Interpreting Black Box Models via Hypothesis Testing

Collin Burns
Department of Computer Science
Columbia University
New York, New York
collin.burns@columbia.edu

Jesse Thomason
Department of Computer Science
University of Washington
Seattle, Washington
thomason.jesse@gmail.com

Wesley Tansey
Department of Systems Biology
Columbia University Medical Center
New York, New York
wesley.tansey@columbia.edu

Abstract

While many methods for interpreting machine learning models have been proposed, they are often ad hoc, difficult to interpret, and come with limited guarantees. This is especially problematic in science and medicine, where model interpretations may be reported as discoveries or guide patient treatments. As a step toward more principled and reliable interpretations, in this paper we reframe black box model interpretability as a multiple hypothesis testing problem. The task is to discover "important" features by testing whether the model prediction is significantly different from what would be expected if the features were replaced with uninformative counterfactuals. We propose two testing methods: one that provably controls the false discovery rate but which is not yet feasible for large-scale applications, and an approximate testing method which can be applied to real-world data sets. In simulation, both tests have high power relative to existing interpretability methods. When applied to state-of-the-art vision and language models, the framework selects features that intuitively explain model predictions. The resulting explanations have the additional advantage that they are themselves easy to interpret.

1 Introduction

When using a black box model to inform high-stakes decisions, one often needs to audit that model. At a minimum, this means understanding which features are influencing its prediction. When the data or predictions are random variables, it may be impossible to determine the important features without some error. In these cases, the reported "important" features should come with some statistical control on the error rate. This last part is critical: if interpreting a black box model is intended to build trust in its reliability, then the method used to interpret it must itself be reliable, robust, and comprehensible.

In this paper, we address the need for reliable interpretation by casting black box model interpretability as a multiple hypothesis testing problem. Given a black box model and an input of interest, we test subsets of features to determine which are collectively important for the prediction. Importance is measured relative to the model prediction if features are replaced with draws from an uninformative counterfactual distribution. Reframing interpretability as hypothesis testing, we develop a framework in which we can control the false discovery rate of important features at a user-specified level.

Within this framework, we propose two hypothesis testing methods: the Interpretability Randomization Test (IRT) and the One-Shot Feature Test (OSFT). The first provably controls the false discovery

Preprint. Under review.

(a) $p = 0.904$ (Original) (b) LIME (c) $p = 0.898$ (-0.006) (d) $p = 0.307$ (-0.597)

Figure 1: Interpretations by the OSFT, one of the methods we propose, and LIME [20]. The ground truth class is "Impala". The impala on the right (bounded by the blue box and replaced by the counterfactual in 1d) was selected as important by the OSFT, while the impala on the left (bounded by the red box and replaced in 1c) was not. In contrast, LIME selects only the impala on the left as important. For each image, p is the predicted probability of the correct class (Impala). The predicted probabilities on the generated counterfactual inputs reassure us that only the impala on the right had a significant effect on the model output, as selected by the OSFT.

rate (FDR), but is computationally intensive. The second is a fast, approximate test that can be used to interpret models on large datasets. In synthetic benchmarks, both tests empirically control the FDR and have high power relative to other methods from the literature. When applied to state-of-the-art vision and language models, the OSFT selects features that intuitively explain model predictions.

Using these methods, one can also visualize why certain features were selected as important. For example, in Fig. 1 we show interpretations of an image classification by both the OSFT and by LIME [20], a popular black box interpretability method. The ground truth label is "Impala", and there are two impala in the image. LIME selects just the one on the left as important. In contrast, the OSFT selects only the impala on the right. Because the framework we propose is based on counterfactuals, we can visualize the counterfactual inputs and how the model changes on those inputs. Fig. 1 shows that the "Impala" class probability drops significantly when the impala on the right is replaced by an uninformative counterfactual, while it decreases only negligibly when the impala on the left is replaced. This reassures us of the validity of the interpretation by the OSFT. It also highlights how it can be misleading to evaluate interpretability methods by visual inspection of the selected features.

2 Background

We focus on prediction-level interpretation. Given a black box model and an input, the goal is to explain the output of the model in terms of features of that input.

Interpreting machine learning models. Most methods for interpreting model predictions are based on optimization. Gradient-based methods like Saliency [22] and DeepLift [21] visualize the saliency of each input variable by analyzing the gradient of the model output with respect to the input example. Black box optimization-based methods that don't assume gradient access include LIME [20], SHAP [18], and L2X [8]. LIME approximates the model to be explained using a linear model in a local region around the input point, and uses the weights of the linear model to determine feature importance scores. SHAP takes a game-theoretic approach by optimizing a kernel regression loss based on Shapley values. L2X selects explanatory features by maximizing a variational lower bound on the mutual information between subsets of features and the model output.

Some existing interpretability methods, like those we present in this paper, are based on counterfactuals. Fong and Vedaldi [11] generate a saliency map by optimizing for the smallest region that, when perturbed (such as by blurring or adding noise), substantially drops the class probability. However, the perturbations used lead to counterfactual inputs that are outside the training distribution. Given the lack of robustness of many modern machine learning models [14], it is unclear how to interpret the resulting explanations. Cabrera et al. [5] introduce an interactive setup for interpreting image classifiers in which users select regions of a given image to inpaint (i.e., delete and fill in with a plausible counterfactual) using a deep generative model. The system then visualizes the change in probabilities for the top classes. Chang et al. [7] similarly use inpainting models but, like Fong and Vedaldi [11], use this to generate a saliency map.

Optimization-based approaches all generally require defining a penalized loss function. Tuning the hyperparameters of these functions is done by visual inspection of the results, and this interactive tuning is often misleading [17, 1]. Optimization may also overestimate the importance of some variables due to the winner's curse [25]. That is, by looking at the impact of variables and selecting for those with high impact, the post-selection assessment of their importance is biased upward. This phenomenon is known in statistics as post-selection inference and requires careful analysis of the penalized likelihood to derive valid inferences [16]. The methods proposed in this paper avoid this issue by taking a multiple hypothesis testing approach.

Multiple hypothesis testing and FDR control. In multiple hypothesis testing (MHT), $\mathbf{z} = (z_1, \ldots, z_N)$ are a set of observations of the outcomes of N experiments. For each observation, if the experiment had no effect ($h_i = 0$) then z_i is distributed according to a null distribution $\pi_0^{(i)}(z)$; otherwise, the experiment had some effect ($h_i = 1$) and z_i is distributed according to some unknown alternative distribution. The null hypothesis for every experiment is that the test statistic was drawn from the null distribution: $H_0^{(i)} : h_i = 0$. For a given prediction $\hat{h}_i$, we say it is a true discovery if $\hat{h}_i = 1 = h_i$ and a false discovery if $\hat{h}_i = 1 \neq h_i$. Let $\mathcal{S} = \{i : h_i = 1\}$ be the set of observations for which there was some effect (true positives) and $\hat{\mathcal{S}} = \{i : \hat{h}_i = 1\}$ be the set of reported discoveries. The goal in MHT is to maximize the true positive rate, also known as *power*, $\text{TPR} \coloneqq \mathbb{E}\left[\frac{\#\{i : i \in \hat{\mathcal{S}} \cap \mathcal{S}\}}{\#\{i : i \in \mathcal{S}\}}\right]$, while controlling an error metric; here we focus on controlling the false discovery rate, $\text{FDR} \coloneqq \mathbb{E}\left[\frac{\#\{i : i \in \hat{\mathcal{S}} \setminus \mathcal{S}\}}{\#\{i : i \in \hat{\mathcal{S}}\}}\right]$. Methods that control FDR ensure that reported discoveries are reliable by guaranteeing that, on average, no more than a small fraction of them are false positives. In the context of black box model interpretation, we seek to control the FDR in the reported set of important features that contributed toward a model's prediction.

3 Interpretability as multiple hypothesis testing

We consider a feature important if its impact on the model output is surprising relative to a counterfactual. We formalize this as a hypothesis testing problem. For each feature, we test whether the observed model output would be similar if the feature was drawn from some uninformative counterfactual distribution. Tests that control the corresponding FDR will then only select features whose effect on the model output is sufficiently extreme with respect to this counterfactual distribution. We focus on contextual importance: we are interested in whether a feature contributes to a prediction in the context of the other features.

