A Geometrical Approach to Finding Difficult Examples in Language

Debajyoti Datta, Shashwat Kumar, Laura Barnes
Systems Engineering
University of Virginia
Charlottesville, VA 22903, USA
{dd3ar, sk9epp, lb3dp}@virginia.edu

P. Thomas Fletcher
Electrical and Computer Engineering
University of Virginia
Charlottesville, VA 22903, USA
ptf8v@virginia.edu

Abstract

A growing body of evidence has suggested that metrics like accuracy overestimate the classifier's generalization ability. Several state of the art Natural Language Processing (NLP) classifiers like BERT and LSTM rely on superficial cue words (e.g., if a movie review has the word "romantic", the review tends to be positive), or unnecessary words (e.g., learning a proper noun to classify a movie as positive or negative). One approach to test NLP classifiers for such fragilities is analogous to how teachers discover gaps in a student's understanding: by finding problems where small perturbations confuse the student. While several perturbation strategies like contrast sets or random word substitutions have been proposed, they are typically based on heuristics and/or require expensive human involvement. In this work, using tools from information geometry, we propose a principled way to quantify the fragility of an example for an NLP classifier. By discovering such fragile examples for several state of the art NLP models like BERT, LSTM, and CNN, we demonstrate their susceptibility to meaningless perturbations like noun/synonym substitution, causing their accuracy to drop down to 20 percent in some cases. Our approach is simple, architecture agnostic and can be used to study the fragilities of text classification models.

1 Introduction

NLP classifiers have achieved state of the art performance in several tasks like sentiment analysis Maas et al. (2011), semantic entailment Bowman et al. (2015), and question answering Rajpurkar et al. (2016). Despite their successes, several studies have pointed out issues in features learnt by such classifiers. Gururangan et al. (2018) discovered that several high performing NLP models were using trivial features like vagueness and negation to perform classification. Geirhos et al. (2020) performed several experiments to discover that "romantic" movies tend to be classified as positive movie reviews due to the presence of unnecessary words like proper nouns, a phenomena they called shortcut learning. Even models like BERT Devlin et al. (2018) rely on superficial cue words like "not" to infer the line of argumentation. Strategies like this enable models to perform prediction without inherently using the semantic meaning of the sentence Niven & Kao (2020). Simple attributes like word lengths are also exploited by models for prediction Poliak et al. (2018).

In order to address these issues, several perturbation based approaches have been proposed. Contrast Sets Gardner et al. (2020) and Counterfactual examples Kaushik et al. (2019) get human annotators to perform minimal token substitutions to construct challenging test sets for the classifier. While useful in addressing biases, manual curation of datasets are often time consuming and require extensive efforts. Using unsupervised training to mitigate issues related to shortcut learning has not been successful in practice Niven & Kao (2020).

While interesting, most traditional formulations treat this input embedding space as flat, thus reasoning that the gradient of the likelihood in the input space gives us the direction that causes the most significant change in likelihood. If we were, however, to consider the discrete likelihood of the class probabilities as the model output and the input as a pullback of this output, the space of output probability distributions is certainly non-linear, and the Euclidean distance metric no longer suffices.

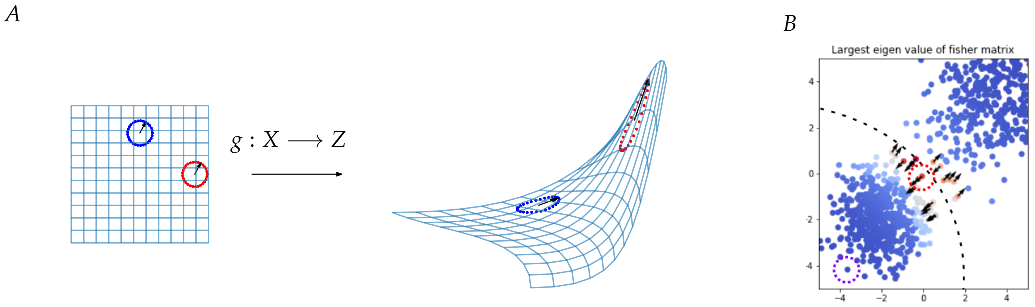
A
$g : X \longrightarrow Z$
B
Largest eigen value of fisher matrix
4
2
0
-2
-4
−4
−2
0
2
4

