Response Time Analysis for Explainability of Visual Processing in CNNs

Eric Taylor*
Vector Institute
University of Guelph
eric.taylor@vectorinstitute.ai

Shashank Shekhar*
University of Guelph
Vector Institute
sshekhar@uoguelph.ca

Graham W. Taylor
University of Guelph
Vector Institute
CIFAR Canada AI Chair
gwtaylor@uoguelph.ca

Abstract

Explainable artificial intelligence (XAI) methods rely on access to model architecture and parameters that is not always feasible for most users, practitioners, and regulators. Inspired by cognitive psychology, we present a case for response times (RTs) as a technique for XAI. RTs are observable without access to the model. Moreover, dynamic inference models performing conditional computation generate variable RTs for visual learning tasks depending on hierarchical representations. We show that MSDNet, a conditional computation model with early-exit architecture, exhibits slower RT for images with more complex features in the ObjectNet test set, as well as the human phenomenon of scene grammar, where object recognition depends on intra-scene object-object relationships. These results cast light on MSDNet's feature space without opening the black box and illustrate the promise of RT methods for XAI.

1. Introduction

The majority of techniques developed for XAI depend on privileged access to the architecture and parameters of the model in question [1]. If XAI as a field is to provide satisfying explanations for decisions and behaviours to all users, researchers will need ways to generate explanations from "outside" the black-box — without having to inspect the model or using it in publicly infeasible circumstances. Explanations from outside the black box are desirable because they empower any user to investigate the cause and consequence of otherwise inscrutable model processes. Democratizing XAI will require the inference of black-box processes from observable behaviours.

The black-box problem in XAI is similar to the challenge faced in building explainable models of a different black box — the human mind. Cognitive psychologists make inferences about mental processes using only experimental stimuli and observed behaviours. Some ML researchers have already begun to model AI decisions using methods from cognitive psychology [2]. For inspiration, we turn to response time (RT) methods for explaining visual processing in humans. Unlike other XAI techniques, RT analyses can be conducted purely "outside" the black box.

We document a new XAI technique for dynamic inference models measuring only the softmax output and system clock. Our method is agnostic to model architecture, its implementation, and the dynamic inference strategy used, making it adaptable to any use-case where variable RTs are available. We test the human phenomena of object recognition from non-canonical viewpoints [3] and scene grammar [4] and show that RT analysis enables testing hypotheses about models' hierarchical feature representations.

2. Background

2.1. RT methods for explaining human vision

RT is an easily-measured external behaviour for humans that corresponds to the complexity of mental operations [5]. However, unlike human brains, most DNNs for computer vision perform a fixed number of operations over a static time interval, resulting in a uniform and uninformative RT distribution. If we want to use RT methods to explain DNN behaviours, we require a *distribution* of RTs.

2.2. Conditional Computation

Conditional computation [6] enables resource-efficient dynamic inference: these models allocate more computational resources and processing time to harder examples and less to easy examples. Viola and Jones [7] introduced the idea of using *cascades* in order to sequentially process input by increasingly complex classifiers. We focus our work on dynamic inference models that are augmented with auxiliary intermediate classifiers, providing a way to "early exit" for easy inputs as soon as the decision criteria is met [8]. Huang et al.'s Multi-Scale DenseNet (MSDNet) [8] is an example of a successful early exit model that utilizes dense

*equal contribution

connections and multi-scale features to achieve near state-of-the-art results on ImageNet [9]. Phuong et al [10] improve upon MSDNet's performance by additionally training with self-distillation from the final classifier into early-exits.

2.3. Scene Grammar

Humans depend on scene grammar, or object-scene congruities, to guide perception and attention [11]. A semantic violation occurs when an object's identity is statistically uncorrelated with that of other scene elements, whereas a syntactic violation occurs when the statistically reliable interposition of objects in a scene is upset. Scene grammar effects also occur in artificial neural networks, with evidence for decreased performance in object and scene classification [12]. If MSDNet composes higher-order object features representing these relationships, they ought to occur in deeper layers, manifesting as a slower RT.

3. Measuring RT from Dynamic Inference

The MSDNet architecture we employed has 38 layers with four scales of feature maps at each layer and five early exit classifiers attached at layers 10, 17, 24, 31, and 38 respectively. The model was pre-trained for image classification on ImageNet and achieves mean top-5 accuracy of 87.8, 90.4, 91.5, 91.8, and 92.3% across its classifiers, respectively. MSD-Net produces variable RTs that correspond to the complexity of representation required for classification by short-circuiting the remaining layers once an intermediate classification reaches a certain confidence threshold (measured as maximum value of classifier logits). This dovetails with the premise of RT methods in that more complex mental processes require more time. RT can therefore be used as a direct correlate of hierarchical processing.

