

## A DETAILS ABOUT THE EVALUATION SCHEME

Here we present the structure of the models used for our experiments, as well as how the evaluation metrics are calculated. Readers may refer to the released code for more details.

### A.1 THE TEXT CLASSIFICATION MODELS FOR IMDB AND YELP P.

Table 4: The structure of text classification models for IMDB and Yelp P.

Layer	# Units	Kernel Size	Stride	Padding	Activation
Embedding	50	--	--	--	--
Dropout (0.2)	--	--	--	--	--
Convolution (1D)	250	3	1	default	ReLU
GlobalMaxPooling	--	--	--	--	--
Fully-Connected	250	--	--	--	ReLU
Dropout (0.2)	--	--	--	--	--
Fully-Connected	2	--	--	--	Softmax

The structure of neural network models to be explained for text classification datasets (including IMDB and Yelp P.) is presented in Table 4. This model structure achieved accuracy scores of about 96.1% and 89.0% on the train set and the test set of IMDB respectively, and about 96.0% and 95.4% on the train set and test set of Yelp P. respectively. The structures of explainers, as well as the approximators of models including L2X, VIBI, AIL, and L2X with  $\hat{y}$  are slightly different from this to incorporate global information when selecting features and to select keywords. During our experiments, we find that while Pretrain prefers powerful approximators, Weight works more stable with simple approximators. This result is consistent with the results in Table 2. Besides, as estimating  $\mathbb{E}(Y|X \odot M)$  requires good calibrations, layers that act differently during training and testing, like *Dropout* and *BatchNormalization* layers, may not be welcomed by explanation models.

### A.2 THE IMAGE CLASSIFICATION MODEL FOR MNIST AND FASION-MNIST

Table 5: The structure of image classification models for MNIST and Fashion-MNIST

Layer	# Units	Kernel Size	Stride	Padding	Activation
Convolution (2D)	32	(3, 3)	1	default	ReLU
Convolution (2D)	64	(3, 3)	1	default	ReLU
MaxPooling (2D)	--	(2, 2)	(2, 2)	--	--
Flatten	--	--	--	--	--
Dropout (0.25)	--	--	--	--	--
Fully-Connected	128	--	--	--	ReLU
Dropout (0.5)	--	--	--	--	--
Fully-Connected	2	--	--	--	Softmax

The structure of neural network models to be explained for image classification datasets (including MNIST and F-MNIST) is presented in Table 5. This model structure achieved accuracy scores of more than 99.8% on both the train set and test set of MNIST, and about 96.1% and 92.2% on the train set and test set of F-MNIST respectively. The structures of explainers as well as the approximators of models including L2X, VIBI, AIL, and L2X with  $\hat{y}$  are slightly different from this to incorporate global information and to select pixels.

### A.3 CALCULATION OF THE EVALUATION METRICS

The same with Chen et al. (2018) and Liang et al. (2020), we used *post-hoc accuracy* for quantitatively validating the effectiveness of the methods. After feature selection, unselected words of texts were filled with <PAD> tokens, and unselected pixels of images were filled with their average

values. Then we used the model to be explained to predict the labels, and compared whether the labels change or not. If the model makes consistent predictions before and after feature selection, the selected features may be informative for the model to make the decisions.

#### A.4 BRIEF INTRODUCTION TO THE BASELINE METHODS

Among the baseline methods, Gradient (Simonyan et al., 2013) takes advantage of the property of neural networks and selects the input features which have the most significant absolute values of gradients. LIME (Ribeiro et al., 2016) explains a model by quantifying the model’s sensitivity to changes in the input. CXPlain (Schwab & Karlen, 2019) involves the real labels  $y$  to compute the loss-function values by erasing each feature to zero and normalizes the loss-function values as the surrogate for ideal explanations for a neural-network to learn. Our methods follow the same paradigm as L2X (Chen et al., 2018), VIBI (Bang et al., 2019), and AIL (Liang et al., 2020), which use hard attention to select a fixed number of features to approximate the output of the original models to be explained. VIBI improves L2X to encourage the briefness of the learned explanation by adding a constraint for the feature scores to a global prior. AIL (Liang et al., 2020) is the state-of-the-art method that use adversarial training to encourage the gap between the predictability of selected/unselected features, additionally, AIL incorporate the outcome it aims to justify explicitly, as well as the gradients of the original model w.r.t. the samples to provide a warm start.

