standard captioning system might provide a description such as "this is a red bird sitting on a tree branch". However, as this description does not mention discriminative features, it could also be applied to a "vermilion flycatcher".

Our first attempt in explanation generation [4] proposes a new model that focuses on the discriminating properties of the visible object, jointly predicts a class label, and explains why the predicted label is appropriate for the image. Namely, our visual explanation generating intelligent machines, a.k.a. agents, learn to fluently justify a class prediction and mention visual attributes, which reflect a strong class prior, e.g. "This is a red headed woodpecker because this bird has a red head and a pointy beak.". To the best of our knowledge, ours is the first

framework to produce deep visual explanations using natural language justications. Our joint vision and language explanation model combines classification and sentence generation by incorporating a loss function that operates over sampled sentences. We show that this formulation is able to focus generated text to be more discriminative and that our model produces better explanations than a description baseline. Our results also conrm that generated sentence quality improves with respect to traditional sentence generation metrics by including a discriminative class label loss during training.

The explanation agent in [4] learns to fluently justify a class prediction. However it may mention visual attributes which reflect a strong class prior, although the evidence may not actually be in the image. This is particularly concerning as ultimately such agents fail in building trust with human users. To overcome this limitation, in [5] we proposed a phrase-critic model to refine generated candidate explanations augmented with flipped phrases which we use as negative examples while training. At inference time, our phrase-critic model takes an image and a candidate explanation as input and outputs a score indicating how well the candidate explanation is grounded in the image. Our explainable AI agent is capable of providing counter arguments for an alternative prediction, i.e. counterfactuals, along with explanations that justify the correct classification decisions. Our model improves the textual explanation quality of fine-grained classification decisions on bird images by mentioning phrases that are grounded in the image. Moreover, our agent detects when there is a mistake in the sen-

Figure 5:
Our phrase-critic agent considers grounded visual evidence to determine if candidate explanations are image relevant. In this example, as many cardinals are red and have a black patch on their faces, mentioning and grounding those properties constitutes an effective factual explanation, i.e. rationalization. Furthermore, in our framework, informing the user of why an image does not belong to another class via the absence of certain attributes constitutes a counterfactual explanation

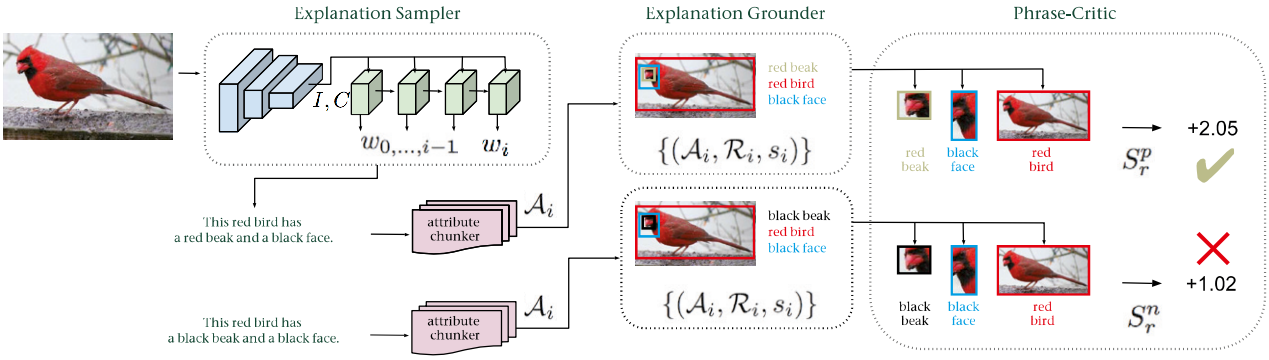


Figure 6: *Our phrase-critic model ensures that generated explanations are both class discriminative and image relevant. We first sample a set of explanations, chunk the sentences into noun phrases and visually ground constituent nouns. Our model assigns a score to each noun phrase-bounding box pair and selects the sentence with the highest cumulative score judging it as the most relevant explanation*

tence, grounds the incorrect phrase and corrects it significantly better than other models. Furthermore, human evaluations show that providing information using clear and plain language it indeed increases trust. In the example figure, as many cardinals are red and have a black patch on their faces, mentioning and grounding those properties constitutes an effective factual explanation, i.e. rationalization. Furthermore, in our framework, informing the user of why an image does not belong to another class, i.e. vermilion flycatcher, via the absence of certain attributes, i.e. black wings, constitutes a counterfactual explanation. Similarly, in [8], we proposed a new model which can jointly generate visual and textual explanations, using an attention mask to localize salient regions when generating textual rationales.

Finally, in the context of self-driving vehicles, we propose a two step framework [6] generating visual and textual explanations of driver's behavior. First, we use a visual (spatial) attention model to train a convolutional network end-to-end from images to the vehicle control commands, i.e., acceleration and change of course. The controllers attention identifies image regions that potentially influence the networks output. Second, we use an attention-based video-to-text model to produce textual explanations of model actions. The attention maps of controller and explanation model are aligned so that explanations are grounded in the parts of the scene that matters to the controller.