The models were implemented with the PyTorch [13] library. To ensure that RT statistics were model- as well as implementation-independent, we relied on the system clock time elapsed during the forward pass of test input. The system used to perform the RT calculations had 2 × E5-2620 v2 Hex-core processors (12 physical CPU cores) and 128 GB RAM. Model parameters were loaded using PyTorch's parallel module onto two 11 GB NVIDIA 2080Ti GPUs and each input was passed individually during test time. Exact values of RT depend on system hardware, so our analysis and all conclusions are based on relative changes in RT in response to different inputs, architectures, or training regimens. Inferences about black-box processes are plausible to the extent that they produce reliable patterns of variance in response to controlled factors.

4. ObjectNet - Experiment & Results

To quantify the over-representation of canonical viewpoints and backgrounds in ImageNet, Barbu & Mayo et al. created ObjectNet, a 50,000-image test set for object recognition tasks that features common objects (many overlapping with ImageNet classes) viewed from non-canonical viewpoints on non-stereotyped backgrounds [3]. To achieve a more diverse set of features, they recruited thousands of workers to photograph objects from specified angles communicated via smartphone. The result is a test set that is more inclusive of objects' non-canonical features. When popular object recognition models are tested on ObjectNet, they exhibit performance decrements of up to 45%.

We profiled MSDNet, pre-trained on ImageNet for object recognition, on the ObjectNet test set, measuring system clock time as RT, top 5 classes for each of 5 classifiers, and the softmax confidence of those classifications. Because ObjectNet is difficult for object recognition models, it is a candidate test case for proof of concept that RT can be used to measure performance in DNNs. More importantly, because ObjectNet deliberately includes many non-canonical and presumably complex object features, which may be over-represented in later layers, it ought to display strong RT effects for models with hierarchical representation and dynamic inference, including MSDNet.

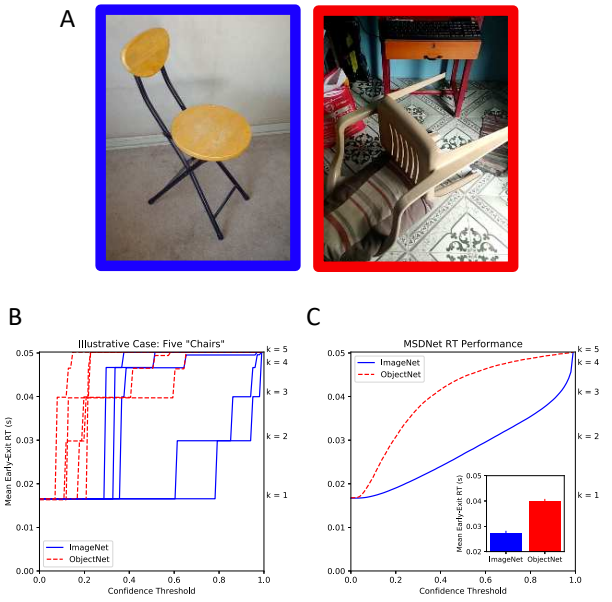

Figure 1. (A) Examples of chairs from ImageNet (top; blue) and ObjectNet (bottom; red). (B) Illustrative example of the step function describing the early-exit RT for five randomly-selected chairs from both ImageNet and ObjectNet. (C) Mean early-exit RT across all test images that had top-5 accuracy in the final auxiliary classifier. The values on the right vertical axis indicate the mean processing time for each of $k = 5$ auxiliary classifiers.

Figure 1 illustrates how quickly MSDNet can make a decision given a range of confidence values between 0 and 1. Results from five randomly-selected images of chairs from

both test sets are plotted to show how confidence propagates through the model, occasionally increasing RT in steps. The best performance would be a reverse-L shape, where the classifier $k = 1$ is sufficiently confident to identify the chair across the full range of thresholds.

RT was aggregated across all images in each set for every level of confidence. These values were submitted to an independent-samples *t*-test to affirm that RT can indeed be used as a reliable indicator of performance (t = 9.29, $p < .001$). Looking at the mean RT across all confidence thresholds, MSDNet processes ImageNet test stimuli 31.11% faster than ObjectNet stimuli with overlapping labels (27.40 ms vs. 39.77 ms). Because ObjectNet is characterized by a range of rotational viewpoints and non-canonical backgrounds, we infer that the higher-order features required to identify these items are better represented across MSDNet's auxiliary classifiers.