## B EXAMPLE EXPLANATION RESULTS

Here, we present the interpretation results of the examples from the tested datasets. Please note that we are trying to explain how machine learning models make decisions. These models are trained on datasets of limited size, and the datasets may be more or less biased, so the interpretation results may be inconsistent with ordinary people’s understanding.

### B.1 IMDB

We first present two examples from IMDB dataset. The results are in Table 6 and Table 7. Explanations to the functional tokens like <START> and <PAD> are omitted.

As shown from the two examples, L2X with  $\hat{y}$  tends to select high-frequency but meaningless words (e.g., *this*, *an*, and *it*), or niche words which are merged as <UNK> token (e.g., *Aussie* and *sporadically*). This phenomenon explains the reason why L2X with  $\hat{y}$  get poor post-hoc accuracy scores during our experiments on text classification datasets, and as analyzed in Section 3, this can be ascribed to the combinatorial shortcuts. After applying the proposed methods (i.e., Pretrain and Weight), the combinatorial shortcut problem has been mitigated successfully, and the explaining model can select better words for explanations. Besides, we can find that after adding the information of  $\hat{y}$  to the query, the explanation models could better capture the key parts corresponding to the original given model to be explained. For example, L2X selects “*is was a vacant effort*” in Table 7 and the given model predicts *negative* with the selected words by L2X, however, the model to be explained actually predicts *positive* for this example, and after adding  $\hat{y}$  and applying the proposed methods, the explanation results become more consistent to the given model.

### B.2 YELP P.

In this section we present two examples from Yelp P. dataset. The results are in Table 8 and Table 9. Similarly, explanations to the functional tokens like <START> and <PAD> are omitted.

Compared with the results of IMDB, we find that the explanation results of L2X seem to be more susceptible to combinatorial shortcut problem, as words like *was*, *a* appear in the explanation results of L2X more frequently. This phenomenon may be ascribed to that the sentences in Yelp P. are shorter than IMDB, thus the explainer could more easily capture the global meaning of the sentences and form combinatorial shortcuts. For L2X with  $\hat{y}$ , the results are consistent with the results of IMDB, i.e., meaningless words and <UNK> tokens are frequently selected. Pretrain and Weight can help mitigate the problem of combinatorial shortcut, and thus improve the performance of interpretability.

Table 6: The results of explanation on Example 1 from IMDB. The model to be explained makes the correct prediction for this sample, i.e., the ground truth label of the review text is *negative*, and the trained model predicts *negative* for it.

