

# Transfer Learning Robustness in Multi-Class Categorization by Fine-Tuning Pre-Trained Contextualized Language Models

Written by Xinyi Liu<sup>1\*</sup> and Artit Wangperawong<sup>2\*</sup>

<sup>1</sup>University of Rochester, xinyi.liu1@ur.rochester.edu

<sup>2</sup>U.S. Bank, artit.wangperawong@usbank.com

## Abstract

This study compares the effectiveness and robustness of multi-class categorization of Amazon product data using transfer learning on pre-trained contextualized language models. Specifically, we fine-tuned BERT and XLNet, two bidirectional models that have achieved state-of-the-art performance on many natural language tasks and benchmarks, including text classification. While existing classification studies and benchmarks focus on binary targets, with the exception of ordinal ranking tasks, here we examine the robustness of such models as the number of classes grows from 1 to 20. Our experiments demonstrate an approximately linear decrease in performance metrics (i.e., precision, recall,  $F_1$  score, and accuracy) with the number of class labels. BERT consistently outperforms XLNet using identical hyperparameters on the entire range of class label quantities for categorizing products based on their textual descriptions. BERT is also more affordable than XLNet in terms of the computational cost (i.e., time and memory) required for training.

## Introduction

Classification involves the prediction of one outcome from two or more possible discrete outcomes. A classification model predicts the probabilities of different possible outcomes of a categorically distributed dependent variable for a set of independent variables as input. Binary classification involves two possible outcomes, whereas multinomial (i.e., multi-class) classification involves three or more possible outcomes. Such models are based on the assumption of independence of irrelevant alternatives (McFadden, Tye, and Train 1977), implying that the probability of preferring one class over another does not depend on the presence or absence of other irrelevant alternatives. For example, adding the class label *Beauty* would not affect the model’s relative output probabilities for classifying the example to be labeled under other categories. In practice, this assumption can be violated when a newly introduced class label correlates with any of the

existing labels, e.g. *Beauty* vs. *Health and Personal Care*, as a given product may reasonably be multi-labeled. Since many interesting classification problems can involve a plethora of class labels for categorization, it is important to understand how the number of classes affects model performance.

Intuitively, one would expect a classification model’s performance to decrease as the number of classes increases. After all, it would be more challenging even for humans to pick the right choice among more available options to categorize an item. Despite what we would imagine is a commonly occurring scenario, there is a dearth of studies or literature that closely examines the robustness of a certain model as the number of classes increases.

## Background and Related Work

Existing studies that compare the performance of different models in multi-class categorization problems usually only examine a chosen couple high number of classes (Uglov et al. 2008; Chamasemani and Singh 2011; Hasan et al. 2016). With such few data points, we would not be confident in extrapolating a trend or judging the performance between different models and parameters. As we show in this work, **Model A** might outperform **Model B** at a certain number of class labels but might underperform by adding or removing just one or two class labels (see Figs. 1 and 2). To conduct a comprehensive study, one would need machine learning models that can independently achieve high performance on each of the individual class labels. Applying transfer learning to state-of-the-art pre-trained models can be a feasible approach. Here we thus report on our experiments involving multi-class categorization by fine-tuning pre-trained contextualized language models to classify e-commerce products based on their textual descriptions on Amazon.

BERT and XLNet are two models that have recently achieved high performance on many natural language benchmarks, including text classification for positive and negative sentiments and ratings (Devlin et al. 2018; Yang et al. 2019). BERT-based models have lead the score boards for many NLP benchmark tasks since October 2018. More recently in June 2019, XLNet was shown to outperform BERT-related models on SQuAD, GLUE, RACE and 17 datasets, achieving record results on a total of 18 tasks. Only one month later in

\*Equal contribution, alphabetically by last name. A.W. proposed fine-tuning BERT and XLNet under identical hyperparameters on Amazon data to evaluate multi-class text classification performance and implemented the initial codebase. X.L. extended the codebase and performed comprehensive research, experiments and analyses.