5. SCEGRAM - Experiment & Results

The SCEGRAM database [14] is a set of images of 62 indoor scenes with carefully curated manipulations of scene grammar. For each scene, there are four images (see Figure 2): consistent scene grammar (CON), inconsistent semantics (SEM), inconsistent syntax (SYN), and inconsistent semantics and syntax (SEMSYN). The semantic and syntactic manipulations are fully crossed. So for a given scene, say a kitchen counter, there are four versions of the image: a pot in a pile of dishes (CON); a clock in a pile of dishes (SEM); a pot affixed to the dishwasher door (SYN); and a clock affixed to the dishwasher door (SEMSYN). All other visual features in the scene are identical, allowing for experimental inferences about the scene grammar manipulations.

We profiled MSDNet object recognition performance on 248 SCEGRAM images, taking the same measurements as in the ObjectNet experiment. If MSDNet composes higher-order object features in later layers, then early-exit decisions should be made using coarse features. If high-confidence classification depends on processing higher-order object features such as inter-object or object-scene relationships, then images with inconsistent scene grammar ought to be processed slower on average than images with consistent scene grammar. Because MSDNet was pre-trained for object recognition, we can predict that RT effects should be specific to semantic, rather than syntactic inconsistencies.

Step functions of RT given the full range of confidence thresholds are displayed in Figure 3. The mean early-exit RT across all 62 scenes for the full range of confidence thresholds was determined to generalize a profile of the relationship between RT, confidence, and scene grammar (see Figure 3B). Visual inspection of these means reveals better performance for scenes with consistent grammar with local troughs around 0.2 and 0.4 confidence. In human subjects experiments with electro-physiological or other time series

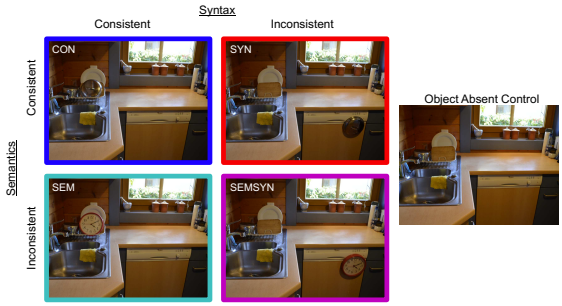

Figure 2. Illustrative example of SCEGRAM's test stimuli. Images feature two orthogonal manipulations: semantic (in)consistency and syntactic (in)consistency. A single scene is therefore presented to the model four times, with different combinations of scene grammar. Each of the four images has an additional clone without the critical object which controls for low-level visual differences with their object-present counterparts. These clones are not photoshopped and have no low-level artifacts; the critical object was physically removed from the photograph.

data, it is common to specify a window of interest within which to compare RTs from different conditions [4]. We defined the boundaries for these windows as the mean confidence required by the model to reach the subsequent classifier's RT (e.g. if RT for $k = 2$ is 0.03 s, what is the mean confidence at which MSDNet reaches 0.03 s?). To characterize whether RT differences were reliable across scenes, we submitted the data to a three-way repeated-measures ANOVA with semantics (consistent, inconsistent), syntax (consistent, inconsistent), and classifier window (thresholds described above) as within-subjects factors. As expected, the classifier produced a strong effect on RT ($F(3,183)$ = 222.59, $p < .001$), indicating that RTs were slower as image processing progressed through the model. The critical result is that, as expected, there was a significant effect of semantics ($F(1,61)$ = 4.87, p = .031), indicating that SCEGRAM images with inconsistent semantic information were classified reliably slower than images with consistent semantics. No other effects reached statistical significance.

We can also predict when scene grammar effects should not emerge on RT. RTs were collected for the object-absent clone images corresponding to each of the same 248 SCEGRAM images used above. These images are identical except that the critical object has been removed, resulting in images with no inconsistencies. As predicted, there was no effect of semantic inconsistency ($F(1,61) = 0.36$, $p = 0.55$). Likewise, training MSDNet with self-distillation, which improves early layers' feature representations and achieves higher accuracy in earlier auxiliary classifiers, shows no scene grammar effects because it performs at a higher level relatively early. This is evident from the linear, diagonal shape in Figure 3D relative to the original bowed shape.