L2X	As an indie filmmaker, I try to at least make a decent film . This piece of ** was beyond low budget. It was shot on video and not 24P mini-DV at least. The look and feel of this was just baaaad. I met the director a few years ago at ShowBiz Expo in LA and he was talking about that book, Film-making for dummies that he was putting together. I thought this little video was going to be something but I guess I was wrong. He could have brought the value up a little by shooting 16mm film instead of that awful video. The plot was stupid as well as the acting and all the fake green screen and sound and the whole nine yards. I had a choice tonight to rent any movie and made the wrong choice. Damn!!!! I did buy JoyRide which was a hell of a movie. Maybe the director should read real motion picture books on film-making and not try to cut corners when trying to make a low budget flick. Maybe he should learn from the masters who made, Night of the living dead and The Evil Dead and Chain saw massacre. just to name a few of the all time low budget great hits. This is one video that should have stayed dead. I cannot call it a film because he did not use film. (Prediction with key words: <i>negative</i> )
L2X with $\hat{y}$	As an indie filmmaker, I try to at least make a decent film . This piece of ** was beyond low budget. It was shot on video and not 24P mini-DV at least. The look and feel of this was just baaaad. I met the director a few years ago at ShowBiz Expo in LA and he was talking about that book, Film-making for dummies that he was putting together. I thought this little video was going to be something but I guess I was wrong. He could have brought the value up a little by shooting 16mm film instead of that awful video. The plot was stupid as well as the acting and all the fake green screen and sound and the whole nine yards. I had a choice tonight to rent any movie and made the wrong choice. Damn!!!! I did buy JoyRide which was a hell of a movie. Maybe the director should read real motion picture books on film-making and not try to cut corners when trying to make a low budget flick. Maybe he should learn from the masters who made, Night of the living dead and The Evil Dead and Chain saw massacre. just to name a few of the all time low budget great hits. This is one video that should have stayed dead. I cannot call it a film because he did not use film. (Prediction with key words: <i>positive</i> )
L2X with $\hat{y}$ (Pretrain)	As an indie filmmaker, I try to at least make a decent film . This piece of ** was beyond low budget. It was shot on video and not 24P mini-DV at least. The look and feel of this was just baaaad. I met the director a few years ago at ShowBiz Expo in LA and he was talking about that book, Film-making for dummies that he was putting together. I thought this little video was going to be something but I guess I was wrong. He could have brought the value up a little by shooting 16mm film instead of that awful video. The plot was stupid as well as the acting and all the fake green screen and sound and the whole nine yards. I had a choice tonight to rent any movie and made the wrong choice. Damn!!!! I did buy JoyRide which was a hell of a movie. Maybe the director should read real motion picture books on film-making and not try to cut corners when trying to make a low budget flick. Maybe he should learn from the masters who made, Night of the living dead and The Evil Dead and Chain saw massacre. just to name a few of the all time low budget great hits. This is one video that should have stayed dead. I cannot call it a film because he did not use film. (Prediction with key words: <i>negative</i> )
L2X with $\hat{y}$ (Weight)	As an indie filmmaker, I try to at least make a decent film . This piece of ** was beyond low budget. It was shot on video and not 24P mini-DV at least. The look and feel of this was just baaaad. I met the director a few years ago at ShowBiz Expo in LA and he was talking about that book, Film-making for dummies that he was putting together. I thought this little video was going to be something but I guess I was wrong. He could have brought the value up a little by shooting 16mm film instead of that awful video. The plot was stupid as well as the acting and all the fake green screen and sound and the whole nine yards. I had a choice tonight to rent any movie and made the wrong choice. Damn!!!! I did buy JoyRide which was a hell of a movie. Maybe the director should read real motion picture books on film-making and not try to cut corners when trying to make a low budget flick. Maybe he should learn from the masters who made, Night of the living dead and The Evil Dead and Chain saw massacre. just to name a few of the all time low budget great hits. This is one video that should have stayed dead. I cannot call it a film because he did not use film. (Prediction with key words: <i>negative</i> )

### B.3 MNIST

Figure 2 shows four examples from MNIST. We find that the given model to be explained predicts all samples with label 8 in the test set correctly, so we choose the least confident sample to present in Figure 2d.

The gray parts of images are unselected pixels, while the white/black parts are selected pixels. We find a very significant tendency that L2X and Weight tend to select white pixels in the image, while L2X with  $\hat{y}$  and Pretrain tend to select black pixels. In Figure 2b, only Weight select the right parts which makes the model to be explained predict 8 for this sample.

### B.4 F-MNIST

Figure 3 shows four examples from Fashion-MNIST. Similarly, the gray parts of images are unselected pixels, and the white/black parts are selected pixels.

Interestingly, we find that L2X tends to select the pixels on the edges. Similar to the examples of L2X in Table 7, the highlighted parts given by L2X may tend to be important generally, however, it may not be the most significant part for explaining the predictions of given models. On the other

Table 7: The results of explanation on Example 2 from IMDB. The model to be explained makes the incorrect prediction for this sample, i.e., the ground truth label of the review text is *negative*, while the trained model predicts *positive* for it.