Visual Explanation Datasets. We propose visual question answering, activity recognition and driving explanations as testbeds for

studying explanations because they are challenging and important visual tasks which have interesting properties for explanation. For fine-grained visual classification, in [9] we collected 5 sentences for each of the images which do not only describe the content of the image, e.g., This is a bird, but also give a detailed description of the bird, e.g., red feathers and has a black face patch. Unlike other image-sentence datasets, every image in this dataset belongs to a class, and therefore sentences as well as images are associated with a single label. This property makes this dataset unique for the visual explanation task, where our aim is to generate sentences that are both discriminative and class-specific. Though these sentences are not originally collected for the visual explanation task, we observe that sentences include detailed and fine-grained category specific information. When ranking human annotations by output scores of our sentence classifier, we find that high-ranking sentences (and thus more discriminative sentences) include rich discriminative details.

" Our explainable AI agent is capable of providing counter arguments for an alternative prediction, i.e. counterfactuals, along with explanations that justify the correct classification decisions. "

Zeynep Akata

For example, the sentence "...mostly black all over its body with a small red and yellow portion in its wing" has a score of 0.99 for Red winged blackbird and includes details specific to this bird variety, such as red and yellow portion in its wing. As ground truth annotations are descriptions as opposed to explanations, not all annotations are guaranteed to include discriminative information. To generate satisfactory explanations, our model learns which features are discriminative from descriptions and incorpo-

rate discriminative properties into generated explanations.

VQA is a widely studied multimodal task that requires visual and textual understanding as well as common-sense knowledge. For complementary VQA pairs that ask the same question of two semantically similar images which have different answers in [8] we collected explanations that focus on the important factors for making a decision. Additionally, we collected annotations for activity recognition that explains a variety of cues, such as pose, global context, and the interaction between humans and objects, e.g., road biking and mountain biking include similar objects like bike and helmet, but road biking occurs on a road whereas mountain biking occurs on a mountain path. Finally, in [6] we proposed a dataset with over 6,984 video clips annotated with driving descriptions, e.g., The car slows down and explanations, e.g., because it is about to merge with the busy highway. Our dataset provides a new test-bed for measuring progress towards developing explainable models for self-driving cars.

Conclusion

Interpretability and explainability in artificial intelligence will have increasing impact in our daily lives. On 10th April 2018, 25 European countries have signed a declaration of cooperation on AI stating that AI can solve key societal challenges, from sustainable healthcare to climate change to cybersecurity. Clearly, the technology is becoming a key driver for economic growth through the digitization of industry and for society as a whole. However, concerns on trust and accountability indicate that humans should remain at the centre of development, deployment and decision making of AI and with that prevent harmful creation and use of AI applications as well as help advance public understanding of AI. This can only achieved by AI systems that are transparent in their decision process. In our research we propose to contribute to the sustainability and trustworthiness of AI-based solutions by designing and implementing vigilant AI systems with improved transparency such that they are more accountable.

References

[1] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label-embedding for image classication. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2016.

[2] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele. Evaluation of Output Embeddings for Fine-Grained Image Classication. In IEEE Computer Vision and Pattern Recognition (CVPR), 2015.

[3] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Longterm recurrent convolutional networks for visual recognition and description. In CVPR, 2015.

[4] L.-A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, and T. Darrell. Generating Visual Explanations. In European Conference of Computer Vision (ECCV), 2016.

[5] L. A. Hendricks, R. Hu, T. Darrell, and Z. Akata. Grounding visual explanations. In European Conference of Computer Vision (ECCV), 2018.

[6] J. Kim, A. Rohrbach, T. Darrell, J. Canny, and Z. Akata. Textual explanations for self driving vehicles. In European Conference of Computer Vision (ECCV), 2018.

[7] M. Malinowski, M. Rohrbach, and M. Fritz. Ask your neurons: A neural-based approach to answering questions about images. In ICCV, 2015.

[8] D. H. Park, L. A. Hendricks, Z. Akata, A. Rohrbach, B. Schiele, T. Darrell, and M. Rohrbach. Multimodal explanations: Justifying decisions and pointing to the evidence. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[9] S. Reed, Z. Akata, H. Lee, and B. Schiele. Learning deep representations of ne-grained visual descriptions. In IEEE Computer Vision and Pattern Recognition (CVPR), 2016.

[10] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative Adversarial Text to Image Synthesis. In International Conference on Machine Learning (ICML), 2016.

[11] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. IJCV, 2015.

[12] E. Schoenfeld, S. Ebrahimi, S. Sinha, T. Darrell, and Z. Akata. Generalized zero- and few-shot learning via aligned variational autoencoders. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[13] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schro, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, Caltech, 2010.

[14] Y. Xian, T. Lorenz, B. Schiele, and Z. Akata. Feature generating networks for zero-shot learning. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[15] Y. Xian, S. Sharma, B. Schiele, and Z. Akata. F-vaegan-d2: A feature generating framework for any-shot learning. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.