Suppose we want to understand the output of a model f on an input $x \in \mathbb{R}^d$ that was sampled from some distribution $P(X)$. For $X \in \mathbb{R}^d$ and $S \subset [d] \coloneqq \{1, \ldots, d\}$, we let X_S denote X restricted to the set S, and let X_{-S} denote X restricted to the features not in S.

Definition 1. *Suppose $T(\cdot)$ is a test statistic and $Q(X_S|X_{-S})$ is some conditional distribution. Let $T_P(f(X))$ be the true distribution of $T(f(x))$, $x \sim P(X)$, and let $T_Q(f(X))$ be the distribution of $T(f(\widetilde{x}))$, where $\widetilde{x} = (\widetilde{x}_S, x_{-s})$ and $\widetilde{x}_S \sim Q(X_S|x_{-s})$. The null hypothesis, H_0, is that $T_P(f(X))$ is stochastically less than $T_Q(f(X))$,*

$$H_0 \colon T(f(x)) \sim T_P(f(X)) \preceq T_Q(f(X)) \,. \tag{1}$$

(A random variable Y is stochastically less than a random variable Z if for all $u \in \mathbb{R}$, $Pr[Y > u] \leq Pr[Z > u]$.) Given $x \in \mathbb{R}^d$, a model f, and a subset of features $S \subset [d]$, we say that x_S is important with respect to the test statistic $T(\cdot)$ and the conditional $Q(X_S|X_{-S})$ if H_0 is false.

The null hypothesis in Eq. (1) covers a family of null distributions for the observed test statistic. Informally, it includes all distributions that put more mass on smaller (i.e., less extreme) statistics than samples from Q would. The distribution corresponding to the pointwise equality null hypothesis,

$$H_0 \colon T(f(X)) \sim T_P(f(X)) \stackrel{d}{=} T_Q(f(X)) \tag{2}$$

will therefore put the most mass on large test statistics of any member of the null family. Consequently, any test statistic for the point null is a conservative statistic for Eq. (1), the familywise null. We use the point null as a proxy for the familywise null, as we can only sample from the former, $T_Q(f(X))$.

Algorithm 1 Interpretability Randomization Test (IRT)

Require: (features $(x_1, \ldots, x_d)$, trained model f, conditional model $Q(X_S|X_{-S})$, test statistic T, target FDR threshold α, subsets of features to test $S_1, ..., S_N \subset [d]$, sample size K)

Compute model output $\hat{y} \leftarrow f(x)$
Compute test statistic $t \leftarrow T(\hat{y})$
for $i \leftarrow 1, \ldots, N$ **do**
  for $k \leftarrow 1, \ldots, K$ **do**
    Sample $\widetilde{x}_{S_i} \sim Q(X_{S_i}|X_{-S_i} = x_{-s_i})$
    Compute model output $\tilde{y}^{(k)} \leftarrow f((\widetilde{x}_{S_i}, x_{-S_i}))$
    Compute the test statistic $\tilde{t}^{(k)} \leftarrow T(\tilde{y}^{(k)})$
  end for
  Compute the p-value

$$\hat{p}_i = \frac{1}{K+1}\left(1 + \sum_{k=1}^{K} \mathbb{I}\left[t \leq \tilde{t}^{(k)}\right]\right)$$

end for
$\tau =$ MHT-Correct$(\alpha, \hat{p}_1, \ldots, \hat{p}_K)$
Return discoveries at the α level: $\{i \colon \hat{p}_i \leq \tau\}$

The definition above applies to any conditional distribution $Q(X_S|X_{-S})$, but it is only a useful notion of interpretability for some distributions. For example, the generated counterfactuals, $\widetilde{X} = (\widetilde{X}_S, x_{-S})$, should lie in the support of the true distribution, $P(X)$. This is important because the model has only been trained on inputs from $P(X)$. As work on robustness and adversarial examples illustrates [14, 12], model behavior on out-of-distribution inputs can be counterintuitive, making the definition of importance with respect to such a distribution potentially misleading.

3.1 The Interpretability Randomization Test

As mentioned, it suffices to control the FDR for the point null, Eq. (2). In general, the null distribution will not be available in closed form, but if we can sample from $Q(X_S|X_{-S})$ then we can repeatedly sample new inputs, calculate a test statistic, and compare it to the original test statistic. Randomization tests build an empirical estimate of the likelihood of observing a test statistic as extreme as that observed under the null distribution. Algorithm 1 details the Interpretability Randomization Test. Adding one to the numerator and denominator ensures that this is a valid p-value for finite samples from H_0 [10], meaning it is stochastically greater than $U(0, 1)$.

When testing multiple features, controlling the error rate requires applying a multiple hypothesis testing correction procedure. The choice of MHT-Correct in Algorithm 1 depends on the goal of inference and the dependence between features. We focus on controlling the FDR via Benjamini-Hochberg (BH) [3], which controls the FDR when the tests are independent or in a large class of positive dependencies [4]. We found this robustness to be sufficient to control the FDR in practice (see Table 1). [1] For completeness, we provide the procedure in Appendix A.1.

3.2 The One-Shot Feature Test

The IRT requires repeatedly sampling counterfactuals, which can be computationally expensive. For instance, in the image and language case studies in Section 4, we generate counterfactuals from deep conditional models. Running these models thousands of times per feature is intractable. For these cases, we propose the One-Shot Feature Test (OSFT), which requires only a single sample from the conditional distribution. The OSFT provably controls the FDR when the features or test statistics are independent; see Appendix A.2 for the proof. As with the IRT using the BH correction procedure, in practice the OSFT controls the FDR in a wider class of scenarios than we can prove theoretically (see Table 1). The OSFT is given in Algorithm 2.

[1] If one wants provable FDR control under arbitrary dependencies, one can instead use the Benjamini–Yekutieli procedure [4], but this essentially just amounts to decreasing the rejection threshold by a logarithmic factor.

Algorithm 2 One-Shot Feature Test (OSFT)

Require: (features $(x_1, \ldots, x_d)$, trained model f, conditional model $Q(X_S|X_{-S})$, test statistic T, target FDR threshold α, subsets of features to test $S_1, ..., S_N \subset [d]$)

Compute test statistic $t \leftarrow T(f(x_1, \ldots, x_d))$
for $i \leftarrow 1, \ldots, N$ **do**
  Sample $\widetilde{x}_{S_i} \sim Q(X_{S_i}|X_{-S_i} = x_{-S_i})$
  Compute model output $\tilde{y}^{(i)} \leftarrow f((\widetilde{x}_{S_i}, x_{-S_i}))$
  Compute the test statistic, $\tilde{t}^{(i)} \leftarrow T(\tilde{y}^{(i)})$
  Compute the difference statistic, $z^{(i)} \leftarrow t - \tilde{t}^{(i)}$
end for
$z^* \leftarrow \underset{z}{\text{argmin}} \left[\frac{1+\#\, z^{(i)} \leq -z}{\#\, z^{(i)} \geq z} \leq \alpha \right]$
Return discoveries at the α level: $\{i\colon z^{(i)} \geq z^*\}$

3.3 Test statistic choice

Some choices of the test statistic, $T(\cdot)$, may be more appropriate for certain tasks and may have higher power than other choices. Two classical statistics are one-sided and two-sided tail probabilities. One-sided tests have a preferred direction of testing, while two-sided tests consider both tails of the null distribution. In the one-sided case, testing for an increase in output can be done by making the test statistic the identity, $T(Y) = Y$. A two-sided IRT statistic requires only modifying the Algorithm 1 to look at both tails of the distribution of $\tilde{t}$. However, the OSFT has no explicit explicit null distribution for each sample. In this case, we can still perform a two-sided test by drawing an extra null variable as a "centering" sample:

$$\bar{X}_i \sim Q(X_i|X_{-i} = x_{-i})\,, \qquad \bar{Y} = f(\bar{X}_i, x_{-i})\,, \qquad T(Y) = (Y - \bar{Y})^2\,. \tag{3}$$

This turns the one-shot procedure into a two-shot procedure.

