Interpreting Black Box Models with Statistical Guarantees

Collin Burns [1] **Jesse Thomason** [2] **Wesley Tansey** [3 4]

Abstract

While many methods for interpreting machine learning models have been proposed, they are frequently ad hoc, difficult to evaluate, and come with no statistical guarantees on the error rate. This is especially problematic in scientific domains, where interpretations must be accurate and reliable. In this paper, we cast black box model interpretation as a 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 randomly-sampled counterfactuals. We derive a multiple hypothesis testing framework for finding important features that enables control over the false discovery rate. We propose two testing methods, as well as analogs of one-sided and two-sided tests. In simulation, the methods have high power and compare favorably against existing interpretability methods. When applied to vision and language models, the framework selects features that intuitively explain model predictions.

1. Introduction

Scientists are increasingly applying black box models, such as deep neural networks, to analyze their data. Medical researchers have used black box models to assess cancer progression (Schaumberg et al., 2018), discover immunotherapy targets (Bobisse et al., 2016), classify Alzheimer's disease status (Sarraf & Tofighi, 2016), and detect diabetic retinopathy (Gulshan et al., 2016). Biologists have used black box models that take DNA and RNA sequences as input to predict cellular processes like regulation (Alipanahi et al., 2015; Kelley et al., 2016) and splicing (Xiong et al., 2015). Chemists have adopted black box models to perform virtual compound screenings and materials design (Goh et al., 2017). A common theme in all of these cases is the need for scientists to base decisions on predictions made out-of-sample by opaque models.

When making a decision, the scientist must be able to audit the black box model to verify that its reasoning matches prior knowledge. At a minimum, this means understanding which features are influencing the prediction. In doing so, we should have strong theoretical guarantees on the interpretations of these models. 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 and robust.

For example, an oncologist may use a black box model that predicts a personalized course of treatment from tumor sequencing data. For the physician to trust the recommendation, it should come with a list of genes explaining the prediction. Those genes can then be cross-referenced with the research literature to verify their association with response to the recommended treatment. However, gene expression data is highly correlated. If the interpretability method does not consider the dependency of different genes, it may report many false positives. This may lead the oncologist to believe the model is incorrectly analyzing the patient, or, worse, to believe the model identified cancer-driving genes that it actually ignores.

In this paper, we address the need for reliable interpretation methods by treating black box model interpretation as a multiple hypothesis testing problem. Given a black box model and an input of interest, we test subsets of features of that input to determine which subsets were collectively "important" for the prediction. Statistically, this corresponds to testing the null hypothesis that the model output was distributed according to an uninformative counterfactual distribution. Specifically, the counterfactual is the output distribution when the features of interest are sampled according to an "uninteresting" but plausible conditional distribution, with the other features held fixed. We derive a multiple hypothesis testing framework based on the ability to sample suitable counterfactuals.

Within this framework, we propose two hypothesis testing methods: the Interpretability Randomization Test (IRT) and

[1]Department of Computer Science, Columbia University, New York, NY, USA [2]Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, WA, USA [3]Data Science Institute, Columbia University, New York, NY, USA [4]Department of Systems Biology, Columbia University Medical Center, New York, NY, USA. Correspondence to: Wesley Tansey <wesley.tansey@columbia.edu>.

the One-Shot Feature Test (OSFT). Both methods provably control the false discovery rate at a specified level, and we find that both methods have similarly high power in synthetic experiments. However, while the IRT can test a single subset of features at a time, it has comparatively high computational cost. In contrast, the OSFT requires testing many subsets of features at once, but is much more efficient than the IRT when doing so.

The methods are also flexible, supporting any choice of test statistic for any black box model. We propose two specific test statistics as analogs of classical one-sided and two-sided hypothesis tests. We show in simulation that these tests empirically have higher power than other popular black box interpretability methods for fixed levels of FDR control. As a case study, we apply the One-Shot Feature Test to interpret the predictions of state-of-the-art vision and language models on benchmark classification tasks.

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.

2.1. Interpreting machine learning models

Most methods for interpreting model predictions take an optimization approach. Gradient-based methods (Simonyan et al., 2014; Selvaraju et al., 2016; Sundararajan et al., 2017) 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 (Ribeiro et al., 2016), SHAP (Lundberg & Lee, 2017), and L2X (Chen et al., 2018). 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.

Other methods for interpretability are based on counterfactuals. Fong & Vedaldi (2017) 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, it is unclear how to interpret the conclusions. Cabrera et al. (2018) introduce an interactive setup for interpreting image classifiers in which users select regions of a given image to in-paint (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. (2018) similarly use in-painting models, mitigating the problem of using out-of-distribution counterfactuals, but like Fong & Vedaldi (2017) 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 (Lipton, 2016; Adebayo et al., 2018). Optimization may also overestimate the importance of some variables due to the winner's curse (Thaler, 1988). 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 (Lee et al., 2016). The methods proposed in this paper avoid this issue by taking a multiple hypothesis testing approach.

2.2. Multiple testing and false discovery rate control

In multiple hypothesis testing (MHT), $\mathbf{z} = (z_1, \ldots, z_N)$ are a set of independent 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 (TPR), also known as *power*,

$$\text{TPR} \coloneqq \mathbb{E}\left[\frac{\#\{i : i \in \hat{\mathcal{S}} \cap \mathcal{S}\}}{\#\{i : i \in \mathcal{S}\}}\right], \tag{1}$$

while controlling the false discovery rate (FDR)–the expected proportion of the predicted discoveries that are actually false positives,

$$\text{FDR} \coloneqq \mathbb{E}\left[\frac{\#\{i : i \in \hat{\mathcal{S}} \backslash \mathcal{S}\}}{\#\{i : i \in \hat{\mathcal{S}}\}}\right]. \tag{2}$$

FDR is the typical error measure targeted in modern scientific analyses. Methods that control for 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 FDR in the reported set of important features that contributed towards a model's prediction.

3. Interpretability as Hypothesis Testing

While there is no agreed upon definition of interpretability (Lipton, 2016), intuitively a feature should be considered "important" if its impact on the model output is surprising relative to some counterfactual, given the context of all other features. We formalize this by testing whether the given model output was sampled from the distribution of outputs in which the feature of interest was resampled from some uninformative counterfactual distribution. If the observed model output is extreme with respect to this distribution, then intuitively the feature must have had a substantial influence on the model.

3.1. Setup

Consider a supervised learning problem for which we have a distribution $P(X, Y)$ over features and labels, where $X = (X_1, \ldots, X_d) \sim P(X|Y)$. Suppose we want to understand the output of a model f_θ on the input $x \in \mathbb{R}^d$ that was sampled from the marginal distribution $P(X)$. Let $\hat{y} := f_\theta(x)$. For $X \in \mathbb{R}^d$ and $S \subset [d] := \{1, \ldots, d\}$ use the notation X_S to denote X restricted to the set S, and X_{-S} to denote X restricted to the features not in S.