Figure 1: **A**: A neural network can be considered as a mapping g between the sentence manifold X on the left (represented as a Euclidean space for simplification) and the statistical manifold of probability distribution over outputs Z. The fisher metric defines a particular Riemannian metric over this manifold. Let us consider an ϵ ball around two sentences x_1 (blue circle) and x_2 (red circle). As we can see from the distortion of the blue circle, for x_1, any local perturbation can only result in a small change over the models output probability distribution which and vice versa for red circle. The eigenvalue of the fisher matrix quantifies this local distortion. **B**: We train a neural network to learn a decision boundary (black) to separate two gaussians and color each data point by the local distortion (blue implying low distortion, red implying higher distortion). As we can see from the colors, perturbing a point close to the decision boundary (red circle, same as red circle in A) results in a larger change over the model's output probability distribution than a point away from the boundary (blue circle, same as blue circle in A).

We show this schematically in Figure 1A, where the transformation g maps elements of the sentence manifold X to a non-linear statistical manifold Z. A natural distance metric to consider on this manifold is the Fisher Information Metric, which along with being a Hessian of the KL divergence, is also a useful distance measure between probability distributions. Furthermore, the Fisher Information Metric is invariant to transformations like changing the model architecture, provided the likelihood remains the same. We thus use the eigenvalues of the Fisher matrix to discover high fragility regions in the statistical manifold. In regions with high λ_{max}, small perturbations can cause large changes in the output probability distribution (Figure 1A, red circle). Linguistically, this corresponds to the classifier being susceptible to meaningless perturbations like noun, synonym substitutions. Similarly, in regions with low λ_{max} no local perturbation can affect the classifier, resulting in the classifier being resilient upto 20% percent of word substitutions.

Our main contribution can thus be summarized as the following. 1. We propose a second order statistic, the log of the largest eigenvalue of Fisher matrix in order to capture linguistic fragilities. 2. We extensively establish the empirical relationship between λ_{max} and success probability of random word substitutions. To the best of our knowledge, this is the first work analyzing properties of the fisher metric to understand classifier fragility in NLP. The rest of the paper is organized as follows: In Section 2, we summarize related work. In Section 3, we discuss our approach of computing the FIM and the gradient-based perturbation strategy. In Section 4, we discuss the results of the eigenvalues of FIM in synthetic data and sentiment analysis datasets with BERT and CNN. We also do extensive quantitative evaluations using 4 other text datasets. Finally, in Section 5, we discuss the implications of studying the eigenvalues of FIM for evaluating NLP models.

2 Related Work

In NLP, improperly trained machine learning models (for instance in the presence of limited/biased data) for classification often rely on spurious statistical patterns of the text and use *shortcut* for learning to classify. These can range from annotation artifacts Goyal et al. (2017); Kaushik & Lipton (2018); Gururangan et al. (2018), spelling mistakes McCoy et al. (2019), or new test conditions that require world knowledge Glockner et al. (2018).

Another issue of language recently has been that static benchmarks (e.g., GLUE by Wang et al. (2018)) tend to saturate quickly because of the availability of ever-increasing compute and harder benchmarks are needed to evaluate NLP models (e.g., SuperGlue Wang et al. (2019)). A more

Table 1: **Top row**: Substituting even a single word for fragile examples (large λ_{max}) causes BERT to change the predicted sentiment. **Bottom Row**: Robust examples (small λ_{max}), however, retain positive sentiment despite multiple substitutions of positive words with negative words.