July 2019, an improved model called RoBERTa was demonstrated based on BERT with more training and removing an unnecessary training task (i.e., next-sentence prediction) (Liu et al. 2019). With RoBERTa, the effect of data supply and hyperparameters were compared between XLNet and BERT, raising the question of whether one model is better than the other. In this study, we evaluated both models trained with fixed data supply and hyperparameters (e.g. learning rate, batch size, training steps etc.). The robustness of BERT and XLNet on multinomial classification was evaluated by categorizing textual product descriptions into product categories using Amazon product data (McAuley et al. 2015).

BERT stands for Bidirectional Encoder Representations from Transformers. This model learns deep bidirectional representations from unlabeled text by jointly conditioning on both left and right contexts in all layers. BERT’s pre-training includes the Masked Language Model (MLM) where some of the tokens in the input are randomly masked for the model to predict the vocabulary id of the masked tokens given both left and right contexts. BERT also utilizes a next-sentence prediction task to train text-pair representations, which was demonstrated by RoBERTa to be unnecessary and disadvantageous for certain tasks. In this paper, we adopted the pre-trained BERT base model that includes 12 layers of a bidirectional transformer encoder (Vaswani et al. 2017) each with a hidden size of 768 units and 12 self-attention heads to comprise a total of 110 million parameters. The first token of each sequence inputted into the model is always a special token: [CLS]. The final hidden state of this token is treated as the final aggregate sequence representation of the sequence for the final classification.

XLNet is a generalized autoregressive pre-training model. Instead of relying on masking the input to train the fused representation for the left and right contexts as in BERT, XLNet learns the bidirectional context by maximizing the expected log likelihood over all permutations of the factorization order. As the permutations can make the tokens from both the left and right contexts available at one side of the positions, XLNet overcomes the limitations of traditional autoregressive models that can be only trained unidirectionally to represent context and then concatenate them. It also avoids potential pretrain-finetune discrepancy caused by data corruption from masking tokens in BERT. Though permuted, XLNet encodes the positions to keep the positional information. XLNet adopts the recurrence mechanism and relative positional encoding scheme of Transformer-XL (Dai et al. 2019) to learn dependency beyond a fixed length and reparameterizes it to remove the ambiguity of the factorization order.

Given a set of target tokens  $T$  and a set of non-target tokens  $N = \mathbf{x} \setminus T$ , BERT and XLNet both maximize  $\log p(T | N)$  but with different formulations:

$$J_{\text{BERT}} = \sum_{x \in T} \log p(x | N), \quad (1)$$

$$J_{\text{XLNet}} = \sum_{x \in T} \log p(x | N \cup T_{<x}), \quad (2)$$

where  $T_{<x}$  denote tokens in  $T$  that have a factorization order

Table 1: Categories from Amazon products numbered in the order in which they are included in the models produced (Figs. 2 and 3).

Order ( $j$ )	Categories
1	Musical instruments
2	Baby
3	Patio, Lawn and Garden
4	Grocery and Gourmet Food
5	Automotive
6	Pet Supplies
7	Office Products
8	Beauty
9	Tools and Home Improvement
10	Toys and Games
11	Health and Personal Care
12	Cell Phones and Accessories
13	Sports and Outdoors
14	Kindle Store
15	Home and Kitchen
16	Clothing, Shoes, and Accessories
17	CDs and Vinyl
18	Movies and TV
19	Electronics
20	Books

prior to  $x$ . We extended both pre-trained models by adding a softmax layer with the appropriate number of class outputs,  $K$ , to the final layer of the pre-trained models, resulting in output probabilities for the  $j^{\text{th}}$  class according to

$$p(y = j | \mathbf{q}) = \frac{e^{\mathbf{q}^T \mathbf{w}_j}}{\sum_{k=1}^K e^{\mathbf{q}^T \mathbf{w}_k}}, \quad (3)$$

where  $K \in [1..20]$ ,  $\mathbf{q}$  is the output from the final layer of the pre-trained model that is input into the softmax function, and  $\mathbf{w}$  is the weight parameter to be trained.

## Experimental Methods

We finetuned the publicly available base models for BERT and XLNet (i.e., **BERT-Base** and **XLNet-Base**, respectively) with the above modifications for text classification on up to 20 different class labels. Both models have 12 transformer layers each with a hidden size of 768 units and 12 self-attention heads.