4 Experiments

We first compare the IRT and OSFT to six baseline interpretability methods on synthetic datasets. We then conduct two case studies applying the OSFT to explain the predictions of a state-of-the-art image classifier (Inception v3) on ImageNet and a state-of-the-art sentiment classifier (BERT) on movie reviews. Code for all experiments is available at `https://github.com/collin-burns/interpretability-hypothesis-testing`.

4.1 Synthetic benchmark

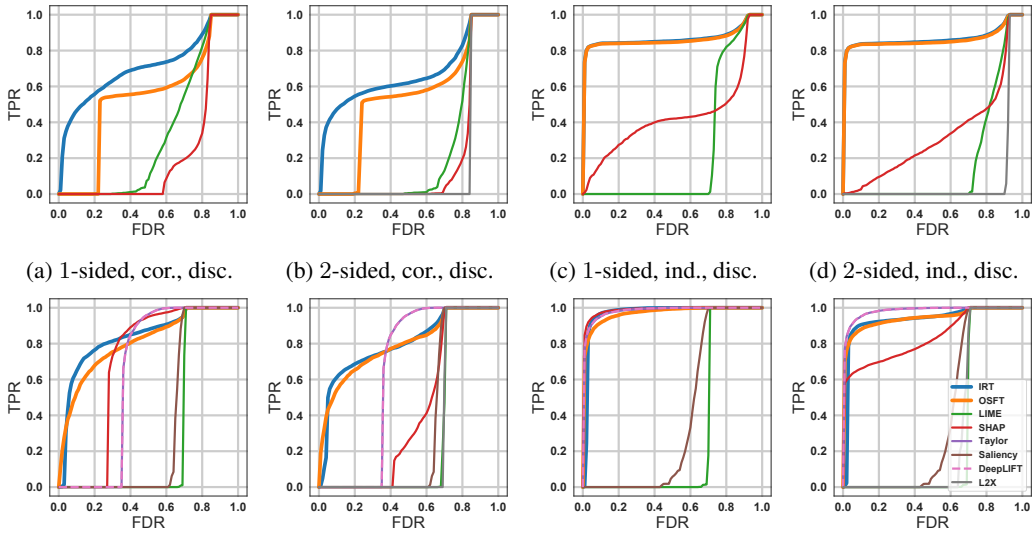
To compare the IRT and OSFT to exising methods, we evaluate how the power varies as a function of the false discovery rate for each method. This requires determining exactly when the null hypothesis is true. In general, this may be infeasible for the null hypothesis given in Eq. (1). However, for certain distributions, the point null given in Eq. (2) is feasible to evaluate. We consider two such distributions: one which has independent features, and the other which has correlated features. We also consider two different models to interpret: a neural network and a discontinuous model.

To empirically evaluate the FDR and TPR, we will use the fact that for each of the following distributions and for both test statistics that we consider, the point null hypothesis, Eq. (2), is equivalent to

$$H_0\colon f(x) \sim f(X) \overset{d}{=} f(\widetilde{X})\,, \tag{4}$$

where, as before, $\widetilde{X} = (\widetilde{X}_S, x_s)$ and $\widetilde{X}_S \sim Q(X_S|x_{-S})$.

Inputs. For the independent distribution, for each feature i, with probability $h = 0.3$ we let $X_i \sim \mathcal{N}(4, 1)$, and with probability $1 - h$, $X_i \sim \mathcal{N}(0, 1)$. We then let $Q(X_i|X_{-i})$ be $\mathcal{N}(0, 1)$. For the correlated distribution, for each feature i, with probability $h = 0.3$, $X_i \sim \mathcal{N}(4, 1)$, and with probability $1 - h$, $X_i \sim \mathcal{N}(m, 1)$, where $m = \sum_{j=1}^{i-1} \beta_j x_j$ and where $\beta_j \sim \mathcal{N}(0, \frac{1}{16})$ for each feature j (fixed for all examples). We then let $Q(X_i|X_{-i})$ be $\mathcal{N}(m, 1)$.



(a) 1-sided, cor., disc. (b) 2-sided, cor., disc. (c) 1-sided, ind., disc. (d) 2-sided, ind., disc.

(e) 1-sided, cor., nn. (f) 2-sided, cor., nn. (g) 1-sided, ind., nn. (h) 2-sided, ind., nn.

Figure 2: The IRT and OSFT have higher power than the baseline methods in most cases, and have comparable power to the best baseline methods in the remaining cases. The curves were averaged over 10 independent runs.

Models. The first model is a paired thresholding model. On an input $X = (X_1, \ldots, X_{2p}) \in \mathbb{R}^{2p}$, the model output is defined to be,

$$f(X) = \sum_{i=1}^{p} w_i \, \mathbf{1}\left[|X_i| \geq t \wedge |X_{i+p}| \geq t\right], \tag{5}$$

for $w \in \mathbb{R}^p$ and $t \geq 0$. We let $w_i = 0.5 + v_i, v_i \sim \text{Gamma}(1, 1)$, and fix $t = 3$.

For each feature $i \in [2p]$, the null hypothesis in Eq. (4) is that $\hat{y} = f(x)$ was sampled from the distribution $f(\widetilde{X}^{(i)})$ where $\widetilde{X}^{(i)} = (\widetilde{X}_i, x_{-i})$ and $\widetilde{X}_i \sim Q(X_i|x_{-i})$. For either data distribution above, when $i \in [p]$ this is false if and only if $|x_{i+p}| \geq t$ (so that feature i can affect the model output at all) and x_i was sampled from the "interesting" distribution $\mathcal{N}(4, 1)$ (since otherwise, by construction, it must have been sampled from $Q(X_i|X_{-i})$, in which case the null would be true). Similarly, for each feature $i \in \{p+1, \ldots, 2p\}$, the null hypothesis is false if and only if $x_{i-p} \geq t$ and x_i was sampled from $\mathcal{N}(4, 1)$. We set $p = 50$, so the number of parameters is 100.

The discontinuous model can only be interpreted by gradient-free interpretability methods. In order to compare our approach to methods that only apply to neural networks (e.g., [21]) or differentiable models (e.g., [22]), we also consider the following setup that mirrors that of Chen et al. [8]. We let $Y := \sum_{i=1}^{d} |X_i|$ be the ground truth response variable, with $d = 25$, and train a two-layer neural network to near-zero test error with this response as the label. Given the test error, we can assume that the network has successfully learned which features are important for the model. We then interpret the trained network. If the network indeed learned the model correctly, then the feature x_i is important if and only if it was sampled from the interesting distribution, $\mathcal{N}(4, 1)$. In particular, each feature is always used by the model. Hence, if x_i was sampled from $\mathcal{N}(4, 1)$, then $f(x)$ was sampled from a different distribution than $f(\widetilde{X})$, so that the null in Eq. (4) is false.

Comparison. For the discontinuous model, we compare against three other black box interpretability methods: LIME [20], SHAP [18], and L2X [8]. For interpreting the neural network, we additionally compare against three methods for interpreting deep learning models: Saliency [22], DeepLIFT [21], and another strong baseline method called Taylor [8]. Taylor computes feature values by multiplying the value of each feature by the gradient of the output with respect to that feature.

L2X directly selects k features to explain a prediction, where k is treated as a hyperparameter. The remaining methods output feature values corresponding to how large of an effect each feature had on

Table 1: Empirical FDR and TPR for the IRT and OSFT on each distribution and model ($\alpha = 0.2$).

	FDR/TPR	
Method, Distribution, Model	1-sided	2-sided
OSFT, Independent, Discontinuous	0.006 / 0.836	0.006 / 0.833
OSFT, Independent, Neural Network	0.212 / 0.962	0.189 / 0.910
OSFT, Correlated, Discontinuous	0.073 / 0.025	0.044 / 0.004
OSFT, Correlated, Neural Network	0.142 / 0.611	0.143 / 0.605
IRT, Independent, Discontinuous	0.002 / 0.393	0.002 / 0.392
IRT, Independent, Neural Network	0.139 / 0.979	0.137 / 0.913
IRT, Correlated, Discontinuous	0.000 / 0.000	0.000 / 0.000
IRT, Correlated, Neural Network	0.129 / 0.716	0.130 / 0.641

the given input. To compare these to the IRT and OSFT, which automatically choose a number of features to select as important, we suppose that these methods are able to control the false discovery rate at a particular level, and measure the true positive rate at that level. Specifically, we plot how the empirical FDR and TPR change as each method increases the number of features it selects. Because the FDR is not necessarily monotonic as the number of selected features increases, for each FDR level we take the maximum TPR achieved for which the FDR is controlled at the specified level.

