# Methods for Estimating and Improving Robustness of Language Models

**Michal Štefánik**
Faculty of Informatics, Masaryk University
stefanik.m@mail.muni.cz

## Abstract

Despite their outstanding performance, large language models (LLMs) suffer notorious flaws related to their preference for simple, surface-level textual relations over full semantic complexity of the problem. This proposal investigates a common denominator of this problem in their weak ability to generalise outside of the training domain. We survey diverse research directions providing estimations of model generalisation ability and find that incorporating some of these measures in the training objectives leads to enhanced distributional robustness of neural models. Based on these findings, we present future research directions towards enhancing the robustness of LLMs.

## 1 Introduction

The advances in language processing that we observe in recent years, mostly led by the instances of large language models (LLMs) based on the transformer architecture (Vaswani et al., 2017) raise a deserved attention of the scientific community. We find studies concluding that LLMs fine-tuned for a specific task can align with, or even outperform human accuracy on complex tasks such as question answering (Rajpurkar et al., 2016), paraphrase identification (Bowman et al., 2015), machine translation (Bahdanau et al., 2016) and others.

In contrast, critical studies demonstrate that many of the models reaching a state-of-the-art on a given task perform poorly on data sets drawn from different distribution(s). This is due to various reasons, such as training data set biases including spurious linguistic correlations (McCoy et al., 2019), different text stylistics or typos (Belinkov and Bisk, 2018), where a broad preference of LLMs towards fitting non-representative, yet easy-to-learn surface-level relations cause them to under-perform even shallow networks (Bojanowski et al., 2016). A lack of generalisation can also be caused by procedural reasons, such as training process instability, causing a convergence to local minima of distinct generalisation quality (McCoy et al., 2020). Low robustness of the consequential model towards out-of-distribution (OOD) samples limits their practical usability to the samples drawn from the training distribution, which is often impossible to ensure.

Despite that the complex language models strike an impression of a black-box, an extensive branch of research demonstrated that internal representations of LLMs correspond well to a human taxonomy in terms of morphological and syntactic decomposition (Clark et al., 2019a), or that the depth of the internal representation correlates well with the complexity of the problem as perceived by humans (Tenney et al., 2019).

The reported agility support the central presumption of this proposal; that LLMs can avoid the problems mentioned above under additional *regularisation*. We argue that such regularisation could also strenghten the implicit property of LLMs learning compositional language features and thus enhance an *interpretability* of their decision-making.

In this proposal, we survey literature from the broader area of neural networks for the reasons for better generalisation of the neural model. We find that many measures reported to correlate well with model's OOD performance can also enhance neural model generalisation when utilised within the model's training objective, as regularisers, or additional components of the training cost function. Inspired by this finding, this proposal outlines a path towards identification and utilisation of generalisation measures aimed to enhance robustness of LLMs towards distribution shift.

**RQ1:** "Can we *estimate* the performance of LLMs on data from OOD, without a collection of annotated data or expert feedback?"

**RQ2:** "Can we *adjust* the process of training LLMs to perform *better* on OOD samples?"



*Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 44 - 51

In Section 2.1 we survey the studies aiming to estimate robustness of neural models with no restrictions on a domain of application. Subsequently, in Section 2.2, we survey the training techniques reported to enhance the robustness of the trained model. Based on these findings, in Section 3 we identify promising directions and respective challenges specific for estimating (§3.1) and enhancing (§3.2) the robustness of LLMs.

### 1.1 Applicability

This proposal grounds the notion of model generalisation to its ability to perform well on samples drawn from distributions different than the training distribution (OOD). In this context, the term of a *distribution*, used interchangeably with *domain*, is commonly described by a specific shared property, such as topic, style, genre, or linguistic register (Ramponi and Plank, 2020).

This proposal focuses on distributional robustness in two branches of applications of current LLMs: *generative tasks*, where the problem is to generate a sequence of tokens, and *discriminative tasks*, where the task is to infer a discrete decision for each token or a sequence of tokens. Generative tasks include summarization, dialogue generation or machine translation, while discriminative tasks include classification, extractive question answering or named entity recognition.

In both cases, we propose to estimate the impact of given adjustment on model generalisation by measuring a difference in the model's performance on a set of distinct OOD domains. We note that such estimation is still only a pointwise estimation of model generalisation as some properties of the domains drawn for evaluation remain uncontrolled.