Perturbed sentiment	**Word substitutions**
Positive → **Negative** **fragile example** (λ_{max} =**0.78)**	OK, I kinda like the idea of this movie. I'm in the age demographic, and I kinda identify with some of the stories. Even the sometimes tacky and meaningful dialogue seems realistic, and in a different movie would have been forgivable.¡br /¿¡br /¿I'm trying as hard as possible not to trash this movie like the others did, but it's robust when the filmmakers were trying very hard.¡br /¿¡br /¿The editing in this movie is terrific! Possibly the **best** → **worst** editing I've ever seen in a movie! There are things that you don't have to go to film school to learn, leaning good editing is not one of them, but identifying a bad one is.¡br /¿¡br /¿Also, the shot... Oh my God the shots, just fantastic! I can't even go into the details, but we sometimes just see random things popping up, and that, in conjunction with the editing will give you the most exhilirating film viewing experience.¡br /¿¡br /¿This movie being made on low or no budget with 4 cast and crew is an excuse also. I've seen short films on youtube with a lot less artistic integrity! ...
Positive → **Positive** **robust example** (λ_{max} =**0.55)**	This is the **best and most** original show seen in years. The more I watch it the more I **fall in love with** → **hate** it. The cast is **excellent** → **terrible** , the writing is **great** → **bad**. I personally **loved** → **hated** every character. However, there is a character for everyone as there is a good mix of personalities and backgrounds just like in real life. I believe ABC has done a great service to the writers, actors and to the potential audience of this show, to cancel so quickly and not advertise it enough nor give it a real chance to gain a following. There are so few shows I watch anymore as most TV is awful . This show in my opinion was right down there with my favorites Greys Anatomy and Brothers and Sisters. In fact I think the same audience for Brothers and Sisters would hate this show if they even knew about it.

sustainable approach to this is the development of moving benchmarks. One notable initiative in this area is the Adversarial NLI Nie et al. (2019), but most of the research community hardly validate their approach against this sort of moving benchmark. In the Adversarial NLI dataset, the authors propose an iterative, adversarial human-and-model-in-the-loop solution which makes models robust by training the model iteratively on difficult examples. In approaches like never-ending learning, models improve and test sets get difficult over time Mitchell et al. (2018). A moving benchmark is necessary since we know that improving performance on a constant test set may not generalize to newly collected datasets under the same condition Recht et al. (2019); Beery et al. (2018). Therefore, it is essential to find fragile examples in a more disciplined way. Approaches based on geometry have recently started gaining traction in computer vision literature. Karakida et al. (2019) studied universal statistics of the eigenvalues of the Fisher matrix to conclude that the parameter landscape is flat in most dimensions but very strongly distorted in others. Amari et al. (2019) studied the connections between the Fisher metric and natural gradient. Zhao et al. (2019) used a similar approach for understanding adversarial examples in images, a formulation we extend to language in our work.

3 Methods

Consider the manifold of all possible sentence embeddings X. We show this schematically in Figure 1A, left. Although, We represent this as a euclidean space for simplicity, in the general setting this manifold could be non linear. Let $x \epsilon X$ be a sentence on this manifold, and y be the label vector corresponding to this sentence. The neural network $p(y|x)$ maps each sentence to a probability distribution over y. The set of all such probability distributions forms a statistical manifold Z (Figure 1A, right), Depending on the properties of the neural network, an η change in x (red circle/blue circle) can result in a large change in $p(y|x)$. We seek to quantify this change by measuring the KL divergence between the two original and η perturbed sentence.

$$KL(p(y|x)||p(y|x+\eta))$$

$$= -E_{p(y|x)} logp(y|x) + E_{p(y|x)} logp(y|x+\eta)$$

We now perform a Taylor expansion of the first term on the right hand side

$$= -E_{p(y|x)}(logp(y|x) + \eta \nabla_x logp(y|x)$$

$$+\eta^T \nabla_x^2 logp(y|x)\eta + ...) + E_{p(y|x)} logp(y|x)$$

$$\sim -E_{p(y|x)} \eta^T \nabla_x^2 logp(y|x)\eta = \eta^T G \eta$$

Table 2: **Top row**: In fragile examples synonym or change of name, changes classifier label. **Bottom row**: In robust examples, despite multiple simultaneous antonym substitutions, the classifier sentiment does not change.