We consider one-sided and two-sided variants for the feature value methods. For the one-sided test, we track how the TPR and FDR vary as the k features with the largest values are selected, for increasing k. For the two-sided variant, we instead select the k features with the largest absolute values. On the other hand, L2X directly selects features that are broadly relevant to the output of the model. This limits L2X to only the two-sided case. We use the default settings for each method. For the IRT, we used $K = 100$ permutations and the same two-sided test statistic as for the OSFT.

Results. Fig. 2 shows the TPR of each method as a function of the FDR, averaged over 10 independent runs with 100 test samples each. The IRT and OSFT have higher power than the baseline methods for FDR levels of interest, except when interpreting the neural network with independent features. In that case, both methods are still competitive with the best baseline methods.

An advantage of the IRT and OSFT not accounted for in Fig. 2 is that they can automatically select features at a given FDR threshold α. To verify that they control the FDR and have high power, in Table 1 we show the FDR and TPR of both methods for each setting described above, where we set $\alpha = 0.2$ (other reasonable choices of α, such as 0.05, give qualitatively similar results). We find that both methods indeed nearly always control the FDR below the target level of 0.2, and often by a large margin. An exception was the OSFT when interpreting the neural network with independent features using the one-sided test. In that case, the empirical FDR was 0.212, barely above the target level. Moreover, both methods usually have reasonably high power; the one notable exception was when interpreting the discontinuous model with correlated inputs. However, in practice, for real models and datasets, we found that low power was not an issue (see Sections 4.2 and 4.3).

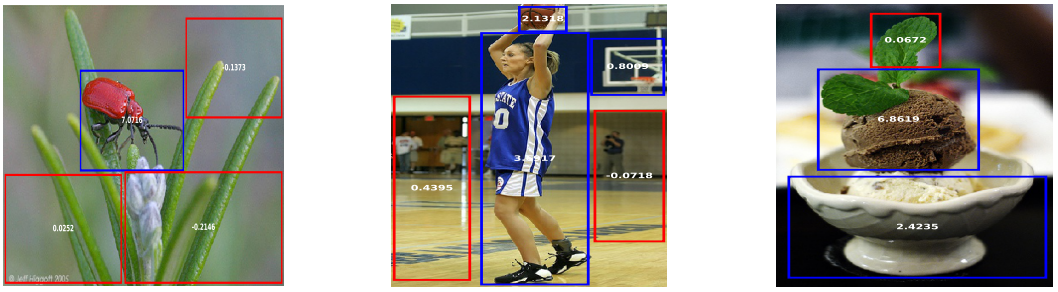
4.2 Interpreting a deep image classifier

Next, we apply the OSFT to interpreting Inception v3 [24], a deep image classifier. We used the pretrained model in the `torchvision` package. As the conditional distribution, $Q(X_S|X_{-S})$, we use a state-of-the-art generative inpainting model [32]. Inpainting models replace subsets of pixels with counterfactuals that are often reasonable proxies for background pixels. We define the model output to be the logits for the predicted class, and use the one-sided statistic, testing subsets of features corresponding to contiguous image patches. In general, pixel feature subsets can be selected in any way, as long as they are non-overlapping, and the best method for doing so will be application dependent. For simplicity here, the patches were selected manually.

We test 50 ImageNet images, some of which were taken from [11] for comparison. Bounding boxes were drawn around objects, parts of objects, and parts of the background. In total, 222 patches were tested at an FDR threshold of $\alpha = 0.2$. Of these patches, 72 (about 32%) were selected as important.

Results. Figure 3 shows 6 of the images and the patches that were tested for each of them. The bounding box color indicates whether the patch was found to be important (blue) or not (red).

(a) Armadillo ($p = 0.970$) (b) Saxaphone ($p = 0.993$) (c) Chocolate sauce ($p = 0.995$)



(d) Leaf beetle ($p = 0.950$) (e) Basketball ($p = 0.974$) (f) Ice cream ($p = 0.923$)

Figure 3: Image classifier interpretations using the OSFT. The threshold for rejecting the null hypothesis is 0.68. Boxes are blue if that bounding box was selected as important, and red otherwise. The number inside the box is the value of the difference statistic for that bounding box.

The value of the difference statistic is printed inside each patch. Intuitively, the bounding boxes corresponding to the ground truth labels are often selected. Interpretations of the remaining 44 images can be found in Appendix B.

While we found most selections to be intuitive, we also investigated some of the counterintuitive features that the OSFT selected, and found that in many cases they were due to especially unrealistic counterfactuals. An example of this is given in Fig. 4 in Appendix B. We expect that such selections will become much less common as generative models improve. In the mean time, one benefit of the framework we propose is that once can visualize the generated counterfactuals. This allows one to easily check whether or not a selection is informative. We view this as an important advantage of the framework we propose: its interpretations are straightforward to interpret.

4.3 Interpreting a deep text classifier

We also apply the OSFT to interpret the Bidirectional Encoder Representations from Transformers (BERT) model [9] for text classification. BERT recently set a new state of the art in text classification performance on the GLUE benchmark [27]. It learns multiple layers of attention instead of a flat attention structure [26], making visualization of its internals complicated. A post-hoc black box interpretability method is therefore more appropriate to understand its predictions, especially since even flat attention visualizations often fail to correspond to model decision semantics [15].

We evaluate on the Large Movie Review Dataset (LMRD) [19], a corpus of movie reviews labeled as having either positive or negative sentiment and split into 25k training and 25k testing examples. We tokenize reviews into WordPieces [29], the sub-word level inputs to the BERT model, and test the significance of each WordPiece. To fit the reviews in memory, we restrict the training set to the 13k reviews that are under 256 WordPieces in length. We then tune a pretrained BERT model to perform sequence classification on this task, achieving 93.1% accuracy at test time. For all pretrained BERT models, we tune from `BERT-Base-Cased` and use the framework provided by `https://github.com/huggingface/pytorch-pretrained-BERT` to train task-specific models. Moreover, input sequences were trimmed to a maximum length of 128 WordPieces, and all training reviews were used for fine-tuning the language model.

For the uninformative conditional distribution, $Q(X_S|X_{-S})$, we use a seperate BERT instantiation for masked language modeling tuned on the LMRD training reviews. We set the FDR threshold α to 0.15 and test on 1000 randomly selected reviews of WordPiece length less than 256 from the test

Table 2: Text classifier interpretations using the OSFT. Selected words are highlighted.

Label	Model	Review
Neg	Neg	Stay away from this movie! It is **terrible** in every way. **Bad** acting, a thin recycled plot and the worst **ending** in film history. Seldom do I watch a movie that makes my adrenaline pump from irritation, in fact the only other movie that immediately springs to mind is another "people in an aircraft in trouble" movie (Airspeed). Please, please don't watch this one as it is utterly and totally **pathetic** from beginning to end. Helge Iversen
Pos	Pos	All i can say is that, i was expecting a wick movie and "Blurred" surprised me on the positive way. Very **nice** teenager movie. All this kinds of situations are normal on school life so all i can say is that all this reminded me my school times and sometimes it's good to watch this kind of movies, because entertain us and travel us back to those golden years, when we were young. As well, lead us to think better in the way we must understand our children, because in the past we were just like they want to be in the present time. Try this movie and you will be very **pleased**. At the same time you will have the guarantee that your time have not been wasted.
Pos	Neg	Not all movies should have that predictable ending that we are all so use to, and it's great to see movies with really unusual twists. However with that said, I was really **disappointed** in l'apartment's ending . In my opinion the ending didn't really **fit** in with the rest of the movie and it basically **destroyed** the story that was being told. **You** spend the whole movie discovering everyone and their feelings but the events in the final 2 minutes of the movie would have impacted majorly on everyones character but the movie **ends** and leaves it all too wide **open**. Overall **though** this movie was very well made, and unlike similar movies such as Serendipity all the scenes were believable and didn't go over the top.
Neg	Pos	**This** **is** **one** **entertaining** **flick**. I suggest you rent it, buy a couple quarts of rum, and invite the whole crew over for this one. My favorite parts were. 1. the gunfights that were so well choreographed that John Woo himself was jealous,. 2. The **wonderful** special effects. 3. the Academy Award winning acting and. 4. The fact that every single gangsta in the **film** seemed to be doing a **bad** "Scarface" impersonation. I mean, Master P as a cuban godfather! This is **groundbreaking** **territory**. **And** **with** **well** written dialogue including lines like "the **only** difference between you and me Rico, is I'm alive and your **dead**," this movie is **truly** a masterpiece. Yeah right.

set, for a total of 95518 WordPieces tested. We used the two-sided test statistic. About 4% of the WordPieces were selected.