## 2 Background

### 2.1 Estimating Model Robustness (RQ1)

Having a set of true labels for some set of OOD samples $X_t$ of target domain(s) $D_t$, the robustness of the model $M$ can be estimated using standard qualitative measures, such as accuracy. This raises questions about the representativeness of the draw of $X_t$: do these cover *all* the domains of application of $M$, and are these domains accurately weighted in evaluation?

The problem is circumvented by generalisation measures based on *latent properties* of $M$, that do not require any labelled data of $D_t$. However, such an approach might come at the price of accuracy: according to Jiang et al. (2020), the Spearman's rank correlation of any unsupervised measure with out-of-distribution accuracy does not exceed $0.5$ on average. The accuracy of the estimator improves using supervised approaches (Stefanik et al., 2021), but these already require some labelled data.

The situation presents a common dilemma in robustness evaluation: Ground-truth evaluation must involve a representative selection of test data. This problem can be avoided with unsupervised estimations based on the model properties, but such proxies are burdened by a certain level of inaccuracy. In the following sections, we review the measures introduced directly for evaluating model generalisation (§2.1.1) and for estimating model's expected output quality (§2.1.2), more commonly used in NLP.

#### 2.1.1 Generalisation Measures

Traditionally, the ability of neural networks to generalise was related to the measures of their *capacity*, where the lower capacity might imply the lower *generalisation gap* (Jiang et al., 2020), i.e. a drop of performance under *distribution shift*. The capacity can be quantified in terms of *complexity* given by a number of model parameters, expressive power or others. A standard example of such a measure is a degree of a polynomial; the higher the degree, the better is the fit, but it comes at the price of generalisation loss. This group of measures is referred to as Vapnik–Chervonenkis dimension (VC-dimension), introduced by Vapnik (1999).

A large body of work aims to find such VC-dimensions that correspond well with OOD performance even with modern, over-parametrised networks. For instance, norm-based approaches (Neyshabur et al., 2015b) propose to use the $p$-norms used in regularisation of the training as the anchor value of generalisation and support this in theory by connecting such measure with a limitation of network capacity. Bartlett et al. (2017) conclude that a *spectral complexity* measure, that is inferred from eigenvalues of a matrix of the network weights, can be used as one of such complexity measures.

A collateral line of work, starting with Shawe-Taylor et al. (1998) show that *generalisation bounds*, denoting a range of expected performance of the given model on an arbitrary test set, can be provably associated with *VC-bounds*. Harvey et al. (2017) show that the *tightness* of such bounds for a linear subset of networks can be theoretically found.

Furthermore, Dziugaite and Roy (2017) propose a method to *optimize PAC-Bayesian bounds*, optimising the model for as tight bounds as possible.

Despite these proofs, error bounds based on VC-dimensions remain *vacuous* in practice (Dziugaite and Roy, 2017; Jiang et al., 2020): such estimates of OOD performance are too wide to be used in practice. Additionally, it is now widely observed (Novak et al., 2018; Neyshabur et al., 2015a), that in practice, an effect of over-parametrisation is in contrast with traditional VC-dimension theory and in multiple cases, over-parametrisation leads to *better* reported generalisation (Neyshabur et al., 2019).

Existing work attempts to ground *error bounds* in the underlying causal model that *describes* the target domains of interest. Meinshausen (2018) introduces a term of *Structural equation model* (SEM) defining the causal interventions consistent with a given *world* and relates domain generalisation to the model's robustness to the *interventions* defined by such SEM. Additionally, given that SEM produces a class of distributions $\mathcal{Q}$, a model $M$ robust on $\mathcal{Q}$ is a *causal inference model* for $\mathcal{Q}$, connecting distributional robustness to a *weak form* of causal inference (Dziugaite et al., 2021). Similarly, Bühlmann (2018) ascribes the ability of causal inference on $\mathcal{Q}$ to any model whose representation is invariant to any domain $D \in \mathcal{Q}$ and proposes a method of selecting a subset of *invariant features* that picks such subset of attributes from a given set.

Practical observations of errors suggest that empirical *error bounds* are in fact significantly tighter than what can be proven in theory. Dziugaite et al. (2021) locate all bounds between the two extremes: theoretically-supported, yet vacuous bounds of methods based solely on the model property (*VC-bounds*) or behaviour (*PAC-Bayesian bounds*) and empirical, yet strictly data- and model-dependent evaluation on sample set(s) $X_t \in D_t$.