Definition 1. *We say that a subset of features $S \subset [d]$ on input x is important with respect to the conditional distribution $Q(X_S|X_{-S})$ if the following null hypothesis, H_0, is false:*

$$H_0 \colon \hat{y} \sim f_\theta(\widetilde{X}) \,. \tag{3}$$

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

While this definition applies to any conditional distribution $Q(X_S|X_{-S})$, it is only a meaningful notion of interpretability for some distributions. One natural such distribution is the complete conditional, $P(X_S|X_{-S})$, where the true label Y is marginalized out. Alternatively, Q could be an approximation to the complete conditional, in which case the same guarantees hold, but with a slightly different notion of importance. Moreover, in some cases there is another natural notion of a background or uninformative distribution; we give such an example in Section 4.1. Crucially, the conditional should be such that the generated inputs, $\widetilde{X} = (\widetilde{X}_S, x_{-S})$, are in the support of the true distribution $P(X)$. This is important because the model has been trained only on inputs from $P(X)$. As work on robustness and adversarial examples illustrates (Goodfellow et al., 2015), model behavior on out-of-distribution inputs can be unreasonable, making the definition of importance with respect to such a distribution potentially misleading.

Algorithm 1 Interpretability Randomization Test (IRT)

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

1: Compute model output $\hat{y} \leftarrow f_\theta(x)$
2: Compute test statistic $t \leftarrow T(\hat{y})$
3: **for** $i \leftarrow 1, \ldots, N$ **do**
4: **for** $k \leftarrow 1, \ldots, K$ **do**
5: Sample $\widetilde{x}_{S_i} \sim Q(X_{S_i}|X_{-S_i} = x_{-s_i})$
6: Compute model output $\tilde{y}^{(k)} \leftarrow f_\theta((\widetilde{x}_{S_i}, x_{-S_i}))$
7: Compute the test statistic $\tilde{t}^{(k)} \leftarrow T(\tilde{y}^{(k)})$
8: **end for**
9: Compute the p-value

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

10: **end for**
11: Sort the $\hat{p}_i$ in ascending order, yielding $\hat{p}^{(1)}, ..., \hat{p}^{(K)}$
12: Compute the largest k such that $\hat{p}^{(k)} \leq \frac{k}{K}\alpha$
13: **Return** discoveries at the α level: $\{i \colon \hat{p}_i \leq \hat{p}^{(k)}\}$

3.2. The Interpretability Randomization Test

In general, the null distribution $f_\theta(\widetilde{X})$ will not be available in closed form, but if we have access to $Q(X_S|X_{-S})$ then we can repeatedly sample new inputs, get the corresponding model outputs, 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 represents a valid p-value for finite samples from H_0 (Edgington & Onghena, 2007), meaning it is stochastically greater than the $Uniform(0, 1)$ distribution.

When testing multiple features, controlling the error rate requires applying a multiple hypothesis testing correction procedure. We focus on controlling the false discovery rate. Given a set of p-values from Algorithm 1 and a target FDR threshold α, we use the Benjamini-Hochberg (Benjamini & Hochberg, 1995) correction technique to control the FDR at the α level (lines $11 - 13$ of Algorithm 1).

3.3. The One-Shot Feature Test

The IRT enables testing any subset of features for importance. In practice, both the black box predictive model and the conditional model may be computationally expensive. In these cases, it may be prohibitively expensive to run multiple trials for each feature. As an efficient alternative, we

Algorithm 2 One-Shot Feature Test (OSFT)

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

Compute test statistic $t \leftarrow T(f_\theta(x_1, \ldots, x_d))$
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 $\tilde{y}^{(i)} \leftarrow f_\theta((\widetilde{x}_{S_i}, x_{-S_i}))$
  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: $\{i\colon z^{(i)} \geq z^*\}$

propose a one-shot testing procedure that only requires a single extra evaluation from the conditional and black box model for each subset of features being tested.

Consider testing N non-overlapping subsets of features $S_1, ..., S_d \subset [d]$ for importance, given $x \in \mathbb{R}^d$ and $\hat{y} = f_\theta(x)$. For each S_i, we wish to test the null hypothesis $H_0^{(i)} : y \sim f_\theta(\widetilde{X}^{(i)})$ where $\widetilde{X}^{(i)} = (\widetilde{X}_{S_i}, x_{-S_i})$ and $\widetilde{X}_{S_i} \sim Q(X_{S_i}|X_{-S_i} = x_{-S_i})$. Rather than building an explicit null distribution, all features can be tested simultaneously by leveraging the fact that the difference statistics are all symmetric about the origin.

Theorem 1. *Let $z^{(i)}$ be the difference between the test statistic from the original input and the test statistic for a single draw from the i^{th} null. Rejecting the null hypotheses in the set $\{H_0^{(i)}\colon z^{(i)} \geq z^*\}$ controls the false discovery rate 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 Theorem 1 and assumption on z^* are the same as for the knockoffs multiple testing procedure (Barber & Candès, 2015; Candes et al., 2018). As Candes et al. (2018) 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 Barber & Candès (2015)). Under $H_0^{(i)}$, $\hat{y}$ has the same distribution as $\tilde{y}^{(i)}$. Since the test statistic is a function of the model output, this implies $\tilde{t}^{(i)} \overset{d}{=} t$. The distribution of $z^{(i)}$ under the null is therefore symmetric about the origin. Hence, $P(z^{(i)} = z) = P(z^{(i)} = -z)$ so that the sign of every $z^{(i)}$ is an independent coin flip. □

Algorithm 2 presents the complete One-Shot Feature TEST (OSFT). Unlike the IRT, the OSFT requires multiple subsets of features to be tested at once, so it is most appropriate for scenarios where a sample contains many features. Alternatively, if an entire dataset is being evaluated by a model, multiple samples can be interpreted simultaneously.

3.4. Test statistic choice

Both of the proposed algorithms above guarantee FDR control under any choice of test statistic. Different choices may be more appropriate for certain tasks and may have higher power. A common design decision in hypothesis testing is whether to perform a one-sided or two-sided test. One-sided tests have a preferred direction of testing. That is, one-sided tests ignore a statistic if it falls on the wrong side of the null, even if it is an unlikely value. Two-sided tests consider both tails of the null distribution. Below, we propose analogs of one-sided and two-sided tests for model interpretation.

One-sided test statistic. The scientist may wish to test which features cause a model to change its output in a specific direction. For example, if a model is performing a multi-class classification, the scientist may wish to inspect which features are associated with an increase in probability for a specific class, while features that decrease probability may not be of interest.