Results. Table 2 visualizes both examples for which the model was correct and examples for which it was incorrect. WordPieces selected as important by the OSFT are highlighted. Intuitively, we find that high-sentiment words like "terrible", "pleased", "disappointed", and "wonderful" tend to be selected as important. Additional example interpretations and additional information about the experiments can be found in Appendix C.

5 Conclusion and future work

We proposed a general framework for reframing model interpretability as a multiple hypothesis testing problem. The framework mirrors the statistical analysis protocol employed by scientists: the null hypothesis test. Within this framework, we introduced the IRT and the OSFT, two hypothesis testing procedures for interpeting black box models. Both methods enable control of the false discovery rate at a user-specified level, making interpretations more reliable.

The methods we proposed in this paper have some important limitations. The most obvious limitation is that they require a way to generate plausible counterfactual inputs while keeping some features held fixed. However, this is already feasible for many types of distributions. For example, image inpainting is a subfield of computer vision that has a long history [13] and much recent work (e.g., [32, 31, 30, 28, 23, 33]), with plausible in-fill models available for many domains. As these generative models improve, so will the framework we proposed. Moreover, some deep language models, like BERT, are masked language models: they are trained, in part, to predict masked words. Consequently, to apply the IRT and OSFT to such models does not require a separate conditional model.

A practical problem that we did not describe in detail is how to choose subsets of features to test. We deliberately chose not focus on this, viewing it as a distraction from our core contributions, and a problem that is application-dependent. In some cases, it is straightforward to test all features individually. In other cases, especially in fields like histology and medical imaging, experts can manually select features of interest to test. In generic vision tasks, one can use an object detector or image segmentation model to select proposed regions. In generic language tasks, one can test over spans of a dependency tree or on selected parts-of-speech, depending on the application.

Finally, the only condition we claim is desirable for the conditional, $Q(X_S|X_{-S})$, is that it should always yield in-sample counterfactuals. However, it is possible that there are other desirable properties of such a conditional. We leave this investigation to future work.

Acknowledgments This work was supported by a seed grant from the Data Science Institute at Columbia University, NIH U54 CA193313. The authors thank Victor Veitch for helpful discussions, Dan Hendrycks for his valuable advice, and the anonymous reviewers for their useful feedback.

References

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian J. Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in Neural Information Processing Systems (NIPS)*, 2018.

[2] Rina Foygel Barber and Emmanuel J Candès. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085, 2015.

[3] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[4] Yoav Benjamini, Daniel Yekutieli, et al. The control of the false discovery rate in multiple testing under dependency. *The annals of statistics*, 29(4):1165–1188, 2001.

[5] Angel Cabrera, Fred Hohman, Jason Lin, and Duen Horng Chau. Interactive classification for deep learning interpretation. *arXiv preprint arXiv:1806.05660*, 2018.

[6] Emmanuel Candes, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold: 'Model-X' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B*, 2018.

[7] Chun-Hao Chang, Elliot Creager, Anna Goldenberg, and David Duvenaud. Explaining image classifiers by adaptive dropout and generative in-filling. *International Conference on Learning Representations (ICLR)*, 2018.

[8] Jianbo Chen, Le Song, Martin J Wainwright, and Michael I Jordan. Learning to explain: An information-theoretic perspective on model interpretation. *International Conference on Machine Learning (ICML)*, 2018.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[10] Eugene Edgington and Patrick Onghena. *Randomization tests*. Chapman and Hall/CRC, 2007.

[11] Ruth Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. *International Conference on Computer Vision (ICCV)*, 2017.

[12] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples. *International Conference on Learning Representations (ICLR)*, 2015.

[13] Christine Guillemot and Olivier Le Meur. Image inpainting : Overview and recent advances. *Signal Processing Magazine, IEEE*, 31:127–144, 2014.

[14] Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *International Conference on Learning Representations (ICLR)*, 2019.

[15] Sarthak Jain and Byron C. Wallace. Attention is not explanation. *North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019.

[16] Jason D Lee, Dennis L Sun, Yuekai Sun, and Jonathan E Taylor. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, 2016.

[17] Zachary Chase Lipton. The mythos of model interpretability. *ICML Workshop on Human Interpretability in Machine Learning (WHI)*, 2016.

[18] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[19] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, 2011.

[20] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?": Explaining the predictions of any classifier. *International Conference on Knowledge Discovery and Data Mining (KDD)*, 2016.

[21] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. *International Conference on Machine Learning (ICML)*, pages 3145–3153, 2017.

[22] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *International Conference on Learning Representations (ICLR) Workshop*, 2014.

[23] Ecem Soganciogl u, Shi Hu, Davide Belli, and Bram van Ginneken. Chest x-ray inpainting with deep generative models. *arXiv preprint arXiv:1809.01471*, 2018.

[24] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[25] Richard H Thaler. Anomalies: The winner's curse. *Journal of Economic Perspectives*, 2(1): 191–202, 1988.

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[27] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

[28] Yi Wang, Xin Tao, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. Image inpainting via generative multi-column convolutional neural networks. *Advances in Neural Information Processing Systems (NIPS)*, 2018.

[29] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, and Klaus Macherey. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

[30] Raymond A. Yeh, Chen Chen, Teck-Yian Lim, Mark Hasegawa-Johnson, and Minh N. Do. Semantic image inpainting with perceptual and contextual losses. *arXiv preprint arXiv:1607.07539*, 2017.

[31] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. *arXiv preprint arXiv:1806.03589*, 2018.

[32] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. Generative image inpainting with contextual attention. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[33] Liuchun Yuan, Congcong Ruan, Haifeng Hu, and Dihu Chen. Image inpainting based on patch-gans. *IEEE Access*, 7:46411–46421, 2019.

Algorithm 3 Benjamini–Hochberg (BH) correction

Require: α, empirical p-values $\hat{p}_1, \ldots, \hat{p}_K$
1: Sort the $\hat{p}_i$ in ascending order, yielding $\hat{p}^{(1)}, \ldots, \hat{p}^{(K)}$
2: Compute the largest k such that $\hat{p}^{(k)} \leq \frac{k}{K}\alpha$
3: **Return** $\tau := \hat{p}^{(k)}$

A Methodological details

A.1 Benjamini–Hochberg correction procedure

First, for the sake of completeness, in Algorithm 3 we provide the Benjamini–Hochberg [3] correction procedure that we use as MHT-Correct for the IRT in all experiments.

A.2 Correctness of the OSFT

Next, we prove that the OSFT controls the false discovery rate when the tests are independent.