### 2.1.2 Quality Estimation

*Quality estimation* (QE) measure predicts model output quality in the absence of ground-truth reference (Fomicheva et al., 2020). Although not commonly used in this manner, QE measures also reflect on model robustness, making this branch of research applicable for OOD performance estimation (**RQ1**).

A significant line of work grounds quality estimation in model *confidence*, which can be estimated using Bayesian networks (Mackay, 1992) where standard *scalar* weights of the network are replaced with random variables, modelling the output distribution. This approach is accurate but not computationally feasible for larger networks. A branch of work *approximates* parametric distributions (Graves, 2011; Tran et al., 2019) making such uncertainty estimation practically feasible.

Model uncertainty can also be computed by ensembling variations of a given model in multiple trials, commonly referred to as *Monte Carlo* (MC) methods. Monte Carlo dropout (Gal and Ghahramani, 2016) applies dropout on inference randomly among multiple inference trials yielding an estimation of the distribution of network output, based on which the uncertainty is approximated. Lee et al. (2015) build such ensembles of estimators using *bagging*, i.e. training the ensembled models on different train sub-sets.

Model-variational methods fit well into the central *PAC-Bayesian* theory (Valiant, 1984), stating that if the error of the classifier can be bound, then also a performance of an ensemble of such classifiers can be upper-bound with arbitrarily-small bound $\epsilon$ (Guedj, 2019).

Confidence estimation can be utilised in enhanced model robustness, where prediction confidence is used as a regularizer of the main objective; in augmentation (Szegedy et al., 2014), confidence calibration (Gong et al., 2021), or in a training for consistency (Xie et al., 2019).

Jiang et al. (2020) propose to measure a regularisation decay of the weights, together with a measure of *sharpness*, reflecting on a volume of change in the model evaluation when the limited surrounding of the learnt parameter space minima is permuted (Keskar et al., 2017). Another introduced measure reflects a *variance of gradients* measured on a train set after a first training iteration. This work is the first large-scale study evaluating correlation of selected generalisation measures with true OOD performance and concludes that the mentioned sharpness and gradient-based measures correlate highest with the measured OOD performance. Consecutively, Dziugaite et al. (2021) support these findings on sharpness-based and PAC-Bayesian measures as the best-correlated in the similar methodology.

An important application of QE techniques lays in neural machine translation, where avoiding *critical errors* in translation remains an open problem. Such errors deviate the meaning of the translation

in a way that may carry health, safety, legal or other implications (Specia et al., 2021). Kim et al. (2017) train a token-level estimator of machine translation output quality concurrently with the neural translation model. Fomicheva et al. (2020) additionally propose to predict output quality from *entropy of attention activations* of transformer model, but they find this approach not more accurate than the one based on simple output entropy (Kim et al., 2017), or than the MC dropout method.

### 2.2 Training Robust Models (RQ2)

A problem of training a model that performs well on out-of-distribution (OOD) samples can be found in the literature under the terms of *out-of-distribution generalisation* (Yi et al., 2021), *domain generalisation* (Gong et al., 2021), *distributional robustness* (Meinshausen, 2018), or simply *generalisation* (Foret et al., 2021). The variety of terminology points to the fact that the standards in this branch of research are not yet clearly set.

Despite imperfect correlations of generalisation measures with measured OOD performance, we find these measures already incorporated in novel training objectives reaching attractive enhancements of model robustness; Neyshabur et al. (2015b) investigate the impact of incorporating norm-based measures into the loss, obtaining generalisation guarantees of $\ell_2$-norm. Foret et al. (2021) enrich the cross-entropy loss with a complementary component reflecting a sharpness of local optimum, based on a difference to local $\epsilon$. Keskar et al. (2017) also demonstrate that the sharpness of the objective's optima corresponds to the model's robustness, and flatter optima can also be reached by noising the update steps by smaller training batch size.