Testing for an increase in output can be done by making the test statistic the identity, $T(Y) = Y$, which is used in all one-sided experiments in Section 4. (Note that testing for a decrease can be done by instead letting $T(Y) = -Y$.) If a feature raises the response of the model, $T(Y) = Y$ will be higher than under the null and the p-value and z-statistics from the two testing algorithms will both be lower. For the IRT, this is straightforward and will not impact model power more than if the lowering features were true null features. For the OSFT, power under this statistic depends on the ratio and magnitude of the feature impacts. In the worst case, if as many features lower the model output by as much as the features that raise it, the one-shot procedure may have zero power. Section 4 investigates this in simulation and on state-of-the-art image and text classifiers, finding that the OSFT with the one-sided test still has high power in practice.

Two-sided test statistic. The scientist may be interested in finding which features have any impact on the model output, regardless of the direction. One approach for doing so would be to modify the IRT to change it from returning a one-sided p-value to a two-sided p-value by looking at both tails of the distribution of $\tilde{t}$. In the one-shot case, there is no explicit null distribution for a given sample; there is only a guarantee of symmetry and centering at zero.

Carrying out a two-sided test can be done by drawing an

extra null variable as a "centering" sample,

$$\begin{aligned} \bar{X}_i \sim \quad & Q(X_i | X_{-i} = x_{-i}) \\ \bar{Y} = \quad & f_\theta(\bar{X}_i, x_{-i}) \\ T(Y) = \quad & (Y - \bar{Y})^2 . \end{aligned} \tag{4}$$

This turns the one-shot procedure into a two-shot procedure. Section 4 demonstrates the two-sided test on a text sentiment classification model.

4. Experiments

We first compare the IRT and OSFT to three previous black box interpretability methods on a synthetic dataset. We then conduct two case studies applying the OSFT to explain the predictions of a state-of-the-art image classifier (Inception) on Imagenet and a state-of-the-art sentiment classifier (BERT) on movie reviews.

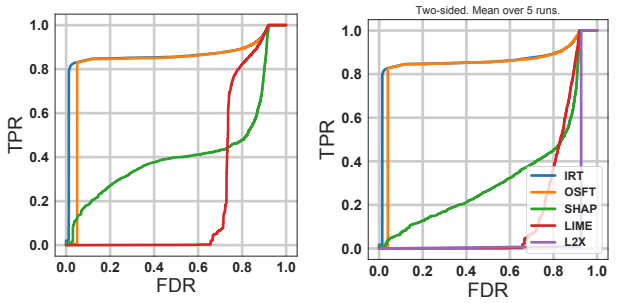
4.1. Synthetic benchmark

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 of these methods. The benchmark task is interpreting a nonlinear 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 weights $w \in \mathbb{R}^p$ and a fixed threshold $t \in \mathbb{R}$. We let $w_i \overset{\text{IID}}{\sim} 0.5 + \text{Gamma}(1, 1)$, and fix $t = 3$. Each feature X_i is drawn IID from a Gaussian mixture model: with probability 0.3 it is sampled from standard normal distribution $\mathcal{N}(0, 1)$, and with probability 0.7 it is sampled from $\mathcal{N}(4, 1)$. Intuitively, this is a simple model for more complicated distributions: a feature is either interesting (e.g., in the case of images it may contain an object related to the underlying label, Y), or it is not, and it is sampled from a different distribution accordingly. In this case, $\mathcal{N}(0, 1)$ serves as a natural background distribution that we can use as our conditional $Q(X_S | X_{-S})$; in particular, its sampled inputs remain well within the support of the true marginal distribution $P(X)$. For each feature $i \in [2p]$, the null hypothesis is that $\hat{y} = f_\theta(x)$ was sampled from the distribution $f_\theta(\widetilde{X}^{(i)})$ where $\widetilde{X}^{(i)} = (\widetilde{X}_i, x_{-i})$ and $\widetilde{X}_i \sim Q(X_i | X_{-i} = x_{-i})$. When $i \in [p]$, this is false if and only if $|x_{i+p}| \geq t$ and x_i was sampled from the interesting distribution $\mathcal{N}(4, 1)$. Similarly, x_{i+p} is important if and only if $|x_i| \geq t$ and $x_{i+p} \sim \mathcal{N}(4, 1)$. We run 5 independent trials, each with 100 examples. We set $p = 50$, so the number of parameters is 100.

We compare against three other methods from the black box interpretability literature:



(a) One-sided test. (b) Two-sided test.

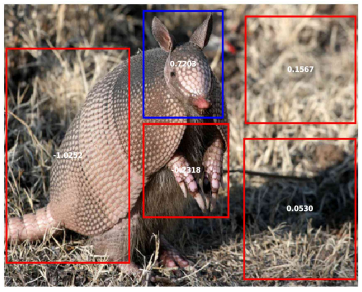
Figure 1. Both the IRT and OSFT have significantly higher power than the baseline methods for nearly all FDR values of interest. Results are averaged over 5 independent trials.

- **LIME** (Ribeiro et al., 2016) builds a linear approximation of the predictive model and uses the coefficients as an importance weights.
- **SHAP** (Lundberg & Lee, 2017) takes a game theoretic approach to importance (Shapley values).
- **L2X** (Chen et al., 2018) optimizes a variational lower bound on the mutual information between the label and each feature.

SHAP and LIME both output real-valued feature importance scores instead of selecting features as important or unimportant. L2X, on the other hand, directly selects k features to explain a prediction, where k is treated as a hyperparameter. To compare these to the IRT and OSFT, which automatically choose a number of features to select, we suppose that SHAP, LIME, and L2X are able to control the false discovery rate at a particular level, and measure the true positive rate at that level. In particular, we see how the FDR and TPR change as these methods smoothly increase the number of features they select.

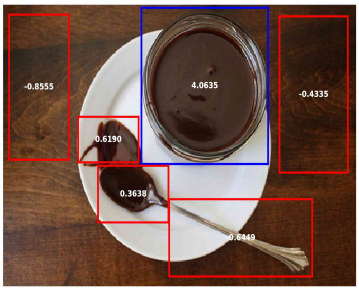
We consider one-sided and two-sided variants for SHAP and LIME. For the one-sided test, we track how the TPR and FDR vary as the k features with the largest values are selected for increasing k. For the two-sided variant, we instead select the k features with the largest absolute values. On the other hand, L2X directly selects features that are broadly relevant to the output of the model. This limits L2X to only the two-sided case.

Results. Fig. 1 shows the TPR of each method as a function of the FDR level. The IRT and OSFT have substantially higher power than the baseline methods in both experiments, with the power of the IRT better than that of the OSFT. LIME and L2X have near-zero power at small FDR thresholds.