Theorem 1. *(OSFT Correctness) Let* $z^{(i)} = t(f(x)) - t(f(\tilde{x}^{(i)}))$*, where* $\tilde{x}^{(i)} = (\widetilde{X}_i, x_{-i})$*, and* $\widetilde{X}_i \sim Q(X_i|x_{-i})$*. If the* $z^{(i)}$ *are independent, then rejecting the null hypotheses in the set* $\{H_0^{(i)} : z^{(i)} \geq z^*\}$ *controls the FDR of the point null at level* α*, if* z^* *is such that*

$$\frac{1 + \#\, z^{(i)} \leq -z^*}{\#\, z^{(i)} \geq z^*} \leq \alpha$$

Proof. The selection procedure in the OSFT and assumption on z^* are the same as for the knockoffs multiple testing procedure [2, 6]. As [6] note, FDR control using the knockoffs selection procedure is guaranteed at the α level as long as the sign of the difference statistics $z^{(i)}$ are i.i.d. coin flips under the null (following Theorems 1 and 2 of [2]). Under the point null for the ith feature, $\tilde{t}^{(i)} \stackrel{d}{=} t$. The distribution of $z^{(i)}$ under the null is therefore symmetric about the origin, so that the sign of every $z^{(i)}$ is indeed an independent coin flip. The claim then follows from [6]. □

B Additional image results

Figs. 5 to 8 show the remaining 44 examples of using the OSFT to interpret an image classifier on Imagenet. Below each image is the predicted class and the corresponding class probability.

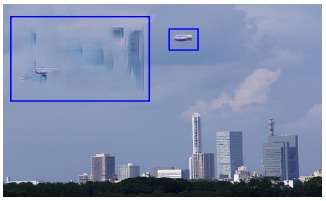
Moreover, we investigated some of the counterintuitive feature selections made by the OSFT, and found that in many cases they were due to poorly generated counterfactuals. We present an example of this in Fig. 4. Fig. 4c is unsurprisingly selected as important because it involves the features corresponding to an airship, which is the true class. We can again verify this by looking at the corresponding counterfactual, which is reasonably realistic. In contrast, in Fig. 4b, the features correspond to an arbitrary part of the sky. However, the generated counterfactual is unrealistic. That we are able to easily visualize and verify the output of the interpretability method is an advantage of the framework we propose, even for cases like this in which it produces an uninformative interpretation.

C Additional language results

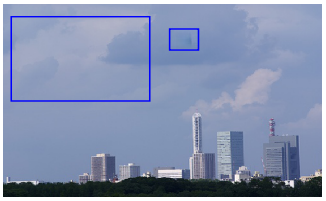
Table 3 shows additional examples of using the OSFT and the two-sided test statistic to interpret the trained BERT model for sentiment classification. As before, the selected WordPieces are highlighted.

Some of language interpretations call attention to the poor counterfactuals sampled for words around sentence boundaries. Language models are typically trained at the sentence level, meaning sentence boundaries lack either left- or right-context. BERT models begin to address this gap by training with pairs of sentences, but still suffer high perplexity at boundaries. This may account for some words that were selected counterintuitively in the example reviews.

(a) $p = 0.970$ (Airship)



(b) $p = 0.811$ (-0.159)

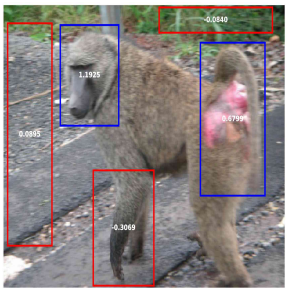


(c) $p = 0.052$ (-0.918)

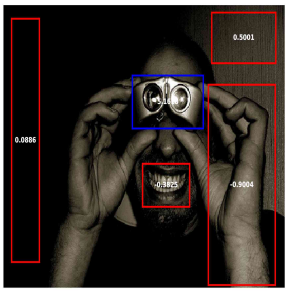
Figure 4: An interpretation illustrating both the benefits and limitations of the approach presented in this work. The subsets of features replaced in both Figs. 4b and 4c were selected as important. (See Fig. 7a.) Fig. 4b illustrates how a poor counterfactual can lead to a (probably) unwarranted selection by the interpretability method, while Fig. 4c illustrates a reasonable counterfactual that shows the importance of the corresponding features in an interpretable way.

Table 3: Additional text classifier interpretations using the OSFT.

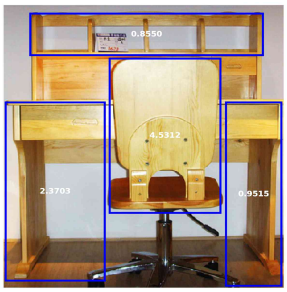
Label	Model	Review
Neg	Neg	the photography is good, the costumes are good, but the editing is bad. The various scenes are cut, or stopped at the wrong times, and the conversations are s-l-o-w and tedious. This **slow**ness **continues** the **entire** show. It is a very **tedious** show to watch.... I believe that more scenes SHOULD HAVE BEEN ADDED, but that would make it a longer show. It is very slow-moving. The writers should have made it JUST a **1**-night show, and not prolong our agony night after night. There is nothing else on, otherwise I'd change the channel (the first night). I feel bad for the **Indians** of the time, and am angry at the white-men for what they did to the Indians, but thats our history.
Neg	Neg	Bathebo, you big dope. This is the WORST piece of **crap** I've seen in a long time. I have just stumbled onto it on late night TV and it is painful to watch. Really painful. How does something like this get made?? Horrible, horrible, horrible! OOOOOO..... The toilet is flushing by itself again! Scary toilet! Scary toilet! Scary toilet! 1992 doesn't seem like that long ago to me, but watching this makes it seem like 1952. I mean its **horrible**. Please don't waste your time on the drivel! Scary old black man telling them not to build the pool in the yard. Scary! Scary! How does this stuff get MADE???
Pos	Pos	Two funeral directors in a Welsh village? English **humour** as opposed to that other stuff from over the Atlantic? **How** could **I resist.** My wife and I saw it on March the 6th for our belated Valentines day celebration and both of us enjoyed some good belly laughs. We were going to see another movie later but decided not to because we wanted the experience of THIS movie to stay with us for the evening. The mortuary scene in the last 20 minutes of the film is worth the **wait**. It raises issues rarely talked about in the community, but I know three funeral directors, and the humour **is** right on **the money** **High**ly **recommended** **and** congratulations to the writers. **Without** you **all** the actors, directors and the others havn't a job on any Monday**.**
Pos	Pos	**Evil** warlord puts a town through pain and **suffering**. Not long before they call upon giant stone samurai Daimaijin for help. Daimaijin **soon** **comes** and **really** **gets** the warlord with all his viscious might. The **revenge** **climax** is really funny as Daimajin squashes guys under his feet and crushes guys with his fist and even drives a spike though a man's heart.
Pos	Neg	I was shocked to learn that Jimmy Caan has left this show, does anyone know why? I **regard** James as one of the all-time greats and wasn't surprised he ended up on TV, which **can** be better than the crap you see on the big screen. The stories are **slick** and the **camera** **faster** than a **speeding** bullet! Mustn't **forget** the rest of the cast: James, Vanessa (yum!) Nikki, Molly, Josh, Mitch.. **Also**, can anyone tell me why on earth there's a **crap** **theme** tune on the DVD sets, but Elvis's JXL remix of A Little Less Conversation is used on the initial NBC broadcasts**?**? **Does** it not make **sense** to use a tune that you would associate with the gambling mecca of America for DVD releases?**?**
Pos	Neg	We just saw this film previewed before release at the Norfolk (VA) Film Forum, and there was general agreement on two matters: **There** were excellent performances in **a** first rate drama by the two leads and by others: and secondly, the **marketing** for this movie will only bring **disaster**. We saw a lur**i**d poster with chains and suggestive commentary implying some sort of wacko sexual relationship between Samuel Jackson and Cristina Ricci, **whereas** the **movie** has some **real** **depth** **and** some **thoughtful** ideas. What's sad is that people looking for near porn will be drawn in to see the film and will be disappointed because it will be too "heavy" for them, while the people who would really enjoy it wouldn't be caught dead walking into the theater showing it. Too bad. A **good** **film** **wasted**.
Neg	Pos	**Based** on a Stephen King novel, NEEDFUL THINGS **provides** the intrigue and eeriness to keep you in your seat. A mysterious man (Max von Sydow) comes to town and soon becomes the most talked about citizen. **Could** it be that the devil himself has set up shop as an antique dealer in a small town in Maine**?** von Sydow is **master**ful and dynamic in this role that dominates the screen. Also starring are Ed Harris and Bonnie Bedelia. Harris is **steady** and Bedelia is deserving of your attention. Also in support are J.T. Walsh and Amanda Plummer. **Not** the best**,** **nor** **the** **worst** **adaptation** of King's horror on the **screen**.
Neg	Pos	Technically abominable (**with** **audible** "pops" between **scenes**) and **awesome**ly **amateur**ish, "Flesh" requires a lot of patience to sit through and will probably turn off most viewers; but the **dialogue** rings **amazing**ly true and Joe Dallesandro, who exposes his body in almost every scene, also gives an utterly **convincing** performance**.** A cu**rio**, to be sure, but the more **polished** "Trash", made two years later, is a definite step forward. I suggest you watch that **instead**. (*1/2)