Objective adjustments creatively utilising PAC--Bayesian measures also confirm reported correspondence of these measures to generalisation. Hinton (2002) proposes a *Product of Experts* (PoE) framework where an ensemble of identical shallow estimators eliminate model-specific biases in a dot product of ensembled outputs, resulting in superior OOD performance. Sanh et al. (2021) show an application of PoE eliminating the systematic biases on adversarial NLI data sets. Dagaev et al. (2021) adopt similar approach in debiasing image classification from *heuristical shortcuts*. Utama et al. (2020) eliminate model reliance on domain-specific attributes in a two-step process: by *identifying* the biased samples by model over-confidence, and their subsequent *down-weighting*.

Rather than encouraging specific model features, others have investigated the impact of specific *training strategies*, which becomes particularly relevant in multi-step training strategies of LLMs. Wang and Sennrich (2020) enhance robustness of the translation by fine-tuning for sentence-level Minimum Risk Training objective instead of the common token-level cross-entropy. Tu et al. (2020) show on adversarial data sets that: a) longer fine-tuning eliminates model fragility on under-represented samples, and b) *multitask learning* has a positive impact on transformer generalisation to adversarial data sets. Compliant results are reported by Xie et al. (2019) with multitask learning for both classification and output consistency to augmented samples, or by Raffel et al. (2020) on generative language multitask learning, or in cross-lingual settings by Clark et al. (2019b); Conneau et al. (2019); Lewis et al. (2020).

Similar results are reported in work addressing dataset biases. Utama et al. (2020); Nie et al. (2019); Teney et al. (2020) report that addressing only one bias in domain adaptation hurts the model generalisation on other domains. On the other hand, Wu et al. (2020) find that addressing multiple biases at once can enhance OOD generalisation, although they draw this conclusion from a single domain.

A different branch of work attempts to enhance the robustness by training strategies that work with knowledge of *domain distinction*. Gong et al. (2021) propose to approximately cover the class of *all* possible *target* domains $D_t$ by *source* domains $D_s$ and to learn the calibration of output probabilities from $D_s$ that will allow to *associate* samples of a new target domain $D_t$ to some known $D_s$. Yi et al. (2021) propose to use the adversarial framework, learning *indistinguishable* final-layer representation for different domains.

## 3 Research Proposal

Following the referenced studies on evaluation and enhancement of the generalisation of neural models, this section outlines directions in measuring and improving robustness of LLMs, respectively.

### 3.1 Estimating Model Robustness (RQ1)

Recently, the measures of generalisation of neural networks struck increasing attention (Jiang et al., 2020; Dziugaite et al., 2021). However, none of the

referenced studies evaluates the measures on the case of LLMs. Especially within a standard *pre-training* + *fine-tuning* framework of modern NLP applications, quality of the measures might differ compared to the experiments on relatively small convolutional networks trained for image classification from scratch.

Hence, we first focus on evaluating the established generalisation measures, such as the ones based on spectral complexity, variance of gradients or sharpness in the case of pre-trained LLMs. A major challenge is to scale such experiments to a representative evaluation framework covering a broad set of tasks, domains, and model types. For instance, other training parameters will likely impact the metrics' quality; such covariates will have to be identified and controlled. However, even extensive evaluation will likely fail to identify some of such covariates; Due to this reason, we will delimit the scope of our results to the estimation and enhancement of robustness *with respect to* the enumerated covariates, even though it contrasts with the methodology of previous work.

We will give preference to the generalisation measures that correspond to linguistic and semantic language properties, as the practical deployment of such measures in evaluation also addresses a desire for enhancing *interpretability* of the LLMs' behaviour. Instances of linguistically-motivated measures can be a *largest common ancestor* between the parse trees of reference and hypothesis of generative model, or a coherence of output of discriminative model when a negation is introduced in the input.

In the evaluation of robustness of generative LLMs, we will prioritise *token-level* measures over conventional segment-level ones such as BLEU, as incorporating accurate token-level measures in training objectives could complement the classic token-level cross-entropy loss in sequence-to-sequence objective with its specific flaws, such as *exposure bias* (Wang and Sennrich, 2020).

The evaluation methodology will closely follow the one of Dziugaite et al. (2021), which reflects on a correlation of the measure with the measured OOD performance. If these measures reach high correlations, they might be applied directly in training regularisation or model selection. Even in cases of measures not reaching a high correlation, these can still bear the potential to improve model robustness (Foret et al., 2021).