For each additional class label, we added 5000 randomly sampled textual descriptions from the  $j^{\text{th}}$  category in the order shown in Table 1. A random 10% of all resulting samples were further held out for testing purposes and excluded from training. Although a product might reasonably fall under multiple categories and sub-categories, only the highest level label is used as the target. We only included descriptions with more than five characters to ensure sufficient textual information for the model inputs.

For tokenization, we used BERT’s pre-trained WordPiece tokenizer (Wu et al. 2016) and XLNet’s pre-trained SentencePiece tokenizer (Kudo and Richardson 2018). We used the

Table 2: Hyperparameter settings used to fine-tune BERT and XLNet pre-trained models.

Hyperparameter	Value
Transformer layers	12
Self-attention heads per layer	12
Hidden size per layer	768
Max sequence length ( $L$ )	128
Train-test split ratio ( $f$ )	0.9
Samples per class ( $N_{samples}$ )	5000
Number of epochs ( $N_{epochs}$ )	3
Mini-batch size ( $B$ )	32
Learning rate ( $\epsilon$ )	2e-5
Dropout rate ( $p_{dropout}$ )	0.1

uncased model for BERT and the cased model for XLNet according to what their respective authors reported to perform better. We developed the fine-tuned models using TensorFlow (Abadi et al. 2015), and then trained it on an Nvidia Tesla T4 GPU for  $n$  training steps, which is given by

$$n = \frac{f \cdot K \cdot N_{samples}}{B} \cdot N_{epochs}, \quad (4)$$

where the hyperparameters  $f$ ,  $N_{samples}$ ,  $N_{epochs}$ , and  $B$  are the train-test split ratio, number of samples per class, number of epochs, and mini-batch size, respectively, defined in Table 2.

Having sampled balanced data from each class  $j$  to avoid bias, we evaluated both models in terms of the macro-averaged precision, recall,  $F_1$  score, and accuracy (Sokolova and Lapalme 2009) of the test set. To compute the macro-average, we first computed the metrics for each class. An example for  $K = 20$  is shown in Fig. 1. For class  $j$ , the precision, recall and  $F_1$  score are calculated respectively as

$$precision_j = \frac{TP_j}{TP_j + FP_j}, \quad (5)$$

$$recall_j = \frac{TP_j}{TP_j + FN_j}, \quad (6)$$

$$F_{1j} = \frac{2 \cdot precision_j \cdot recall_j}{precision_j + recall_j}, \quad (7)$$

where  $TP_j$  denotes the number of true positive predictions,  $FP_j$  denotes the number of false positive predictions and  $FN_j$  denotes the number of false negative predictions for class  $j$ . We then calculated the average of the precisions, recalls and  $F_1$  scores over all the classes as the macro-averaged precision, recall and  $F_1$  score. The formulas with  $K$  total classes are respectively

$$precision = \frac{\sum_{j=1}^K precision_j}{K}, \quad (8)$$

$$recall = \frac{\sum_{j=1}^K recall_j}{K}, \quad (9)$$

$$F_1 = \frac{\sum_{j=1}^K F_{1j}}{K}, \quad (10)$$

and the accuracy of the test set is calculated as

$$accuracy = \frac{\sum_{j=1}^K TP_j}{(1 - f) \cdot K \cdot N_{samples}}. \quad (11)$$