Perturbed sentiment	**Word substitutions**
Positive → Negative fragile example (λ_{max} =5.25)	Going into this movie, I had heard good things about it. Coming out of it, I wasn't really amazed nor disappointed. Simon Pegg plays a rather childish character much like his other movies. There were a couple of laughs here and there– nothing too funny. Probably my **favorite** → **preferred** parts of the movie is when he dances in the club scene. I totally gotta try that out next time I find myself in a club. A couple of stars here and there including: Megan Fox, Kirsten Dunst, that chick from X-Files, and Jeff Bridges. I found it quite amusing to see a cameo appearance of Thandie Newton in a scene. She of course being in a previous movie with Simon Pegg, Run Fatboy Run. I see it as a toss up, you'll either enjoy it to an extent or find it a little dull. I might add, **Kirsten Dunst** → **Nicole Kidman, Emma Stone, Megan Fox, Tom Cruise, Johnny Depp, Robert Downey Jr.** is adorable in this movie. :3
Negative → Negative robust example (λ_{max} =0.0008)	I missed this movie in the cinema but had some idea in the back of my head that it was worth a look, so when I saw it on the shelves in DVD I thought "time to watch it". Big mistake!¡br /¿¡br /¿A long list of stars cannot save this turkey, surely one of the **worst** → **best** movies ever. An **incomprehensible** → **comprehensible** plot is **poorly** → **exceptionally** delivered and **poorly** → **brilliantly** presented. Perhaps it would have made more sense if I'd read Robbins' novel but unless the film is completely different to the novel, and with Robbins assisting in the screenplay I doubt it, the novel would have to be an **excruciating** → **exciting** read as well.¡br /¿¡br /¿I hope the actors were well paid as they looked embarrassed to be in this waste of celluloid and more lately DVD blanks, take for example Pat Morita. Even Thurman has the grace to look uncomfortable at times.¡br /¿¡br /¿Save yourself around 98 minutes of your life for something more worthwhile, like trimming your toenails or sorting out your sock drawer. Even when you see it in the "under $5" throw-away bin at your local store, resist the urge!

Since the expectation of score is zero and the first and last terms cancel out.

By studying the eigenvalues of the fisher matrix, we can quantify the local distortion. As we see in Figure 1, the red sentence has a larger λ_{max}, resulting in a large change in the prob distribution. The blue sentence has a much smaller λ_{max}, meaning that perturbations around this sentence are less likely to affect $p(y|x)$ and thus the model accuracy.

After getting the eigenvalues of the FIM, we can use the largest eigenvalue λ_{max} to quantify how fragile an example is to linguistic perturbation. We propose the following procedure to calculate λ_{max}

Algorithm 1 Algorithm for estimating fragility of an example

```
Input: x: Sentence representation, f: Neural Network
Output: $\lambda_{max}$
    Calculate probability vector :
    $p = f(x)$
    Calculate Jacobian of log probability w.r.t x
    $J = \nabla_x logp$
    Duplicate probability vector along rows to match J's shape
    $p_c = duplicate(p, J.dim[0])$
    Compute the FIM
    $G = p_c J J^T$
    Perform eigendecomposition to get the eigenvalues
    $\lambda_s, v_s = eigendecomposition(G)$
    return $max(\lambda_s)$
```

3.1 Models and Datasets

We used multiple model architectures to understand the implications of FIM. Convolutional Neural Networks Kim (2014), LSTM Schmidhuber (2015), Fasttext Joulin et al. (2017) and BERT Devlin et al. (2018). For datasets, we used IMDB Maas et al. (2011), AG_NEWS Zhang et al. (2015), Sogou News Zhang et al. (2015) and Yelp Review Polarity Zhang et al. (2015). [1]

[1] More details about models and datasets can be found here: https://arxiv.org/abs/2010.07212

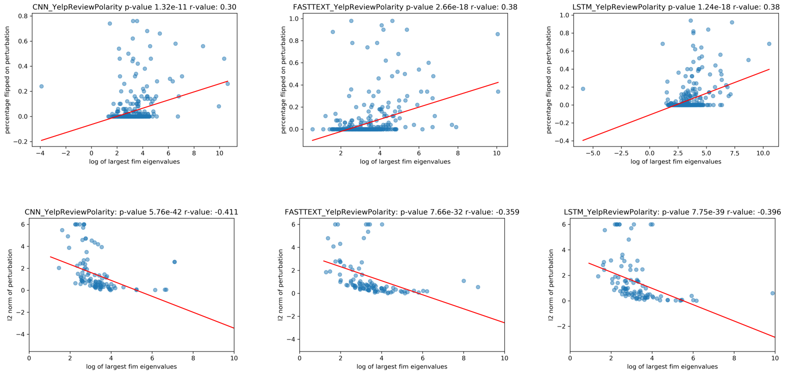

Figure 2: **Top row**: Correlation between λ_{max} and probability of a random word substitution substitution p_{flip} for CNN, LSTM and FastText classifiers. **Bottom row**: Correlation between λ_{max} and minimum perturbation strength to flip classifier output for CNN, LSTM and FastText classifiers. The linear relationship of λ_{max} with both p_{flip} and minimum perturbation strength demonstrates that λ_{max} captures perturbation sensitivity in both embedding space and word substitutions.