### 3.2 Training Robust Models (RQ2)

Following the referenced examples adjusting training objectives with accurate generalisation measures (§2.2), e.g. norm-based measures (Neyshabur et al., 2015b), PAC-Bayesian measures (Sanh et al., 2021; Dagaev et al., 2021; Utama et al., 2020), or sharpness measure (Foret et al., 2021), we will use the accurate generalisation measures of LLMs (§3.1) as *regularizers* and *complementary objectives* of the training.

Locatello et al. (2019) theoretically prove that full distributional robustness is not possible without an *explicit* exposition of both the data and the model biases. Recently, Bengio et al. (2020) theoretically and empirically demonstrated that the model could *utilise* data biases to expose the underlying causal structure of the data in an experiment where such a structure is preliminarily known.

We will introduce training objectives that expose domain-specific data biases to the model in more explicit ways. The most direct approach is to complement the task-specific objective with another objective of distinguishing the domain(s) of origin. The domain-distinctive objective can shape a form of a binary classifier or a similarity loss of selected model representations (e.g. KL-divergence (Kullback and Leibler, 1951)).

We will investigate the impact of the *pre-training*, and *fine-tuning* objectives on the model's eventual robustness over multiple application tasks, domains and architectures, in a methodology similar to the *generalisation measures* evaluation of (Dziugaite et al., 2021).

Additionally, we will *replace* or *complement* the objectives of generative LLMs with token-level measures well-correlated with the OOD performance and compare the resulting models with computationally-expensive sentence-level objectives optimising the measures such as BLEU as their objectives.

In the case of discriminative models, we will evaluate robustness to surface-level heuristics using adversarial datasets like HANS (McCoy et al., 2019), or PAWS (Zhang et al., 2019) designed to expose the commonly-learnt biases of LLMs. For generative LLMs, we will evaluate a performance of the model on domain(s) *different* from the training domain; for instance, we will train a translation model on *subtitles* parallel corpus and evaluate on a domain of *news articles*. We will also evaluate the trained model(s) for its inclination to *critical*

*errors* as a probability of generating a translation containing a severe error (Specia et al., 2021) in *enforced* generation.

## 4 Conclusion

Our work outlines potential directions in enhancing distributional robustness of LLMs to mitigate a performance drop under distribution shift. We survey and identify accurate generalisation measures (§2.1) and find multiple studies demonstrating that utilisation of these measures in the training objectives positively impacts model robustness (§2.2).

Following this observation, we propose to identify generalisation measures best-suitable for LLMs (§3.1) and outline ways how to utilise these measures in the training process. Additionally, we identify a set of other methods reported to enhance OOD performance of LLMs that we propose to compare to in the outlined methodology for evaluating generalisation measures.

Similarly, we propose methodologies for robustness estimation of both generative and discriminative LLMs (§3.2); These methodologies are based on a quality assessment on the domains covered by the enclosed set of variables, and on the robustness towards the data set(s) constructed to expose enclosed set of models' biases.

## References


Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural Machine Translation by Jointly Learning to Align and Translate.

Peter L. Bartlett, Dylan J. Foster, and Matus Telgarsky. 2017. Spectrally-Normalized Margin Bounds for Neural Networks. In *Proc. of the 31st International Conference on Neural Information Processing Systems*, NIPS '17, pages 6241–6250, Red Hook, NY, USA. Curran Associates Inc.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and Natural Noise Both Break Neural Machine Translation. In *Proc. of International Conference on Learning Representations*.

Yoshua Bengio, Tristan Deleu, Nasim Rahaman, Nan Rosemary Ke, Sebastien Lachapelle, Olexa Bilaniuk, Anirudh Goyal, and Christopher Pal. 2020. A Meta-Transfer Objective for Learning to Disentangle Causal Mechanisms. In *Proc. of International Conference on Learning Representations*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching Word Vectors with Subword Information. ArXiv:1607.04606.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proc. of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. ACL.

Peter Bühlmann. 2018. Invariance, Causality and Robustness. *CoRR*, 1812.08233v1.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019a. What Does BERT Look At? An Analysis of BERT's Attention. ArXiv:1906.04341.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019b. What does BERT look at? an analysis of BERT's attention. In *Proc. of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. ACL.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised Cross-lingual Representation Learning at Scale. ArXiv:1911.02116.