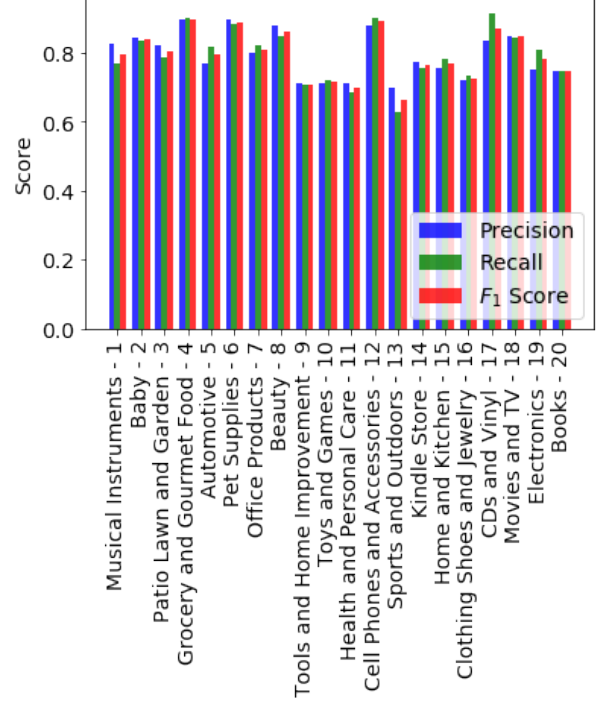


Figure 1: XLNet precisions, recalls, and  $F_1$  scores of each category before averaging for  $K = 20$ .

## Results and Discussion

All performance metrics on the test set decreases approximately linearly as the number of classes increases. The trend suggests that probability of incorrectly categorizing an item is proportional to the number of classes, i.e.

$$p(y \neq j|x_j) \propto K. \quad (12)$$

For completeness, the trivial case of a single class is included. Since we used balanced data for each class, the macro-averaged recall should equal accuracy. However, due to rounding and floating point precision, negligible differences exist. It is also not surprising to see that the precisions and  $F_1$  scores are similar to the accuracies. The linearly decreasing trend is preferable over an otherwise faster decreasing function, such as an exponential decay. However, it is possible that the initial linear trend can become non-linear with more classes.

Fitted lines result in coefficients of determination ( $R^2$ ) close to 1, which indicate good fits in the range of interest. However, the dispersion of the data points appears to increase with the number of classes. Extrapolating performance would be therefore less accurate with higher number of classes. Deviations from the fitted line could be due to a