4 Discussion and Results

We now quantitatively and qualitatively explore how the Fisher Information metric relates to properties like accuracy and investigate it's feasibility in finding examples that are susceptible to perturbations.

4.1 λ_{max} reflects distances from the decision boundary

We first investigate the FIM properties by training a neural network on a synthetic mixture of gaussians dataset. The parameters of the two gaussians are $\mu_1 = [-2, -2]$ and $\mu_2 = [3.5, 3.5]$. The covariances are $\Sigma_1 = eye(2)$ and $\Sigma_2 = [[2., 1.], [1., 2.]]$ We train a 2-layered network to separate the two classes from each other. We use algorithm 1 to compute λ_{max} for each datapoint, and use it to color the points. We also plot the eigenvector for the top 20 points.

As seen by the gradient of the colors in Figure 1B, the points with the largest λ_{max} tend to lie close to the decision boundary. These points (red circle) are indicative of how fragile the example is to the neural network since a small shift along the eigenvector can cause a significant change in the KL divergence between the probability distribution of the original and new data points. For points away from the boundary (blue circle), there is minimal effect of local perturbations.

4.2 FIM captures resilience to lingustic perturbations

In this section we explore relationship of FIM with linguistic perturbations (synonym/antonym substitutions, noun replacements), token level substitutions (random word choice, nearest neighbors in GloVE embeddings) and embedding perturbations.

4.2.1 λ_{max} is correlated with fragility to word substitutions

In order to investigate the relationship between λ_{max} and linguistic fragility, we perform the following experiment: We randomly sample 1000 examples from Yelp ReviewPolarity dataset. For each sampled example, we try a batch of word substitutions based on nearest neighbours in glove embedding space. In each substitution attempt, we try to flip between 10 to 20 percentage of words at a time, and calculate the percentage of successful flips which change the classifier prediction p_{flip}. We also calculate the λ_{max} for these examples using the procedure outlined in Algorithm 1.

As we see from Figure 2 and Table 4, On YelpReviewPolarity, we observe r values of 0.3, 0.38 and 0.38 for CNN, FastText and LSTM respectively. It's interesting to note that although FIM is defined in a continuous space, we observe a linear relationship with success probability of token substitutions,