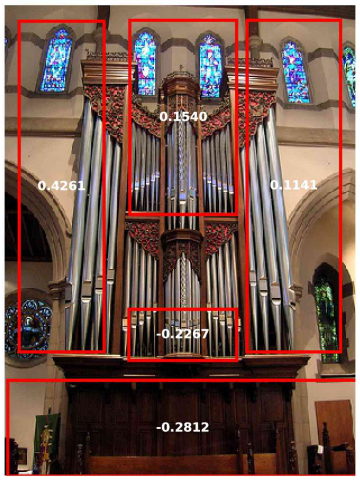


(a) Armadillo ($p = 0.970$)

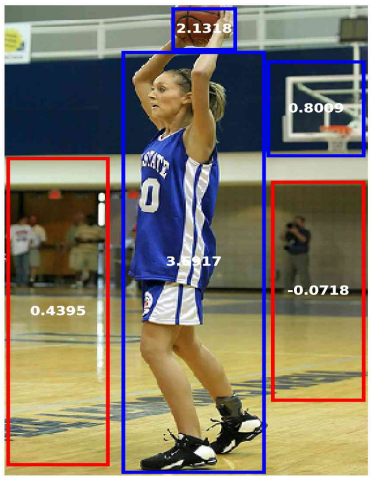
(b) Saxaphone ($p = 0.993$)



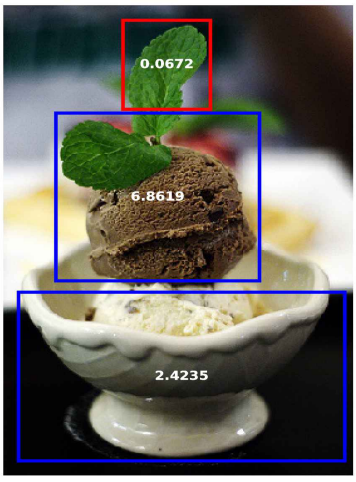
(c) Chocolate sauce ($p = 0.995$)



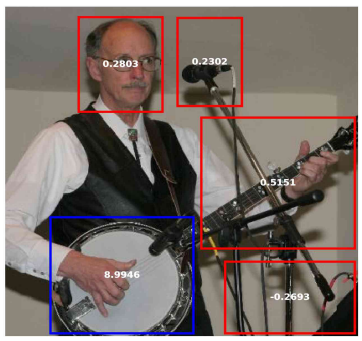
(d) Organ ($p = 0.959$)



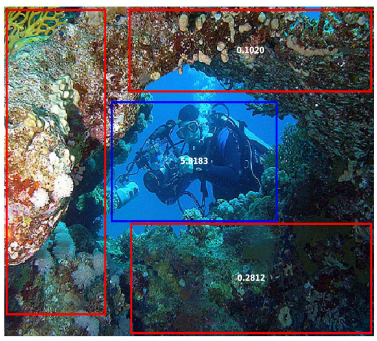
(e) Basketball ($p = 0.974$)



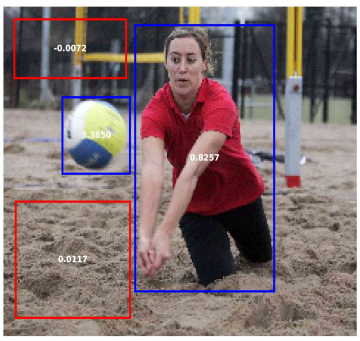
(f) Ice cream ($p = 0.923$)



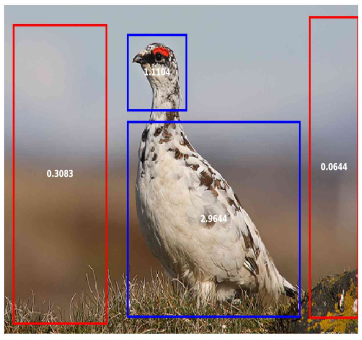
(g) Banjo ($p = 0.997$)



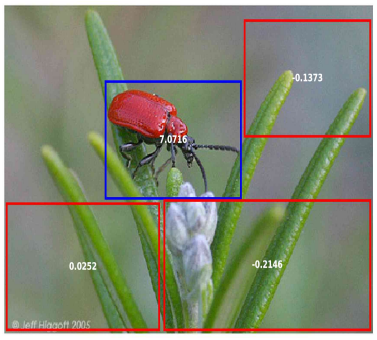
(h) Scuba diver ($p = 0.945$)



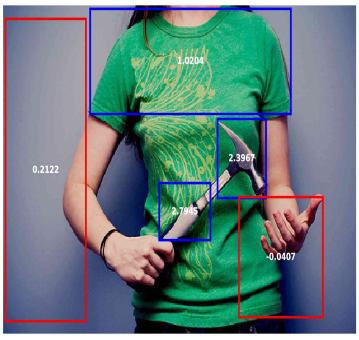
(i) Soccer ball ($p = 0.465$)



(j) Ptarmigan ($p = 0.871$)



(k) Leaf beetle ($p = 0.950$)



(l) Hammer ($p = 0.695$)

Figure 2. Image classifier interpretations using the one-sided test statistic $T(Y) = Y$. The threshold for rejecting the null hypothesis is 0.6799. 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.

4.2. Interpreting a deep image classifier

We apply the OSFT to interpreting Inception (Szegedy et al., 2015), a deep image classifier. To approximate sampling from the complete conditional distribution $P(X_S|X_{-S})$, we use a state-of-the-art generative inpainting model (Yu et al., 2018). We define the model output to be the logits for the predicted class, and use the one-sided statistic $T(Y) = Y$, 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 instance, if the image is a microscopy slide, a cell segmentation model (e.g., Kraus et al., 2016) could be used to automatically select individual cells for testing. Alternatively, a histologist could examine the slide and select candidate regions. For simplicity here, the patches were selected manually.

We test 50 Imagenet images, some of which were taken from Fong & Vedaldi (2017) 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 2 shows 12 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); the value of the difference statistic is printed inside the patch. The remaining 38 images and the selected patches for each can be found in the appendix.

In order to test whether the model primarily relies on a small portion of the true object, in several cases the bounding boxes were selected to cover only part of that object. This tends to lead to the covered portion being labeled as not important. Intuitively, this is because a patch is only considered important if it covers a sufficiently large portion of the region that that the model uses and that cannot be recovered from the rest of the image. For example, in Fig. 2d, in order to test whether the model primarily relies on just one part of the organ, or if it mistakenly uses part of the surrounding background, none of the bounding boxes cover the full organ. Because of this, it is unsurprising that no bounding box is selected as important.