L2X	Aussie Shakespeare for 18-24 set. With blood, blood and more blood, and good dose of nudity. This will not be for every one on <b>may</b> levels, to violent for some too cheap for most. Done on low budget they try and do there best but it <b>only</b> works sporadically. And this macbeth just <b>seem</b> to be lacking, it's just <b>not</b> compelling. Although there is some good acting on the part of most you don't get into there heads especially mecbeths. The <b>best</b> performance came from Gary sweet and the strangest mick molly. If your into Shakespeare then see it, but if you like your cheese mature you will love it. It not a bad film but it not that good either. Sam Peckenpah would of loved it, that is if it was filmed as a western. I was expecting a lot from this, as I loved romper stomper. But this <b>is</b> was a <b>vacant</b> effort. (Prediction with key words: <i>negative</i> )
L2X with $\hat{y}$	Aussie Shakespeare for 18-24 set. With blood, blood <b>and</b> more blood, <b>and</b> good <b>dose</b> of nudity. This will not be for every one on <b>may</b> levels, to violent for some too cheap for most. Done on low budget they try <b>and</b> do there best but it only works <b>sporadically</b> . And this <b>macbeth</b> just seem to be lacking, it's just not compelling. Although there is some good acting on the part of most you don't get into there heads especially mecbeths. The best performance came from Gary sweet <b>and</b> the strangest mick molly. If your into Shakespeare then see it, but if you like your cheese mature you will love it. It not a bad film but it not that good either. Sam Peckenpah would of loved it, that is if it was filmed as a western. I was expecting a lot from this, as I loved <b>romper</b> stomper. But this is was a <b>vacant</b> effort. (Prediction with key words: <i>positive</i> )
L2X with $\hat{y}$ (Pretrain)	Aussie Shakespeare for 18-24 set. With blood, blood and more blood, and good dose of nudity. This will not be for every one on <b>may</b> levels, to violent for some too cheap for most. Done on low budget they try <b>and</b> do there <b>best</b> but it only works sporadically. And this macbeth just seem to be lacking, it's just not compelling. Although there is some good acting on the part of most you don't get into there heads <b>especially</b> mecbeths. The <b>best performance came</b> from Gary sweet and the strangest mick molly. If your into Shakespeare then see it, but if you like your cheese mature you will <b>love</b> it. It not a bad film but it not that good either. Sam Peckenpah would of <b>loved</b> it, <b>that</b> is if it was filmed as a western. I was expecting a lot from this, <b>as I loved</b> romper stomper. But this is was a vacant effort. (Prediction with key words: <i>positive</i> )
L2X with $\hat{y}$ (Weight)	Aussie Shakespeare for 18-24 set. With blood, blood and more blood, and good dose of nudity. This will not be for every one on <b>may</b> levels, to violent <b>for</b> some too cheap for most. Done on low budget they <b>try</b> and do there best but it only works sporadically. And this macbeth just <b>seem</b> to be lacking, <b>it's</b> just not compelling. Although there is some <b>good</b> acting on the part of most <b>you</b> don't get into there heads especially mecbeths. The best performance came from Gary sweet and the strangest mick molly. If your into Shakespeare then see it, but if <b>you like</b> your cheese mature you will love it. It not a bad film but it not that good either. Sam Peckenpah would of loved it, that is if it was filmed as a western. I was expecting a lot from this, as I loved romper stomper. But <b>this</b> is was a vacant effort. (Prediction with key words: <i>positive</i> )

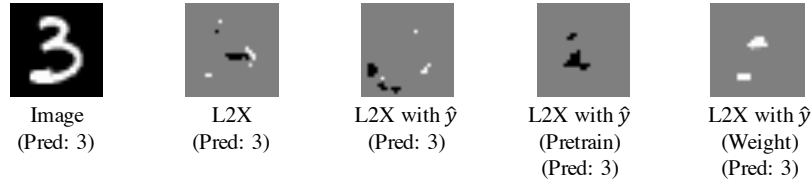
Table 8: The results of explanation on Example 1 from Yelp P. The model to be explained makes the correct prediction for this sample, i.e., the ground truth label of the review text is *positive*, and the trained model predicts *positive* for it.

L2X	Good choice if <b>you</b> are looking for <b>a</b> pricier Italian menu. They feature veal entrees and a pretty good assortment of Italian seafood dishes. Their homemade gnocchi and cavatelli are wonderful as well. The service is <b>typically</b> so-so, but the great food keeps me coming back. (Prediction with key words: <i>positive</i> )
L2X with $\hat{y}$	Good choice if you are looking for a pricier Italian menu. They feature veal entrees and a pretty good assortment of Italian seafood dishes. Their homemade gnocchi and cavatelli are wonderful <b>as well</b> . The service is typically <b>so-so</b> , but the great food keeps me coming back. (Prediction with key words: <i>positive</i> )
L2X with $\hat{y}$ (Pretrain)	Good choice if you are looking for a pricier Italian menu. They feature veal entrees and a pretty good assortment of Italian seafood <b>dishes</b> . Their <b>homemade</b> gnocchi and cavatelli are <b>wonderful</b> as <b>well</b> . The service is typically so-so, but the <b>great</b> food keeps me coming back. (Prediction with key words: <i>positive</i> )
L2X with $\hat{y}$ (Weight)	<b>Good choice</b> if you are <b>looking</b> for a pricier Italian menu. They feature veal entrees and a pretty good assortment of Italian seafood dishes. Their <b>homemade</b> gnocchi and cavatelli are wonderful as well. The service is typically so-so, but the <b>great</b> food keeps me coming back. (Prediction with key words: <i>positive</i> )