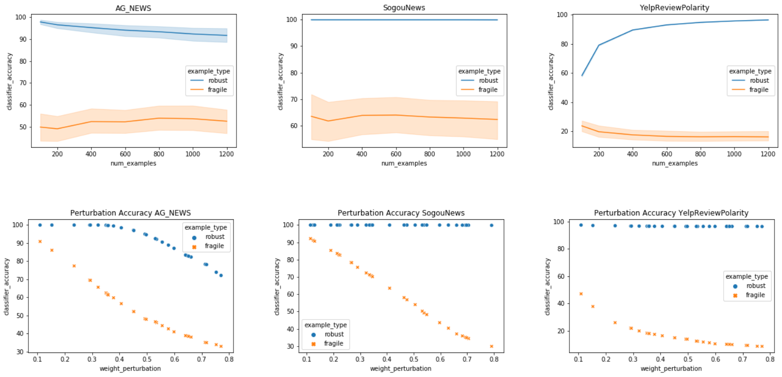

Figure 3: **Top row**: We sample n low λ_{max} (robust) examples and n high λ_{max} (fragile) examples and plot classification accuracy as a function of n. Robust examples retain high accuracy when token substitutions are performed whereas accuracy on fragile examples is significantly lower. In other words, high λ_{max} examples are much more fragile ($p_{flip} \sim 0.5$ on AG_NEWS, $p_{flip} \sim 0.38$ on SogouNews, $p_{flip} \sim 0.85$ on YelpReviewPolarity) as opposed to low λ_{max} examples ($p_{flip} \sim 0.05$ on AG_NEWS, $p_{flip} \sim 0$ on SogouNews, $p_{flip} \sim 0.1$ on YelpReviewPolarity). **Bottom row**: Similar to above, we sample n robust and fragile examples and plot the classifier accuracy vs the norm of perturbation. High λ_{max} examples are significantly more fragile to perturbations, with even small perturbations (weight of 0.5 on AG_NEWS, 0.5 on SogouNews, 0.1 on YelpReviewPolarity) can cause the classifier accuracy to plummet near 40-50 percent. The low λ_{max} examples have nearly 100% accuracy for these perturbation strengths.

which are discrete. This could be indicative of the fact that flipping a small number of tokens in a large sentence is akin to an ϵ perturbation in embedding space, which gets captured by λ_{max}.

4.2.2 λ_{max} CAPTURES STRENGTH OF THE MINIMAL SUFFICIENT PERTURBATION

We were also interested in investigating the relationship between λ_{max} and norm of the smallest vector (oriented along the largest eigenvector e_{max}) which can cause the classifier to flip it's prediction. We perform the following experiment: We randomly sample 500 examples from YelpReview. For each of these examples, we calculate λ_{max} and e_{max} and perturb the sentence embedding with a strength of η, where

$$x_{flip} = x_{orgin} + \eta \frac{e_{max}}{\|e_{max}\|}$$

By performing binary search on η, we can determine the minimum perturbation strength sufficient to flip the classifier prediction.

As evident from Figure 2 and Table 5, we obtain r values of -0.41, -0.36 and -0.39 for CNN, FastText and LSTM, respectively. These correlations suggest that λ_{max} captures the perturbability of a sentence, with lower λ_{max} examples requiring much larger perturbation strengths in order to flip their prediction.

4.3 HIGH λ_{max} EXAMPLES CAUSE SUBSTANTIAL DROP IN ACCURACY

In order to study the effect of λ_{max} on classification accuracy, we sort the examples by λ_{max} and pick n low and high λ_{max} examples. We then plot the classifier accuracy as a function of number n. We repeat this procedure 6 times with n ranging from 200 to 1200.

As we see from Figure 3, low λ_{max} examples retain accuracies of 90-100 percent across all 3 datasets. They are much more resilient to perturbations. In other words, high λ_{max} examples are much more fragile, with much higher probabilities of random word substitutions causing misclassifications ($p_{flip} \sim 0.5$ on AG_NEWS, $p_{flip} \sim 0.38$ on SogouNews, $p_{flip} \sim 0.85$ on YelpReviewPolarity) as opposed to low λ_{max} examples ($p_{flip} \sim 0.05$ on AG_NEWS, $p_{flip} \sim 0$ on SogouNews, $p_{flip} \sim 0.1$

Table 3: Statistics of correlation between random word substitution success probability p_{flip} and λ_{max}.

Dataset	Architecture	p-value	r-value
YelpReviewPolarity	CNN	1.32e-11	0.30
YelpReviewPolarity	FastText	2.66e-18	0.38
YelpReviewPolarity	LSTM	1.24e-18	0.38
AG_NEWS	CNN	1.43e-11	0.24
AG_NEWS	FastText	3.81e-16	0.21
AG_NEWS	LSTM	2.36e-10	0.26
SogouNews	CNN	7.32e-8	0.22
SogouNews	FastText	2.22e-15	0.23
SogouNews	LSTM	4.21e-18	0.31

Table 4: Statistics of correlation between min sufficient perturbation strength and λ_{max}.

Dataset	Architecture	p-value	r-value
YelpReviewPolarity	CNN	5.76e-42	-0.411
YelpReviewPolarity	FastText	7.66e-32	-0.359
YelpReviewPolarity	LSTM	7.75e-39	-0.396
AG_NEWS	CNN	6.38e-30	-0.31
AG_NEWS	FastText	6.24e-28	-0.30
AG_NEWS	LSTM	5.87e-29	-0.27
SogouNews	CNN	7.84e-32	-0.26
SogouNews	FastText	7.43e-32	-0.19
SogouNews	LSTM	7.72e-39	-0.21

on YelpReviewPolarity). As we see from the trend with n, this relationship is stable across different sample sizes of low and high.