Nikolay Dagaev, Brett D. Roads, Xiaoliang Luo, Daniel N. Barry, Kaustubh R. Patil, and Bradley C. Love. 2021. A Too-Good-to-be-True Prior to Reduce Shortcut Reliance. *CoRR*, abs/2102.06406v2.

Gintare K. Dziugaite and Daniel M. Roy. 2017. Computing Nonvacuous Generalization Bounds for Deep (Stochastic) Neural Networks with Many More Parameters than Training Data. *CoRR*, abs/1703.11008v1.

Gintare Karolina Dziugaite, Alexandre Drouin, Brady Neal, Nitarshan Rajkumar, Ethan Caballero, Linbo Wang, Ioannis Mitliagkas, and Daniel M. Roy. 2021. In Search of Robust Measures of Generalization. *CoRR*, abs/2010.11924v2.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. Unsupervised Quality Estimation for Neural Machine Translation. *Transactions of the ACL*, 8:539–555.

Pierre Foret, Ariel Kleiner, H. Mobahi, and Behnam Neyshabur. 2021. Sharpness-Aware Minimization for Efficiently Improving Generalization. *CoRR*, abs/2010.01412v1.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proc. of the 33rd International Conference on Machine Learning*, volume 48 of *Proc. of Machine Learning Research*, pages 1050–1059, New York, USA. PMLR.

Yunye Gong, Xiaoyu Lin, Yi Yao, Thomas G. Dietterich, Ajay Divakaran, and M. Gervasio. 2021. Confidence Calibration for Domain Generalization under Covariate Shift. *CoRR*, abs/2104.00742v2.

Alex Graves. 2011. Practical Variational Inference for Neural Networks. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.

Benjamin Guedj. 2019. A Primer on PAC-Bayesian Learning. *CoRR*, abs/1901.05353v3.

Nick Harvey, Christopher Liaw, and Abbas Mehrabian. 2017. Nearly-tight VC-dimension bounds for piecewise linear neural networks. In *Proc. of the Conference on Learning Theory*, volume 65 of *PMLR*, pages 1064–1068. PMLR.

Geoffrey E. Hinton. 2002. Training Products of Experts by Minimizing Contrastive Divergence. *Neural Computation*, 14(8):1771–1800.

Yiding Jiang, Behnam Neyshabur, H. Mobahi, Dilip Krishnan, and Samy Bengio. 2020. Fantastic Generalization Measures and Where to Find Them. *CoRR*, abs/1912.02178v1.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. 2017. On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. *CoRR*, abs/1609.04836v1.

Hyun Kim, Hun-Young Jung, Hongseok Kwon, Jong-Hyeok Lee, and Seung-Hoon Na. 2017. Predictor-Estimator: Neural Quality Estimation Based on Target Word Prediction for Machine Translation. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 17(1).

S. Kullback and R. A. Leibler. 1951. On Information and Sufficiency. *Annals of Mathematical Statistics*, 22:79–86.

Stefan Lee, Senthil Purushwalkam, Michael Cogswell, David J. Crandall, and Dhruv Batra. 2015. Why M Heads are Better than One: Training a Diverse Ensemble of Deep Networks. *CoRR*, abs/1511.06314v1.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proc. of the 58th Annual Meeting of the ACL*, pages 7871–7880. ACL.

Francesco Locatello, Stefan Bauer, Mario Lucic, S. Gelly, B. Schölkopf, and Olivier Bachem. 2019. Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations. *CoRR*, 1811.12359v4.

David John Cameron Mackay. 1992. *Bayesian Methods for Adaptive Models*. Ph.D. thesis, California Institute of Technology, USA. UMI Order No. GAX92-32200.

R. Thomas McCoy, Junghyun Min, and Tal Linzen. 2020. BERTs of a feather do not generalize together: Large variability in generalization across models with similar test set performance. *CoRR*, abs/1911.02969v2.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *Proc. of the 57th Annual Meeting of the ACL*, pages 3428–3448, Florence, Italy. ACL.

Nicolai Meinshausen. 2018. Causality from a Distributional Robustness Point of View. In *Proc. of IEEE Data Science Workshop (DSW 2018)*, pages 6–10.