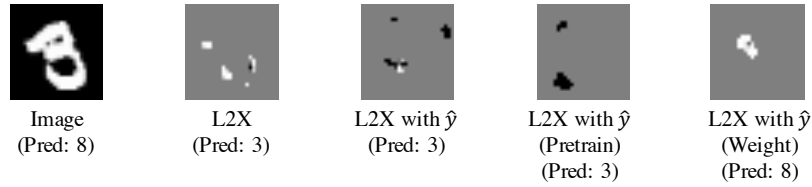
side, the explanations given by Weight tend to be of the neckline parts and the shoulder parts of the clothing, and maybe more informative to explain the predictions of machine learning models. For example, in Figure 3c, L2X selects the edges of the given image, while the model to be explained judges that the select features suggest that the image belongs to a Pullover.

Table 9: The results of explanation on Example 2 from Yelp P. The model to be explained makes the incorrect prediction for this sample, i.e., the ground truth label of the review text is `positive`, while the trained model predicts `negative` for it.

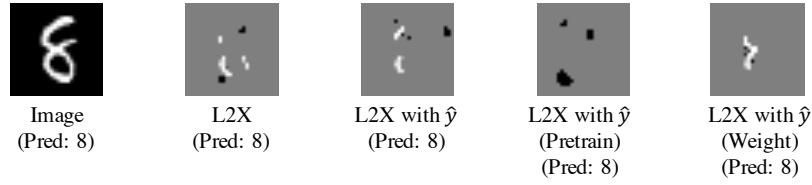
L2X	With so many businesses in Southpark, <code>you'd</code> think there would be more places to eat for a quick bite. Unfortunately, its either white table cloth restaurants, fast <code>food</code> and a few disappointing sandwich places. Thankfully, American Roadside opened and it has help to add a little diversity to the lunchtime landscape. It's a burger place where everything is made to order. I've had lunch there a dozen times or so since its opened and the food <code>was</code> the best the first week. <code>Then</code> it got worse and now its heading back to great. I'm not sure what was going on there but even at its worst, it was still lightyears ahead of the other burger place with the Taco name. Choose the onion rings over the fries. (Prediction with key words: <code>negative</code> )
L2X with $\hat{y}$	With so many businesses in <code>Southpark</code> , you'd think there would be more places to eat for a quick bite. Unfortunately, its either white table <code>cloth</code> restaurants, fast food and a few disappointing sandwich places. Thankfully, American <code>Roadside</code> opened and it has help to add a little <code>diversity</code> to the lunchtime landscape. It's a burger place where everything is made to order. I've had lunch there a dozen times or so since its opened and the food was the best the first week. Then it got worse and now its heading back to great. I'm not sure what was going on there but even at its worst, it was still lightyears ahead of the other burger place with the <code>Taco</code> name. Choose the onion rings over the fries. (Prediction with key words: <code>positive</code> )
L2X with $\hat{y}$ (Pretrain)	With so many businesses in Southpark, you'd think there would be more places to eat for a quick bite. <code>Unfortunately</code> , its either white table cloth restaurants, fast food and a few <code>disappointing</code> sandwich places. Thankfully, American Roadside opened and it has help to add a little diversity to the lunchtime landscape. It's a burger place where everything is made to order. I've had lunch there a dozen times or so since its opened and the food was the best the first week. <code>Then</code> it got <code>worse</code> and now its heading back to great. I'm <code>not</code> sure what was going on there but even at its worst, it was still lightyears ahead of the other burger place with the Taco name. Choose the onion rings over the fries. (Prediction with key words: <code>negative</code> )
L2X with $\hat{y}$ (Weight)	With so many businesses in Southpark, you'd think there would be more places to eat for a quick bite. Unfortunately, its either white table cloth restaurants, fast food and a few disappointing sandwich places. Thankfully, American Roadside opened and it has help to add <code>a little</code> diversity to the lunchtime landscape. It's a burger place where everything is made to order. I've had lunch there a dozen times or so since its <code>opened</code> and the food was the best the first week. Then it got worse and now its heading back to great. I'm not sure what <code>was going</code> on there but even at its worst, it was still lightyears ahead of the other burger place with the Taco name. Choose the onion rings over the fries. (Prediction with key words: <code>negative</code> )