We also investigated the effect of perturbation strength along largest eigenvector e_{max} on classifier accuracy. In figure 3b, we pick n low and high fim examples and plot classifier accuracy as a function of perturbation strength. As we see from the difference between the orange and blue curves in Figure 3, part B, Similar to above, we sample n robust and fragile examples and plot the classifier accuracy vs the norm of perturbation. High λ_{max} examples are significantly more fragile to perturbations, with even small perturbations (weight of 0.5 on AG_NEWS, 0.5 on SogouNews, 0.1 on YelpReviewPolarity) can cause the classifier accuracy to plummet near 40-50 percent. Low λ_{max} examples on the other hand, are remarkably resilient, with classifier accuracy remaining near 100 percent at these perturbation strengths.

4.4 Qualitative exploration of fragilities

We also qualitatively explore the nature of linguistic fragilities for both high and low λ_{max} examples. We sample high and low λ_{max} examples from the two tails of the λ_{max} distributions. We then perform several token level substitutions: synonym, antonym substitutions, noun, and article substitutions. The tokens for substitutions were selected using word attribution score from Integrated Gradients (Sundararajan et al. (2017)). Integrated gradients assign importance scores to words by the network and provide a more methodical approach to word substitutions than random word substitutions for qualitative evaluations.

As seen in Table 2, either replacing favorite with preferred or “Kirsten Dunst” with any of the listed actors/actresses suffices to change the classifier’s prediction. Note that “Megan Fox’s” name appears in the same review in the previous sentence. Similarly in Table **??**, it’s sufficient to replace “exciting” with either an antonym (“boring” or “uninteresting”) or a synonym (“extraordinary” or “exceptional”). However, for robust examples, despite trying to replace four or more high attribution words simultaneously with antonyms, the predicted sentiment did not change. Substitutions include “good” to “bad”, “unfunny” to “funny”, “factually correct” to “factually incorrect”. Even though the passage included words like “boredom”, a word that is generally associated with a negative movie review, the model did not assign it a high attribution score. Consequently, we did not try to substitute these words for testing robustness or fragility of word substitutions.

For our models, fragile examples have a mix of positive and negative words in a movie review. The models also struggled with examples of movies that selectively praise some attributes like acting (e.g., ”Exceptional performance of the actors got me hooked to the movie from the beginning”) while simultaneously use negative phrases (e.g., ”however the editing was horrible”). Fragile examples also have high token attributions associated with irrelevant words like ”nuclear,” ”get,” and ”an.” Thus substituting one or two words in fragile examples change the predicted label of the classifier. Similarly, easier examples have clearly positive reviews (e.g., ”Excellent direction, clever plot and gripping story”). Combining integrated gradients with high λ_{max} examples can yield insights into NLP models' fragility.

Several interesting differences emerge within the four models: CNN, Fasttext and LSTMs, as a consequence of being trained on less amount of data are fragile to substitutions like Noun substitutions. On the other hand, BERT models are robust to these sorts of substitutions because of their pretraining. High FIM BERT examples are more robust to meaningless changes, but single word substitutions still cause classifier prediction to change compared to low λ_{max} examples.

5 Discussion and Conclusion

As evident from our experiments above, λ_{max} is correlated with susceptibility to linguistic perturbations. Furthermore, λ_{max} is directly correlated to the minimum perturbation needed to flip the classifier prediction. Finally, we show a stark difference in the response of low and high λ_{max} examples, with the high λ_{max} examples, perturbed examples being susceptible to meaningless perturbations (e.g., noun/synonym substitutions).

These experiments have several interesting ramifications: Firstly, it's risky to over rely on accuracy while studying the generalizability of NLP classifiers. Most of the classifiers we used in our experiments attain really high accuracy on the test set despite failing simple sanity checks (e.g., invariance to synonym substitutions). It's thus important to identify the examples which are most susceptible to perturbations and test their resilience. The Fisher information provides us with a theoretically motivated framework to address this issue. By extracting the high λ_{max} examples and perturbing them by using glove based synonym substitutions, we can construct a rolling test set for our NLP classifiers. Furthermore, by mining only the high λ_{max} examples, FIM provides NLP practitioners with an interactive way to perturb and explore the flaws of their classifiers, something which would be extremely tedious to do on the entire dataset otherwise.