4.3. Interpreting a deep text classifier

As a second case study, we apply the OSFT to interpret the Bidirectional Encoder Representations from Transformers (BERT) model (Devlin et al., 2018) for text classification. This model recently set a new state of the art in text classification performance on the GLUE benchmark (Wang et al., 2018). BERT learns multiple layers of attention instead of a flat attention structure (Vaswani et al., 2017), making visualization of its internals complicated. Consequently, a post-hoc black box interpretability method is necessary to understand its predictions.

We evaluate on the Large Movie Review Dataset (LMRD) (Maas et al., 2011), 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 (Wu et al., 2016), the subword 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.

We approximate the complete conditional distribution using 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 set, for a total of 95518 WordPieces tested. We used the two-sided test statistic from Eq. (4), which selected about 4% of the WordPieces.

Results. In Table 1, we visualize both examples for which the model was correct and examples for which it was incorrect, highlighting WordPieces selected as important by the OSFT. More example review visualizations can be found in the appendix. Intuitively, we find that high-sentiment words like “terrible”, “pleased”, “disappointed”, and “wonderful” tend to be selected as important.

Some of these examples call attention to weaker estimation of the complete conditional distribution around sentence boundaries, which can lead to an uncontrolled error rate for rejecting H_0. 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 leads to a poorer estimation of the complete conditional distribution around punctuation tokens, leading to some unintuitive highlighted words (e.g., “open”) in the example reviews.

5. Discussion

Scientists need to understand predictions from machine learning models when making decisions. In medicine, a treatment based on a black box prediction could lead to patient harm if the prediction was based on poor evidence or flawed reasoning. In biology, a set of low quality predictions may lead scientists to waste time and funding exploring a potential new drug target that was simply an artifact of the correlation structure of the data. To ensure reliability of models, scientists must be able to audit and confirm their

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.

Table 1. The true label and model output for a review sentiment classification task on example reviews. Text classifier interpretations from the OSFT are shown using the two-sided test statistic; Eq. (4). Each WordPiece was tested, and WordPieces for which the null hypothesis was rejected are shown in **blue**, indicating that the WordPiece was important for the model prediction.

reasoning in a principled manner.

This paper proposed a model interpretation framework that mirrors the analysis protocol employed by scientists: the null hypothesis test. The framework is a first step towards more statistically rigorous interpretation of black box models. By viewing model interpretation as a hypothesis testing problem, scientists can control the error rate at fixed levels and gain confidence in their interpretations.

Both the IRT and OSFT rely on the assumption that a meaningful null distribution $Q(X_S|X_{-S})$ is known or well-estimated. If it is poorly estimated, the error rate for rejecting H_0 will not be controlled. Conditional randomization tests like the IRT have been shown to be robust to model misspecification and estimation of the conditional from finite data, both empirically and theoretically (Berrett et al., 2018; Tansey et al., 2018). Similarly, the OSFT procedure is closely related to the Model-X Knockoffs approach of Candes et al. (2018), for which theoretical robustness results have also been established (Barber et al., 2018). These theoretical results, combined with the empirical results in Section 4, suggest that sufficient estimation of the conditional is tractable in practice.

An added benefit of the testing framework is that it can leverage three distinct fields of machine learning research. First, any state-of-the-art classifiers and regressors can be used, because the framework is model-agnostic. Second, in the case of high-dimensional data like images and language, autoregressive models for joint probability estimation in these tasks can be reused as approximate null models. Third, though we manually selected image regions in Section 4.2, anomaly and object detection methods can be used to automatically find candidate features to test. Analogously, dependency and syntactic parsing could be used to detect spans of candidate word features for language models. Each of these fields has a substantial literature, and future gains in these methods will translate to improving the performance of the IRT and OSFT.

The test statistics proposed in Section 3.4 reflect traditional one-sided and two-sided tests, but the framework admits any test statistic. There may be other statistics with higher power or that better fit a specific interpretation task. Samples could also be stratified into groups so that feature importance could be analyzed at the group level. For example, a cancer patient cohort may be segmented first by the progress grade of their tumor, then model predictions from gene expression data could be analyzed for each group. We plan to explore these ideas in 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, and the anonymous reviewers for their valuable feedback.

References

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I. J., Hardt, M., and Kim, B. Sanity checks for saliency maps. *Advances in Neural Information Processing Systems (NIPS)*, 2018.

Alipanahi, B., Delong, A., Weirauch, M. T., and Frey, B. J. Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. *Nature biotechnology*, 33(8):831, 2015.

Barber, R. F. and Candès, E. J. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5): 2055–2085, 2015.

Barber, R. F., Candès, E. J., and Samworth, R. J. Robust inference with knockoffs. *arXiv preprint arXiv:1801.03896*, 2018.

Benjamini, Y. and Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, pp. 289–300, 1995.

Berrett, T. B., Wang, Y., Barber, R. F., and Samworth, R. J. The conditional permutation test. *arXiv preprint arXiv:1807.05405*, 2018.

Bobisse, S., Foukas, P. G., Coukos, G., and Harari, A. Neoantigen-based cancer immunotherapy. *Annals of Translational Medicine*, 4(14), 2016.

Cabrera, A., Hohman, F., Lin, J., and Chau, D. H. Interactive classification for deep learning interpretation. *arXiv preprint arXiv:1806.05660*, 2018.

Candes, E., Fan, Y., Janson, L., and Lv, J. Panning for gold: Model-X knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B*, 2018.

Chang, C., Creager, E., Goldenberg, A., and Duvenaud, D. Explaining image classifiers by adaptive dropout and generative in-filling. *International Conference on Learning Representations (ICLR)*, 2018.

Chen, J., Song, L., Wainwright, M. J., and Jordan, M. I. Learning to explain: An information-theoretic perspective on model interpretation. *International Conference on Machine Learning (ICML)*, 2018.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Edgington, E. and Onghena, P. *Randomization tests*. Chapman and Hall/CRC, 2007.

Fong, R. and Vedaldi, A. Interpretable explanations of black boxes by meaningful perturbation. *International Conference on Computer Vision (ICCV)*, 2017.

Goh, G. B., Hodas, N. O., and Vishnu, A. Deep learning for computational chemistry. *Journal of Computational Chemistry*, 38(16):1291–1307, 2017.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *International Conference on Learning Representations (ICLR)*, 2015.

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., and Cuadros, J. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Journal of the American Medical Assocation*, 316(22):2402–2410, 2016.

Kelley, D. R., Snoek, J., and Rinn, J. L. Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Research*, 2016.

Kraus, O. Z., Ba, J. L., and Frey, B. J. Classifying and segmenting microscopy images with deep multiple instance learning. *Bioinformatics*, 32(12):i52–i59, 2016.

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, 2016.