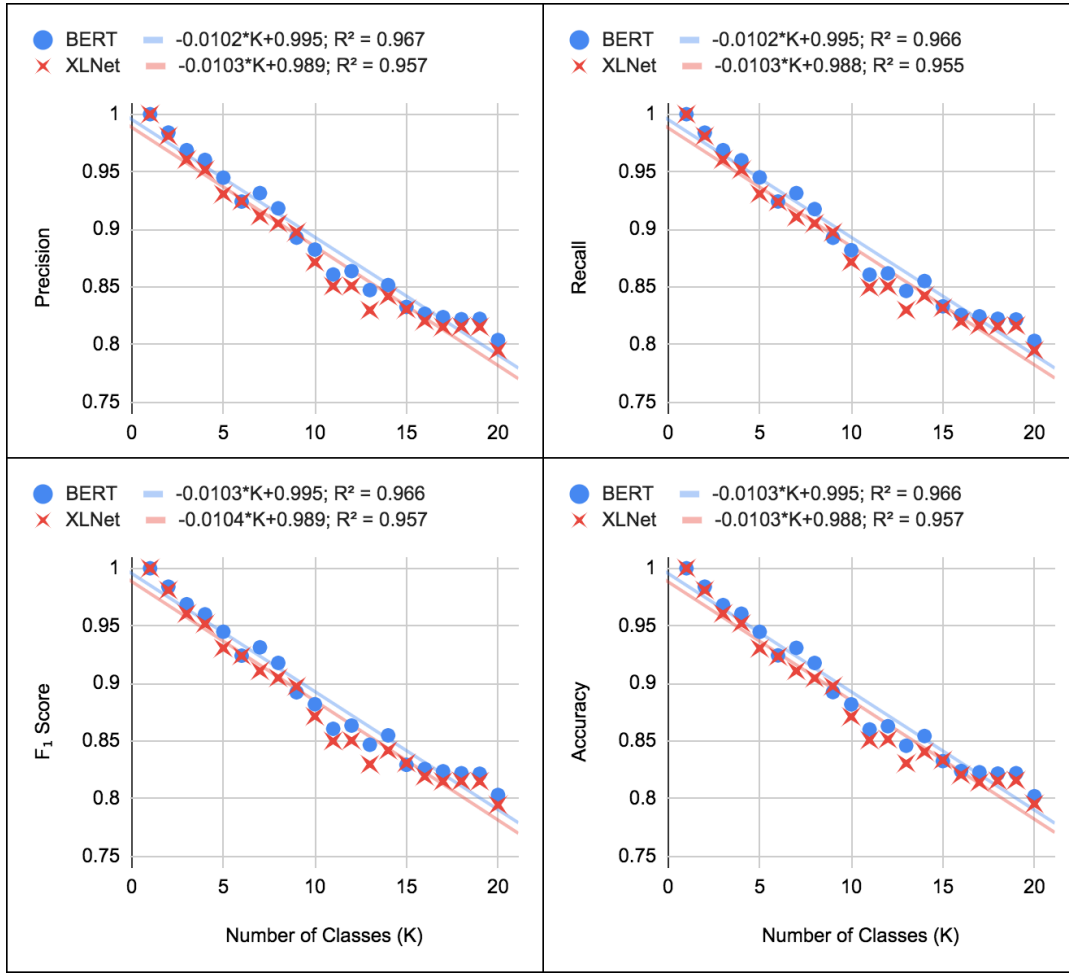


Figure 2: Macro-averaged precision, recall,  $F_1$  score, and accuracy of the test set versus the number of classes  $K$  for pre-trained **BERT-Base** and **XLNet-Base** language models fine-tuned for multi-class categorization. For completeness, the trivial case of a single class is included.

variety of factors, including underfitting, overfitting, violating the assumption of independence of irrelevant alternatives as described above, and multi-label tendencies of a particular class.

Multi-label tendencies refers to an item reasonably having more than one label. Since this study is strictly multi-class categorization, the models have to choose one label for each item even if there is definition overlap between two or more class labels. For example, there could be items appropriately labeled as either *Health and Personal Care* or *Beauty*.

Some classes perform more poorly than others, dragging down the macro-averaged metrics below the fitted line. As shown in Figs. 2 and 3, this is most evident when  $K = 11$  and  $K = 13$ . A simple reason is that the respective latest included categories are more difficult to predict, i.e. *Health and Personal Care* and *Sports and Outdoors* (Fig. 1). Both BERT and XLNet exhibit the same trends, though BERT outperforms XLNet on all performance metrics over the entire range of class label quantities (Fig. 2). We show that a more performant model can reduce this discrepancy by reproduc-

ing the experiments described above on **XLNet-Large** (see Fig. 3).

XLNet-Large is a much larger model than either XLNet-Base or BERT-Base, containing 24 transformer layers each with a hidden size of 1024 units and 16 self-attention heads. It also took approximately quadruple the time and triple the memory required for training XLNet-Base under the same experimental conditions described above. An exception was that we had to reduce the batch size to 8 to avoid memory limits. The number of training steps was maintained according to the formula described in Eqn. 4. Despite our efforts, we were unable to obtain the results for all the number of classes due to computational resource limitations. Nevertheless, XLNet-Large is more robust in terms of performance with higher number of class labels. XLNet-Large is also better than BERT-Base in maintaining the linear trend; for example, see  $K = 6$  in Fig. 3. Even with the performance boost and discrepancy reduction offered by using this larger model, however, BERT-Base remains competitive over many values of  $K$  examined.