Behnam Neyshabur, Zhiyuan Li, Srinadh Bhojanapalli, Yann LeCun, and Nathan Srebro. 2019. The role of over-parametrization in generalization of neural networks. In *Proc. of International Conference on Learning Representations*.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. 2015a. In Search of the Real Inductive Bias: On the Role of Implicit Regularization in Deep Learning. ArXiv:1412.6614.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. 2015b. Norm-based capacity control in neural networks. In *Proc. of The 28th Conference on Learning Theory*, volume 40 of *PMLR*, pages 1376–1401, Paris, France. PMLR.

Yixin Nie, Yicheng Wang, and Mohit Bansal. 2019. Analyzing Compositionality-Sensitivity of NLI Models. *CoRR*, abs/1811.07033v1.

Roman Novak, Yasaman Bahri, D. Abolafia, Jeffrey Pennington, and J. Sohl-Dickstein. 2018. Sensitivity and Generalization in Neural Networks: an Empirical Study. ArXiv:1802.08760.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(146):1–67.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proc. of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, USA. ACL.

Alan Ramponi and Barbara Plank. 2020. Neural unsupervised domain adaptation in NLP—A survey. In *Proc. of the 28th International Conference on Computational Linguistics*, pages 6838–6855, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Victor Sanh, Thomas Wolf, Yonatan Belinkov, and Alexander M. Rush. 2021. Learning from others' mistakes: Avoiding dataset biases without modeling them. *CoRR*, abs/2012.01300v1.

J. Shawe-Taylor, P.L. Bartlett, R.C. Williamson, and M. Anthony. 1998. Structural risk minimization over data-dependent hierarchies. *IEEE Transactions on Information Theory*, 44(5):1926–1940.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Chrysoula Zerva, Zhenhao Li, Vishrav Chaudhary, and André F. T. Martins. 2021. Findings of the WMT 2021 shared task on quality estimation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 684–725, Online. Association for Computational Linguistics.

Michal Stefanik, Vít Novotný, and Petr Sojka. 2021. Regressive ensemble for machine translation quality evaluation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1041–1048, Online. Association for Computational Linguistics.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, D. Erhan, I. Goodfellow, and R. Fergus. 2014. Intriguing properties of neural networks. *CoRR*, abs/1312.6199v4.

Damien Teney, Kushal Kafle, Robik Shrestha, Ehsan Abbasnejad, Christopher Kanan, and A. V. Hengel. 2020. On the Value of Out-of-Distribution Testing: An Example of Goodhart's Law. *CoRR*, abs/2005.09241v1.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT Rediscovers the Classical NLP Pipeline.

Dustin Tran, Michael W. Dusenberry, Mark van der Wilk, and Danijar Hafner. 2019. Bayesian Layers: A Module for Neural Network Uncertainty. In *Proc. of the 33rd International Conference on NIPS*, Red Hook, USA. Curran Associates Inc.

Lifu Tu, Garima Lalwani, Spandana Gella, and He He. 2020. An Empirical Study on Robustness to Spurious Correlations using Pre-trained Language Models. *Transactions of the ACL*, 8:621–633.

Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. 2020. Towards Debiasing NLU Models from Unknown Biases. In *Proc. of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7597–7610, Online. ACL.

L. G. Valiant. 1984. A Theory of the Learnable. In *Proc. of the Sixteenth Annual ACM Symposium on Theory of Computing*, STOC '84, pages 436–445, New York, USA. ACM.

Vladimir N. Vapnik. 1999. *The Nature of Statistical Learning Theory*, second edition. Springer.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *Proc. of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Chaojun Wang and Rico Sennrich. 2020. On exposure bias, hallucination and domain shift in neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3544–3552, Online. Association for Computational Linguistics.

Mingzhu Wu, N. Moosavi, Andreas Rücklé, and Iryna Gurevych. 2020. Improving QA Generalization by Concurrent Modeling of Multiple Biases. *CoRR*.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. 2019. Unsupervised Data Augmentation. *CoRR*, abs/1904.12848v1.

Mingyang Yi, Lu Hou, Jiacheng Sun, Lifeng Shang, Xin Jiang, Qun Liu, and Zhi-Ming Ma. 2021. Improved OOD Generalization via Adversarial Training and Pre-training. In *Proc. of the 38th ICML*, volume 139, pages 11987–11997. PMLR.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase Adversaries from Word Scrambling. In *Proc. of the 2019 Conf. NAACL-HLT*, pages 1298–1308, Minneapolis, USA. ACL.