Lipton, Z. C. The mythos of model interpretability. *ICML Workshop on Human Interpretability in Machine Learning (WHI)*, 2016.

Lundberg, S. and Lee, S. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems (NIPS)*, 2017.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. Learning word vectors for sentiment analysis. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150, 2011.

Ribeiro, M. T., Singh, S., and Guestrin, C. "Why should I trust you?": Explaining the predictions of any classifier. *International Conference on Knowledge Discovery and Data Mining (KDD)*, 2016.

Sarraf, S. and Tofighi, G. Classification of Alzheimer's disease using fMRI data and deep learning convolutional neural networks. *arXiv preprint arXiv:1603.08631*, 2016.

Schaumberg, A. J., Rubin, M. A., and Fuchs, T. J. H&E-stained whole slide image deep learning predicts SPOP mutation state in prostate cancer. *BioRxiv*, 2018.

Selvaraju, R. R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., and Batra, D. Grad-CAM: Why did you say that? Visual explanations from deep networks via gradient-based localization. *International Conference on Computer Vision (ICCV)*, 2016.

Simonyan, K., Vedaldi, A., and Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *International Conference on Learning Representations (ICLR) Workshop*, 2014.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. *International Conference on Machine Learning (ICML)*, 2017.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. Going deeper with convolutions. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

Tansey, W., Veitch, V., Zhang, H., Rabadan, R., and Blei, D. M. The holdout randomization test: Principled and easy black box feature selection. *arXiv preprint arXiv:1811.00645*, 2018.

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

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in Neural Information Processing Systems (NIPS)*, 2017.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., and Macherey, K. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

Xiong, H. Y., Alipanahi, B., Lee, L. J., Bretschneider, H., Merico, D., Yuen, R. K., Hua, Y., Gueroussov, S., Najafabadi, H. S., and Hughes, T. R. The human splicing code reveals new insights into the genetic determinants of disease. *Science*, 347, 2015.

Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., and Huang, T. S. Generative image inpainting with contextual attention. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

A. Supplementary Material

A.1. Additional language examples

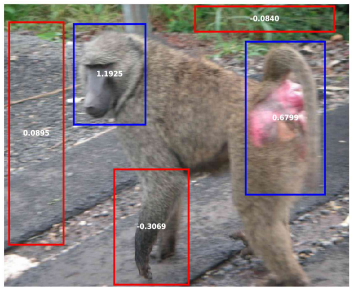
Table 2 shows some additional examples of using the OSFT to interpret the trained BERT model for sentiment classification.

A.2. Additional image examples

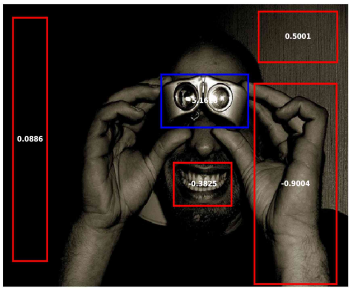
Figs. 3 to 6 show the remaining 38 examples of using the OSFT to interpret an image classifier on Imagenet. The label for each image is the predicted class and predicted class probability.

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 luri**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 intri**gue** 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)

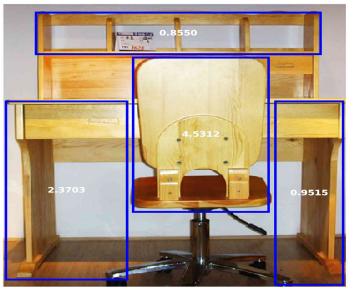
Table 2. The true label and model output for a review sentiment classification task on additional example reviews. Text classifier interpretations from the OSFT are shown using the two-sided test statistic; Eq. (2). Each WordPiece was tested, and WordPieces for which the null hypothesis was rejected are shown in **blue**, indicating that the WordPiece was important for the model prediction.



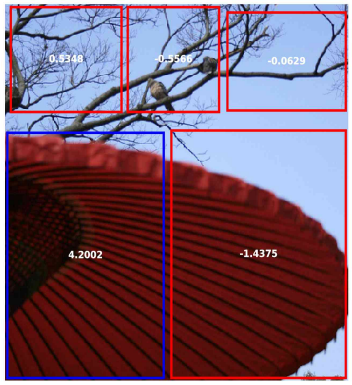
(a) Baboon ($p = 0.935$)



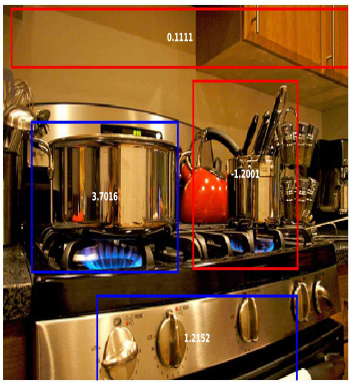
(b) Binoculars ($p = 0.920$)



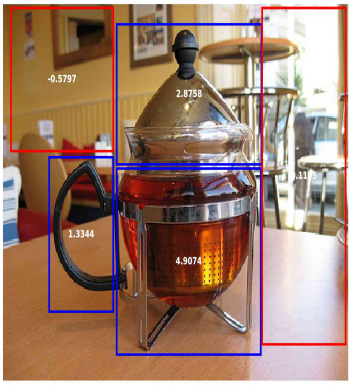
(c) Desk ($p = 0.695$)



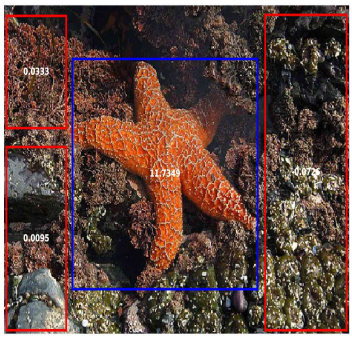
(d) Umbrella ($p = 0.925$)



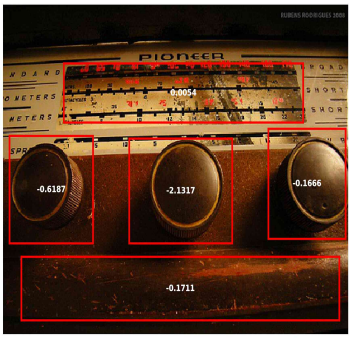
(e) Stove ($p = 0.671$)



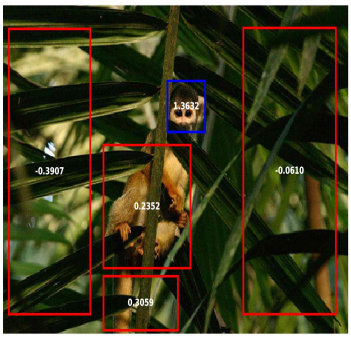
(f) Teapot ($p = 0.819$)