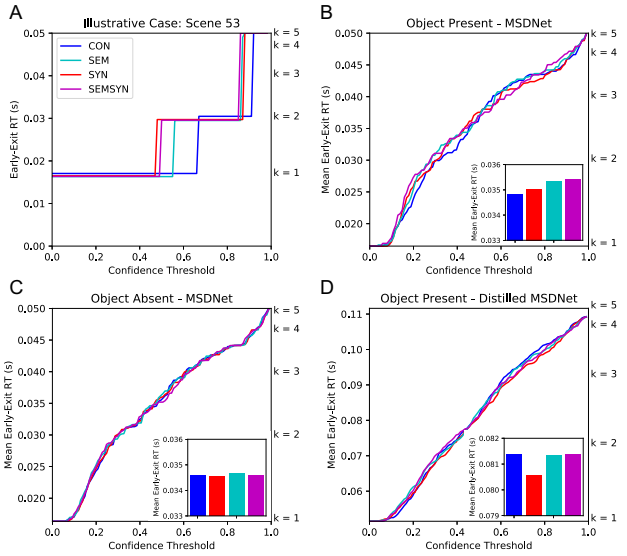

Figure 3. (A) Illustrative example of the step function describing the early-exit RT for a single scene in SCEGRAM. (B) Mean early-exit RT across all SCEGRAM scenes, grouped by scene grammar condition. Consistent scene grammar (CON) has visibly faster RT for responses around 0.2 and 0.4 confidence. The values on the right vertical axis indicate the mean processing time for each of $k = 5$ auxiliary classifiers. (C) Same analysis for the object-absent clone images in SCEGRAM. These are technically all semantically consistent. As predicted, they share the same RT profile as CON. (D) Same analysis for distilled MSDNet.

6. Conclusions and Future Work

These experiments demonstrate the value of RT analyses as a method to probe the inner workings of otherwise opaque models. We were able to test *a priori*, falsifiable hypotheses about the relationship between input space and response time using two different test sets. We showed that classification that depends on access to higher-layer features takes longer for dynamic inference models using conditional computation. These analyses could be used to form expectations for when and how models should perform in situations where explanations are desirable, but privileged access to a model is denied.

Like humans, MSDNet composes features hierarchically, with distributed complexity across its intermediate classifiers. However, the RT analysis failed to reveal an interaction between SCEGRAM condition and processing window, so we cannot make any statements about *when* the semantic features are processed in MSDNet. In contrast, humans show a clear double dissociation in neurophysiological recordings: semantic violations trigger an early neural signature, whereas syntactic violations occur later [4].

In future, we would like to extend our RT analyses to other dynamic inference models as well as to other visual tasks and their controls, starting with the effect of rotation, background and viewpoint variations on object detection (published but yet to be released for ObjectNet).

Acknowledgement

The authors would like to thank Mary Phuong for sharing her code and pre-trained models.

References

[1] I. Rahwan *et al.*, "Machine behaviour," *Nature*, vol. 568, no. 7753, 2019.

[2] S. Ritter *et al.*, "Cognitive psychology for deep neural networks: A shape bias case study," in *Proceedings of the 34th International Conference on Machine Learning*, vol. 70, 2017.

[3] A. Barbu *et al.*, "Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models," in *Advances in Neural Information Processing Systems*, 2019.

[4] M. L.-H. Võ *et al.*, "Differential Electrophysiological Signatures of Semantic and Syntactic Scene Processing," *Psychological Science*, vol. 24, no. 9, 2013.

[5] F. C. Donders, "On the speed of mental processes," *Acta psychologica*, vol. 30, 1868.

[6] Y. Bengio *et al.*, "Estimating or Propagating Gradients Through Stochastic Neurons for Conditional Computation," *arXiv:1308.3432 [cs]*, 2013.

[7] P. Viola *et al.*, "Rapid object detection using a boosted cascade of simple features," in *IEEE computer society conference on computer vision and pattern recognition (CVPR)*, vol. 1, 2001.

[8] G. Huang *et al.*, "Multi-Scale Dense Networks for Resource Efficient Image Classification," *arXiv:1703.09844 [cs]*, 2018.

[9] J. Deng *et al.*, "Imagenet: A large-scale hierarchical image database," in *IEEE conference on computer vision and pattern recognition*, 2009.

[10] M. Phuong *et al.*, "Distillation-based training for multi-exit architectures," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[11] M. L.-H. Võ *et al.*, "Reading scenes: How scene grammar guides attention and aids perception in real-world environments," *Current Opinion in Psychology*, vol. 29, 2019.

[12] A. Bayat *et al.*, "Scene Grammar in Human and Machine Recognition of Objects and Scenes," in *Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018.

[13] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, 2019.

[14] S. Öhlschläger *et al.*, "Scegram: An image database for semantic and syntactic inconsistencies in scenes," *Behavior research methods*, vol. 49, no. 5, 2017.