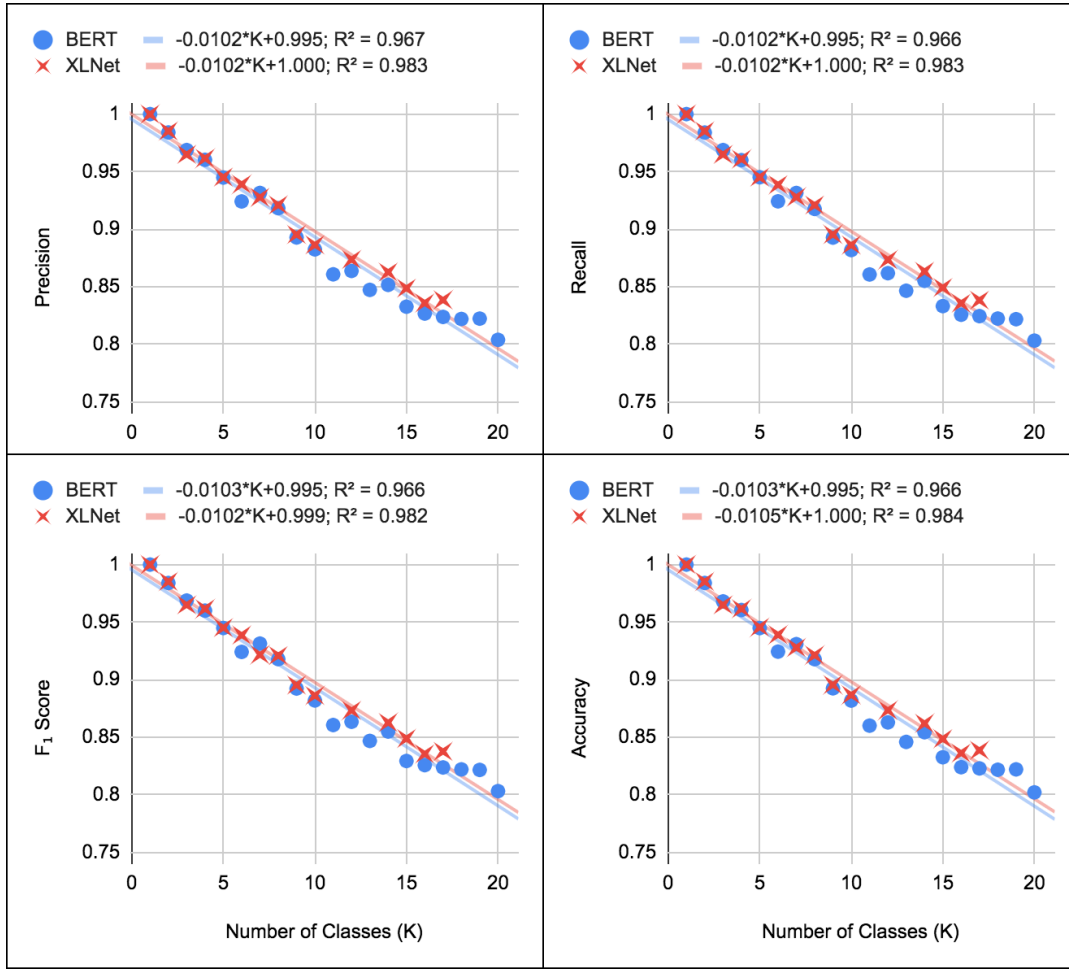


Figure 3: Macro-averaged precision, recall,  $F_1$  score, and accuracy of the test set versus the number of classes  $K$  for pre-trained **BERT-Base** and **XLNet-Large** language models fine-tuned for multi-class categorization. For completeness, the trivial case of a single class is included. Note that XLNet-Large data points were not available for all the number of classes due to computational resource limitations.

## Conclusion and Future Work

For the first time, we conducted experiments to understand the effect of including additional classes on multi-class categorization over a considerable range. We used identical hyperparameters to fairly compare fine-tuned BERT and XLNet models. BERT-Base outperformed XLNet-Base over the entire range studied with about 60% less training time and about 85% less training memory. In order for XLNet to be competitive with BERT, we increased the number of parameters for the model by implementing XLNet-Large, quadrupling time and tripling memory requirements. However, BERT still remained competitive. In all cases, the models exhibited approximately linearly decreasing trends in performance metrics with the number of class labels.

To improve performance further, future work can explore the hyperparameter space. The number of unique training samples per class can be increased to the tens of thousands using the same Amazon product dataset. More computational resources can be applied. For instance, **BERT-Large** with

whole-word masking, more training steps, and larger batch sizes can be used. Multi-modal fusion models can be developed by combining our fine-tuned models with image classifiers, which can also be based on transfer learning of state-of-the-art models (Wirojwatanakul and Wangperawong 2019). Character-level models based on transformers can also be considered (Wangperawong 2018). In all cases, it would also be useful to study more than 20 classes to confirm if the linear trend persists.