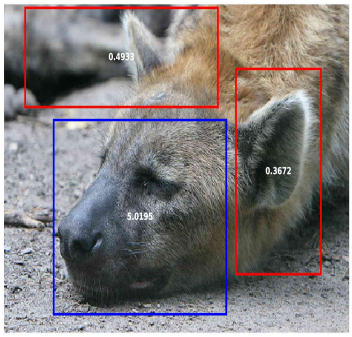
(g) Starfish ($p = 0.995$)



(h) Radio ($p = 0.762$)

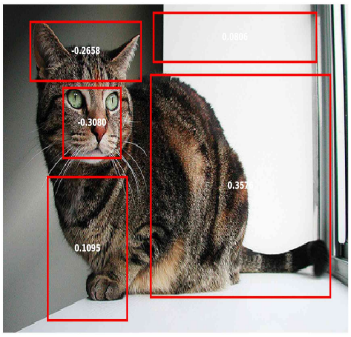


(i) Squirrel monkey ($p = 0.689$)



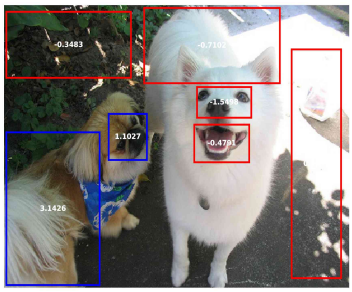
(j) Hyena ($p = 0.961$)

(k) Toaster ($p = 0.992$)

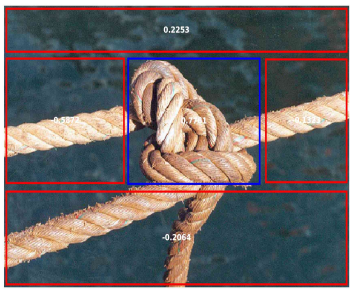


(l) Tabby cat ($p = 0.922$)

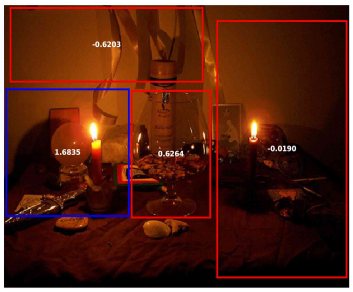
Figure 3.



(a) Pekinese ($p = 0.484$)

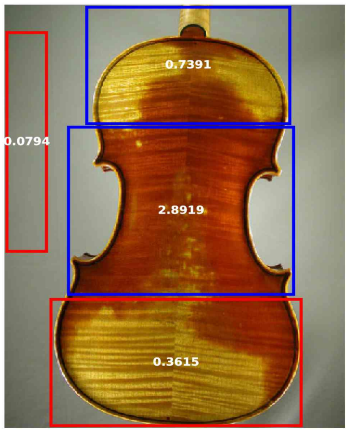


(b) Knot ($p = 0.994$)

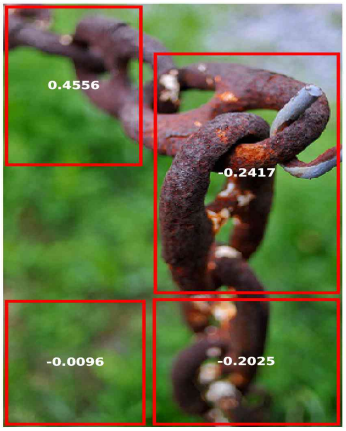


(c) Altar ($p = 0.791$)

(d) Gown ($p = 0.863$)



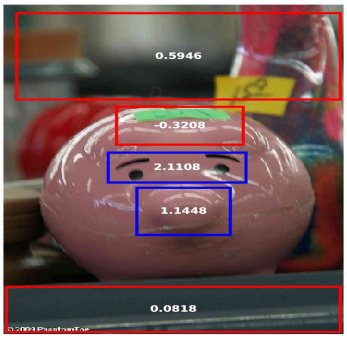
(e) Violin, fiddle ($p = 0.991$)



(f) Binoculars ($p = 0.679$)

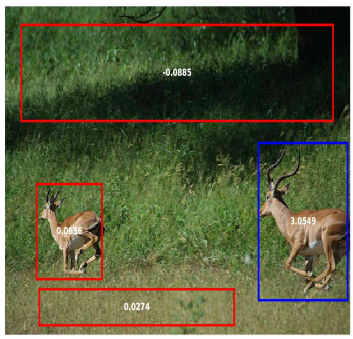
(g) Street sign ($p = 0.883$)

(h) Menu ($p = 0.924$)

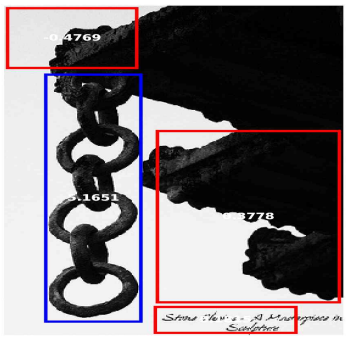


(i) Piggy bank ($p = 0.930$)

(j) Harp ($p = 0.999$)

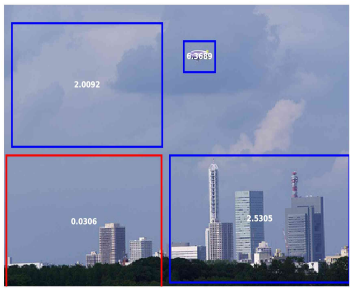


(k) Impala ($p = 0.679$)

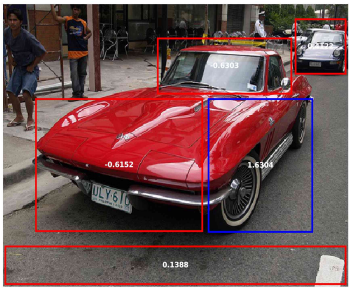


(l) Hook, claw ($p = 0.431$)

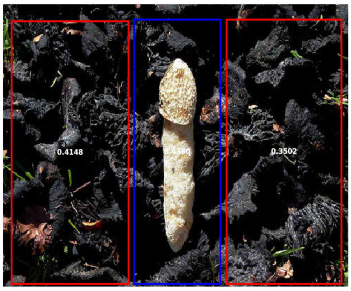
Figure 4.