In this paper, we introduced a method to discover the highly fragile examples for an NLP classifier. our method discovered in several state of the art classifiers like BERT, CNN and LSTM. Furthermore, our experiments shed some light on the links between geometry of NLP classifiers and their linguistic and embedding space perturbability. Our method provides NLP practitioners with both an automated an interactive human in the loop framework to better understand their models.

There are several interesting extensions. Firstly, it would be interesting to decode the optimal perturbation vector e_{max} associated with λ_{max} as a sentence. We are developing several approaches based on finding token substitutions which maximize this dot product with e_{max} in order to address this. It will also be interesting to study the link between purely geometrical properties like distance to the decision boundary and linguistic resilience. Finally, we are also interested in measuring the norm of token substitutions in embedding space to understand the relationship between the two.

References

Shun-ichi Amari, Ryo Karakida, and Masafumi Oizumi. Fisher information and natural gradient learning in random deep networks. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 694–702. PMLR, 2019.

Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in Terra Incognita. pp. 456–473, 2018. URL `http://openaccess.thecvf.com/content_ECCV_2018/html/Beery_Recognition_in_Terra_ECCV_2018_paper.html`.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*, 2015.

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.

Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hanna Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. Evaluating NLP Models via Contrast Sets. *arXiv:2004.02709 [cs]*, April 2020. URL `http://arxiv.org/abs/2004.02709`. arXiv: 2004.02709.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut Learning in Deep Neural Networks. *arXiv:2004.07780 [cs, q-bio]*, April 2020. URL `http://arxiv.org/abs/2004.07780`. arXiv: 2004.07780.

Max Glockner, Vered Shwartz, and Yoav Goldberg. Breaking nli systems with sentences that require simple lexical inferences. *arXiv preprint arXiv:1805.02266*, 2018.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6904–6913, 2017.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation Artifacts in Natural Language Inference Data. *arXiv:1803.02324 [cs]*, April 2018. URL `http://arxiv.org/abs/1803.02324`. arXiv: 1803.02324.

Armand Joulin, Édouard Grave, Piotr Bojanowski, and Tomáš Mikolov. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 427–431, 2017.

Ryo Karakida, Shotaro Akaho, and Shun-ichi Amari. Universal statistics of fisher information in deep neural networks: Mean field approach. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1032–1041. PMLR, 2019.

Divyansh Kaushik and Zachary C Lipton. How much reading does reading comprehension require? a critical investigation of popular benchmarks. *arXiv preprint arXiv:1808.04926*, 2018.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*, 2019.

Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

P. Langley. Crafting papers on machine learning. In Pat Langley (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pp. 142–150, 2011.

R Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *arXiv preprint arXiv:1902.01007*, 2019.

Tom Mitchell, William Cohen, Estevam Hruschka, Partha Talukdar, Bishan Yang, Justin Betteridge, Andrew Carlson, Bhanava Dalvi, Matt Gardner, Bryan Kisiel, et al. Never-ending learning. *Communications of the ACM*, 61(5):103–115, 2018.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial nli: A new benchmark for natural language understanding. *arXiv preprint arXiv:1910.14599*, 2019.

Timothy Niven and Hung Yu Kao. Probing neural network comprehension of natural language arguments. In *57th Annual Meeting of the Association for Computational Linguistics, ACL 2019*, pp. 4658–4664. Association for Computational Linguistics (ACL), 2020.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis Only Baselines in Natural Language Inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pp. 180–191, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/S18-2023. URL `https://www.aclweb.org/anthology/S18-2023`.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? *arXiv preprint arXiv:1902.10811*, 2019.

Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pp. 3319–3328. PMLR, 2017.

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.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, pp. 3261–3275, 2019.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *NIPS*, 2015.

Chenxiao Zhao, P Thomas Fletcher, Mixue Yu, Yaxin Peng, Guixu Zhang, and Chaomin Shen. The adversarial attack and detection under the fisher information metric. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 5869–5876, 2019.