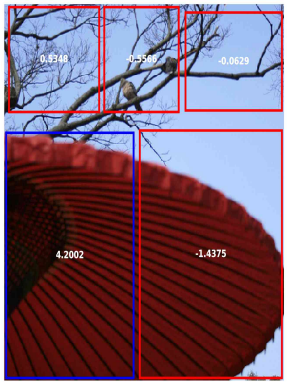
(a) Baboon ($p = 0.935$)



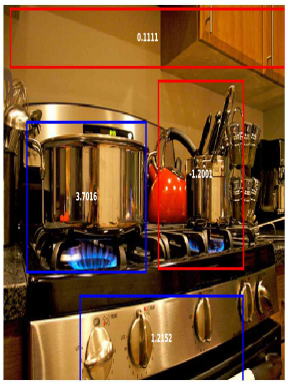
(b) Binoculars ($p = 0.920$)



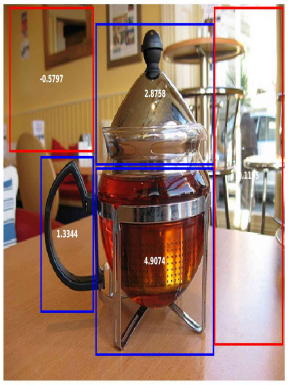
(c) Desk ($p = 0.695$)



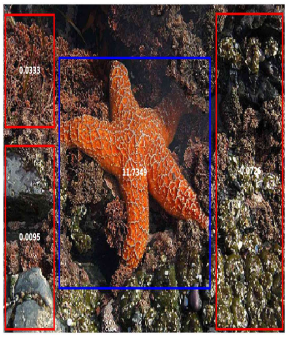
(d) Umbrella ($p = 0.925$)



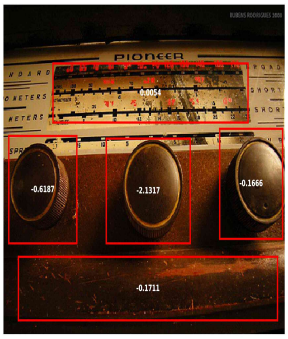
(e) Stove ($p = 0.671$)



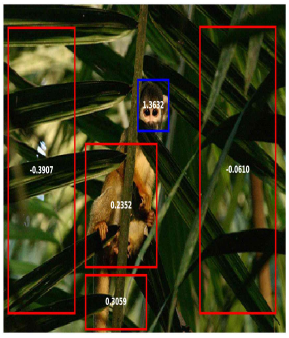
(f) Teapot ($p = 0.819$)



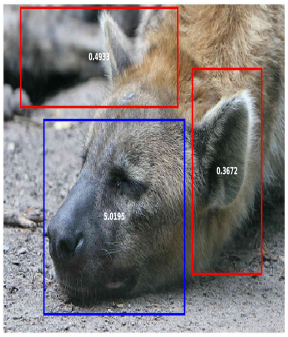
(g) Starfish ($p = 0.995$)



(h) Radio ($p = 0.762$)

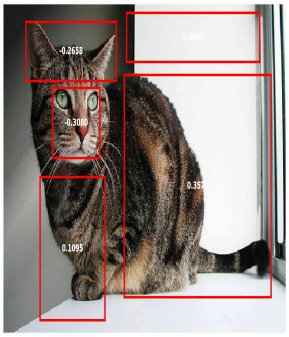


(i) Squirrel monkey ($p = 0.689$)



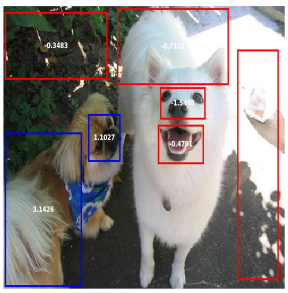
(j) Hyena ($p = 0.961$)

(k) Toaster ($p = 0.992$)

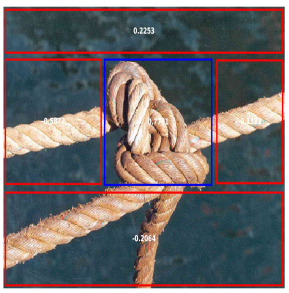


(l) Tabby cat ($p = 0.922$)

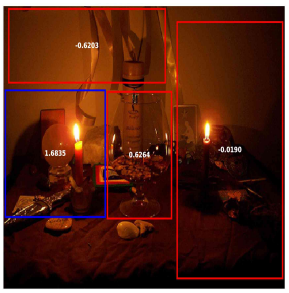
Figure 5



(a) Pekinese ($p = 0.484$)

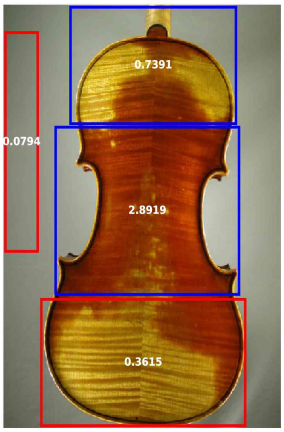


(b) Knot ($p = 0.994$)

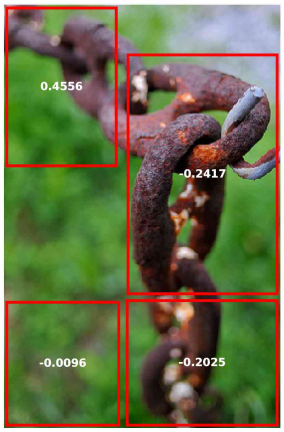


(c) Altar ($p = 0.791$)

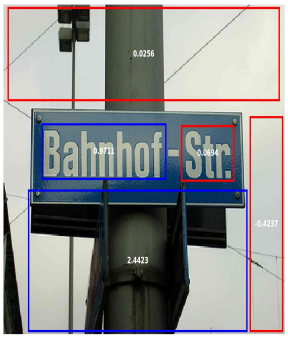
(d) Gown ($p = 0.863$)



(e) Violin, fiddle ($p = 0.991$)

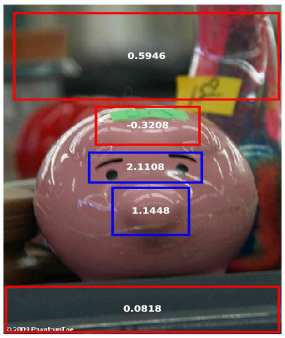


(f) Binoculars ($p = 0.679$)

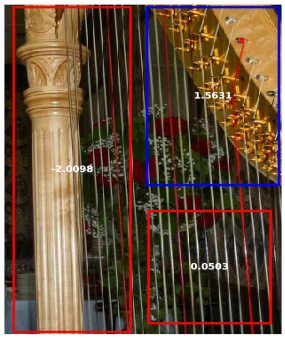


(g) Street sign ($p = 0.883$)

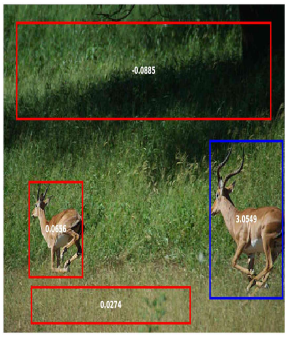
(h) Menu ($p = 0.924$)