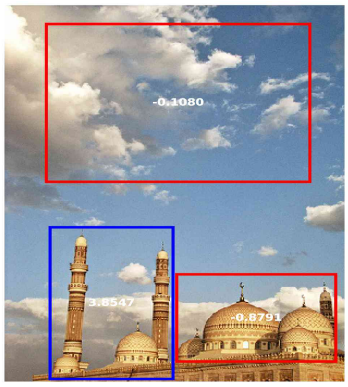
(a) Airship, dirigible ($p = 0.969$)



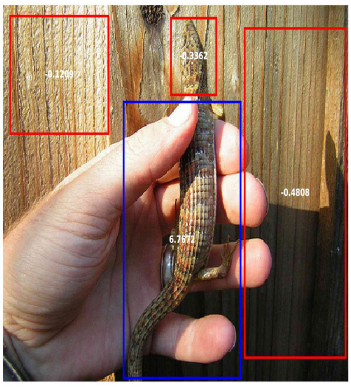
(b) Convertible ($p = 0.619$)



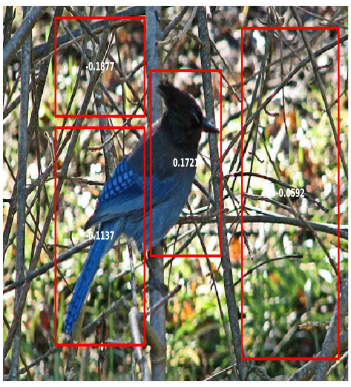
(c) Stinkhorn, carrion fungus ($p = 0.803$)



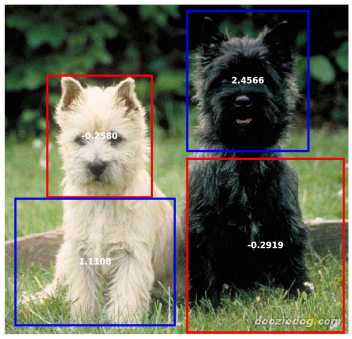
(d) Mosque ($p = 0.749$)



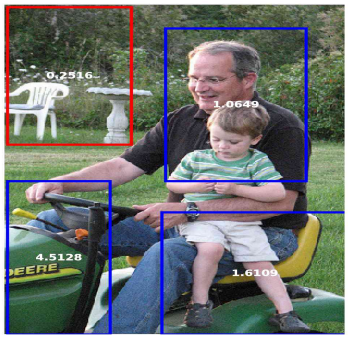
(e) Alligator lizard ($p = 0.954$)



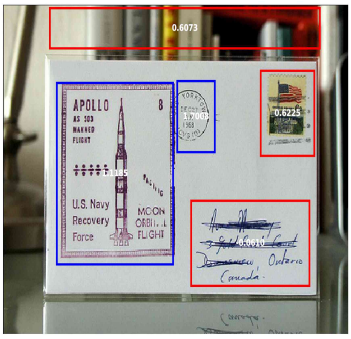
(f) Jay ($p = 0.961$)



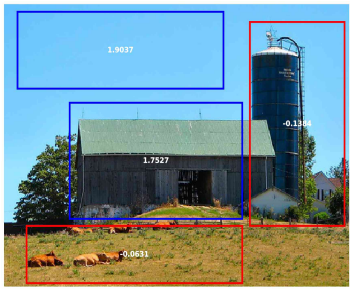
(g) Cairn terrier ($p = 0.816$)



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

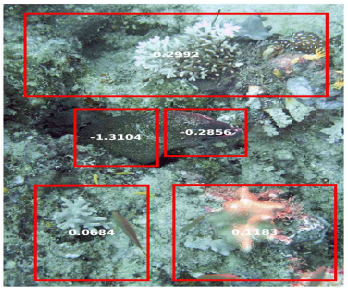


(i) Envelope ($p = 0.059$)



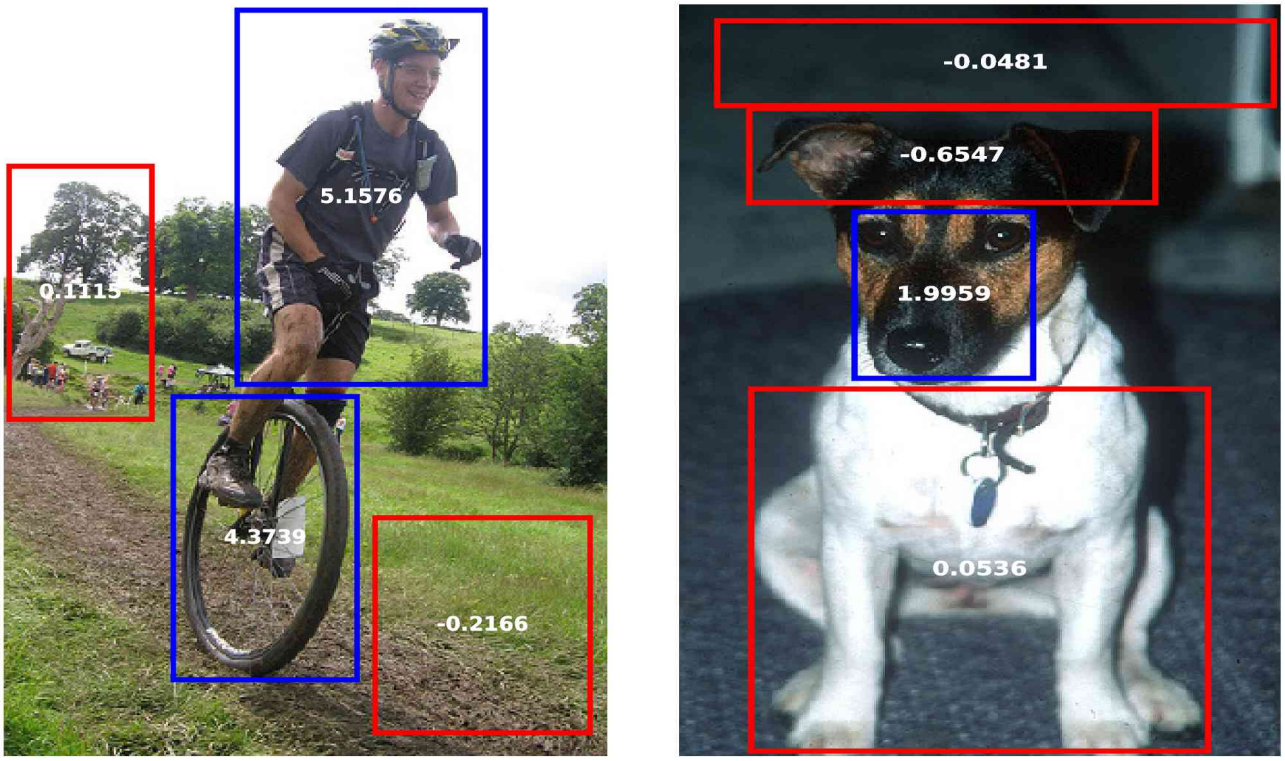
(j) Barn ($p = 0.968$)

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



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

Figure 5.



(a) Unicycle ($p = 0.996$)

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

Figure 6.