## References

- [Abadi et al. 2015] Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G. S.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Goodfellow, I.; Harp, A.; Irving, G.; Isard, M.; Jia, Y.; Jozefowicz, R.; Kaiser, L.; Kudlur, M.; Levenberg, J.; Mané, D.; Monga, R.; Moore, S.; Murray, D.; Olah, C.; Schuster, M.; Shlens, J.; Steiner, B.; Sutskever, I.; Talwar, K.; Tucker, P.; Vanhoucke, V.; Vasudevan, V.; Viégas, F.; Vinyals, O.; Warden, P.; Wattenberg, M.;

- Wicke, M.; Yu, Y.; and Zheng, X. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.
- [Chamasemani and Singh 2011] Chamasemani, F. F., and Singh, Y. P. 2011. Multi-class support vector machine (svm) classifiers – an application in hypothyroid detection and classification. In *Proceedings of the 2011 Sixth International Conference on Bio-Inspired Computing: Theories and Applications*, BIC-TA '11, 351–356. Washington, DC, USA: IEEE Computer Society.
- [Dai et al. 2019] Dai, Z.; Yang, Z.; Yang, Y.; Carbonell, J. G.; Le, Q. V.; and Salakhutdinov, R. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *CoRR* abs/1901.02860.
- [Devlin et al. 2018] Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR* abs/1810.04805.
- [Hasan et al. 2016] Hasan, M.; Kotov, A.; Idalski Carcone, A.; Dong, M.; Naar, S.; and Brogan Hartlieb, K. 2016. A study of the effectiveness of machine learning methods for classification of clinical interview fragments into a large number of categories. *J. of Biomedical Informatics* 62(C):21–31.
- [Kudo and Richardson 2018] Kudo, T., and Richardson, J. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *CoRR* abs/1808.06226.
- [Liu et al. 2019] Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR* abs/1907.11692.
- [McAuley et al. 2015] McAuley, J. J.; Targett, C.; Shi, Q.; and van den Hengel, A. 2015. Image-based recommendations on styles and substitutes. *CoRR* abs/1506.04757.
- [McFadden, Tye, and Train 1977] McFadden, D.; Tye, W. B.; and Train, K. 1977. *An application of diagnostic tests for the independence from irrelevant alternatives property of the multinomial logit model*. Institute of Transportation Studies, University of California Berkeley.
- [Sokolova and Lapalme 2009] Sokolova, M., and Lapalme, G. 2009. A systematic analysis of performance measures for classification tasks. *Information processing & management* 45(4):427–437.
- [Uglov et al. 2008] Uglov, J.; Jakaite, L.; Schetinin, V.; and Maple, C. 2008. Comparing robustness of pairwise and multi-class neural-network systems for face recognition. *EURASIP Journal on Advances in Signal Processing* 2008:64.
- [Vaswani et al. 2017] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. *CoRR* abs/1706.03762.
- [Wangperawong 2018] Wangperawong, A. 2018. Attending to mathematical language with transformers. *CoRR* abs/1812.02825.
- [Wirojwatanakul and Wangperawong 2019] Wirojwatanakul, P., and Wangperawong, A. 2019. Multi-label product categorization using multi-modal fusion models. *CoRR* abs/1907.00420.
- [Wu et al. 2016] Wu, Y.; Schuster, M.; Chen, Z.; Le, Q. V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; Klingner, J.; Shah, A.; Johnson, M.; Liu, X.; Kaiser, L.; Gouws, S.; Kato, Y.; Kudo, T.; Kazawa, H.; Stevens, K.; Kurian, G.; Patil, N.; Wang, W.; Young, C.; Smith, J.; Riesa, J.; Rudnick, A.; Vinyals, O.; Corrado, G.; Hughes, M.; and Dean, J. 2016. Google’s neural machine translation system: Bridging the gap between human and machine translation. *CoRR* abs/1609.08144.
- [Yang et al. 2019] Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J. G.; Salakhutdinov, R.; and Le, Q. V. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR* abs/1906.08237.