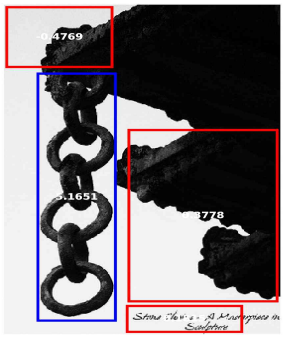
(i) Piggy bank ($p = 0.930$)



(j) Harp ($p = 0.999$)

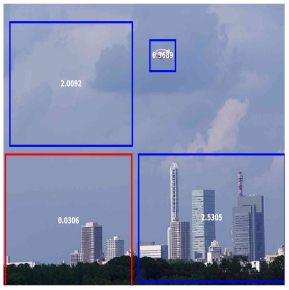


(k) Impala ($p = 0.679$)

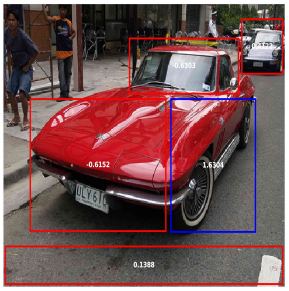


(l) Hook, claw ($p = 0.431$)

Figure 6

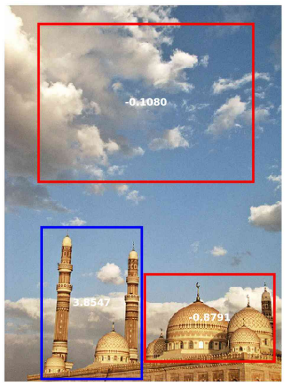


(a) Airship, dirigible ($p = 0.969$)

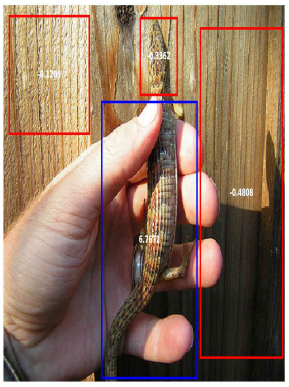


(b) Convertible ($p = 0.619$)

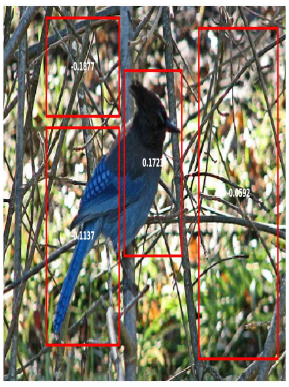
(c) Stinkhorn, carrion fungus ($p = 0.803$)



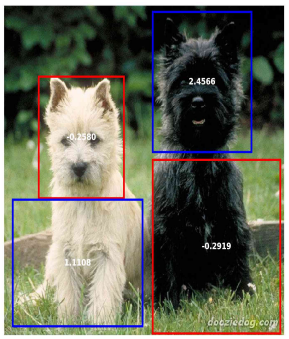
(d) Mosque ($p = 0.749$)



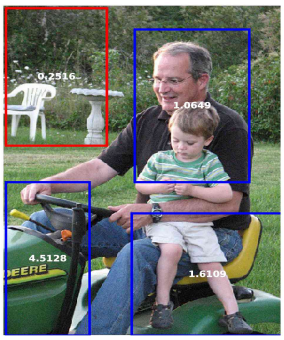
(e) Alligator lizard ($p = 0.954$)



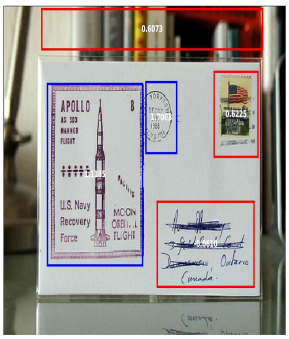
(f) Jay ($p = 0.961$)



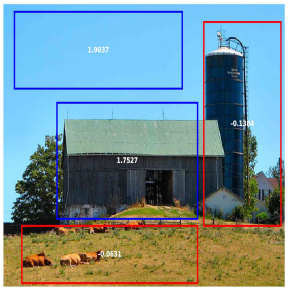
(g) Cairn terrier ($p = 0.816$)



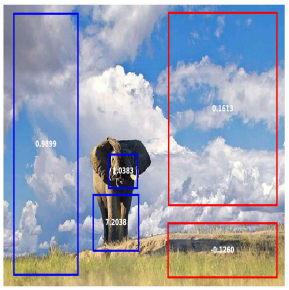
(h) Lawn mower ($p = 0.923$)



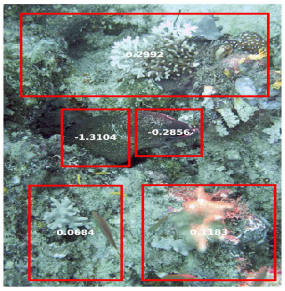
(i) Envelope ($p = 0.059$)



(j) Barn ($p = 0.968$)

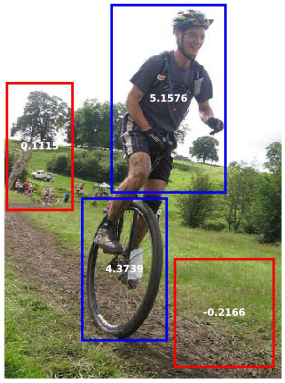


(k) African elephant ($p = 0.896$)

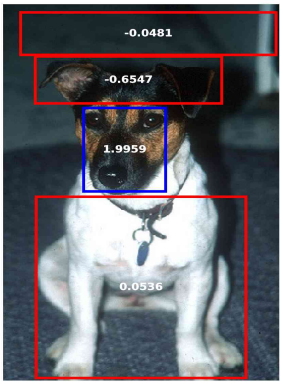


(l) Coral reef ($p = 0.502$)

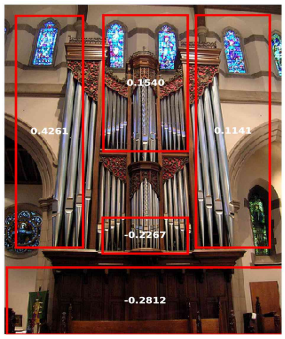
Figure 7



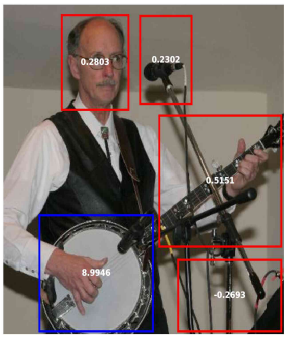
(a) Unicycle ($p = 0.996$)



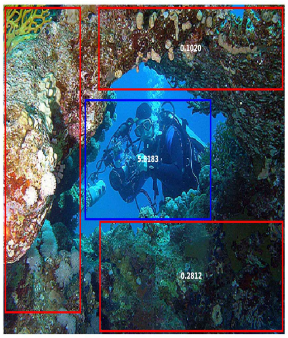
(b) Toy Terrier ($p = 0.927$)



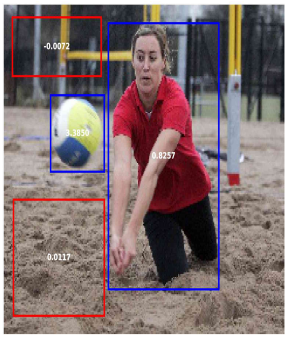
(c) Organ ($p = 0.959$)



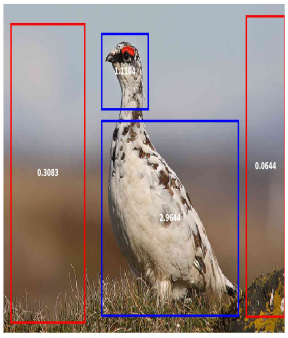
(d) Banjo ($p = 0.997$)



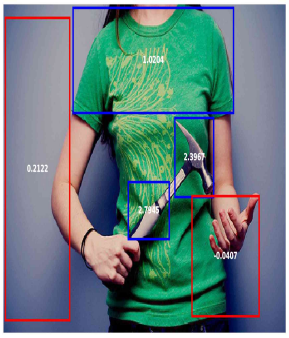
(e) Scuba diver ($p = 0.945$)



(f) Soccer ball ($p = 0.465$)



(g) Ptarmigan ($p = 0.871$)



(h) Hammer ($p = 0.695$)

Figure 8