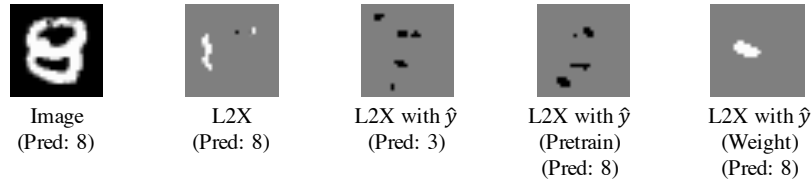
(a) The ground truth label of the image is 3, and the model to be explained predicts 3 for it.



(b) The ground truth label of the image is 3, but the model to be explained predicts 8 for it.

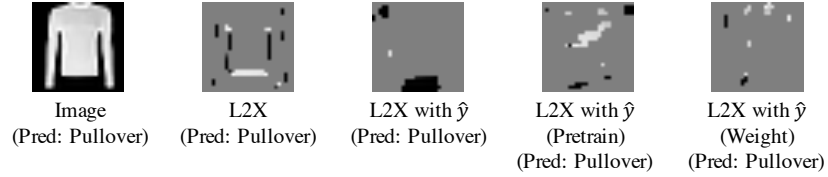


(c) The ground truth label of the image is 8, and the model to be explained predicts 8 for it.

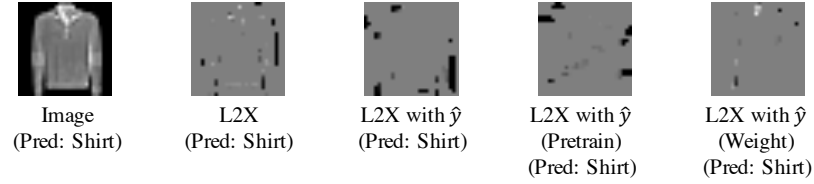


(d) The ground truth label of the image is 8, and the model to be explained predicts 8 but with very low confidence.

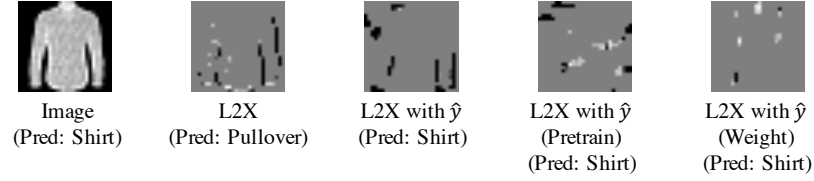
Figure 2: The result of explanation on images from MNIST.



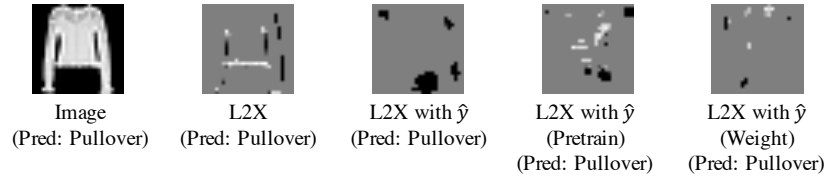
(a) The ground truth label of the image is `Pullover`, and the model to be explained predicts `Pullover` for it.



(b) The ground truth label of the image is `Pullover`, but the model to be explained predicts `Shirt` for it.



(c) The ground truth label of the image is `Shirt`, and the model to be explained predicts `Shirt` for it.



(d) The ground truth label of the image is `Shirt`, but the model to be explained predicts `Pullover` for it.

Figure 3: The result of explanation on images from Fashion-MNIST.