Why External Validity Matters for Machine Learning Evaluation: Motivation and Open Problems

Thomas I. Liao*
me@thomasliao.com

Rohan Taori
Stanford University
rtaori@stanford.edu

Ludwig Schmidt
University of Washington
schmidt@cs.uw.edu

Abstract

New machine learning methods often fail to perform as expected on datasets similar to benchmarks reported on in their respective papers. These performance gaps pose a challenge for evaluation: both researchers and practitioners expect (or hope) that machine learning models which perform well on a dataset designed for a task perform well on other datasets matched to that task. We argue that external validity, the relationships between tasks and the learning problems which instantiate them, is understudied. We highlight the ways in which algorithm developers and benchmark creators fail to address this concern of external validity, suggest some remedies, and identify open questions in external validity which would help the community better build benchmarks and understand model performance.

1 Introduction

Evaluating machine learning models on benchmarks provides limited utility unless performance on the benchmark is informative about performance on a task. For example, researchers may not be interested in performance on ImageNet (Russakovsky et al., 2015) as such, but rather in image classification as a broadly-applicable task. Recent work reveals a string of examples where the bridge between benchmarks and tasks crumbles, from models exploiting inadvertent spurious heuristics in data (Oakden-Rayner et al., 2020; DeGrave et al., 2021; Goyal et al., 2017; McCoy et al., 2019; Kaushik et al., 2019), to examples where chosen metrics actually disagree with human judgment (Zhang & Toral, 2019; Edunov et al., 2019; Kryściński et al., 2019; Fabbri et al., 2020; Callison-Burch et al., 2006).

The notion of validity has a rich history in other fields; here, we use the concepts as adapted in (Liao et al., 2021). More specifically, *external validity* is the generalizability of methods and findings from one learning problem to another. *Learning problems* are the combination of a dataset and evaluation metric, e.g., ImageNet with top-1 accuracy forms a learning problem. Learning problems instantiate one or more *tasks*, or abstract problems, defined in either natural language or in a formal manner; e.g. image classification is a task. Evaluation questions regarding multiple benchmarks, such as extrapolating performance from one benchmark to another, are inherently different from *internal validity* concerns involving only a single benchmark, like characterizing model performance across random seeds, or comparing against competitive baselines. We illustrate an example task hierarchy and associated learning problems in Figure 1.

While the community has increased attention to the internal validity of research findings, e.g. pushing for reporting statistical uncertainty, external validity claims made by papers are often under-scrutinized, and the topic of external validity itself is understudied. We (i) provide some representative examples of external validity failures identified by recent papers; (ii) suggest methodological changes for paper authors to improve the rigor of external validity claims; (iii) highlight what we think are interesting and unanswered questions in the study of external validity.

*Corresponding author

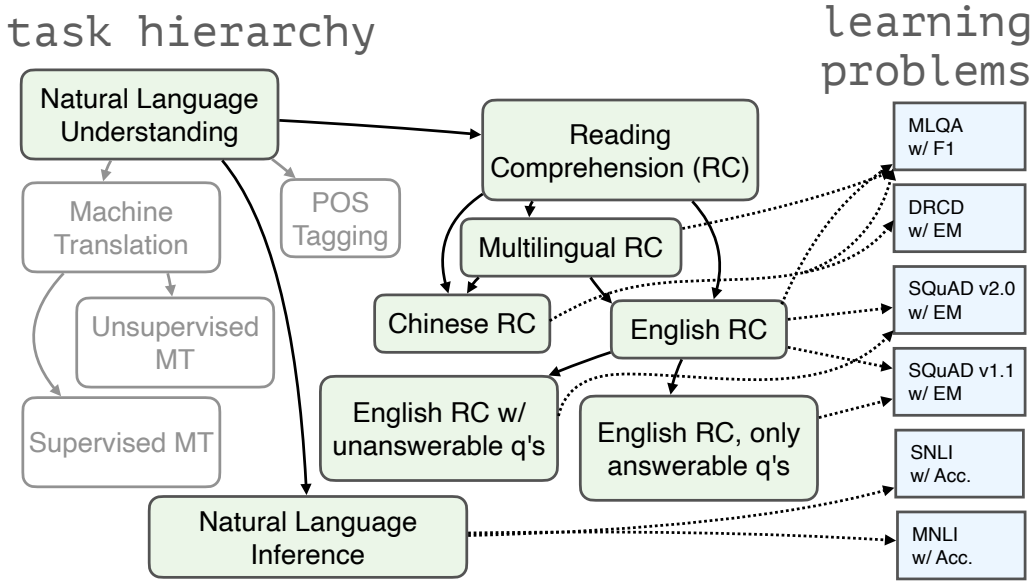

Figure 1: External validity is the generalizability of methods and findings from one learning problem to another. Learning problems combine a dataset and a metric to instantiate one or more tasks. While learning problems are concrete artefacts, tasks are abstract problems related to each other hierarchically. External validity failures like spurious correlations reveal the limitations of evaluating on a single learning problem. Datasets – MLQA (Lewis et al., 2019); DRCD (Shao et al., 2018); SQuAD v2.0 (Rajpurkar et al., 2018); SQuAD v1.1 (Rajpurkar et al., 2016); SNLI (Bowman et al., 2015); MNLI (Williams et al., 2017).

2 Motivating external validity

Recent works provide a long list of evaluation failures in machine learning papers.

Question answering models trained on the SQuAD dataset (Rajpurkar et al., 2016), sampled from Wikipedia, perform poorly when evaluated on data from other source domains, like Reddit posts and New York Times articles (Miller et al., 2020). Similarly, text summarization models originally trained on XSum (Narayan et al., 2018) or MLSum (Scialom et al., 2020) perform poorly when tested on newer data involving COVID-19 (Mille et al., 2021). One interpretation is that the original benchmarks overstate model performance - another is that the original benchmarks do not hold very strong external validity for other learning problems.

Reading comprehension models trained on certain bAbI (Weston et al., 2015) or CBT tasks (Hill et al., 2015) can predict the correct answer surprisingly well with only the text excerpt and ignoring the question - sometimes better than models trained with both (Kaushik & Lipton, 2018). Likewise, on the MNLI natural language inference dataset (Williams et al., 2017), heuristics comparing whether words in the hypothesis exist in the premise can achieve surprisingly strong performance (McCoy et al., 2019). In other words, supposedly critical components of these evaluation benchmarks can be ignored and still achieve good or even superior performance compared to state of the art models.

Both BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004), widely-used metrics for machine translation and automatic summarization, fail to correlate strongly with expert judgments of quality (Kryściński et al., 2019; Fabbri et al., 2020; Callison-Burch et al., 2006). Indeed, BLEU has contradicted human rankings in the past (Edunov et al., 2019; Zhang & Toral, 2019). The consequence is that improving ROUGE or BLEU scores on benchmarks may actually lead to reduced performance as determined by humans.

These cases demonstrate that even if models are rigorously evaluated on a single learning problem, performance cannot be readily extrapolated to the context of another learning

problem. If benchmarks are spotlights for model performance, these examples suggest that they illuminate a much narrower field of view than expected, and are often cast askew.

3 Suggestions for algorithm and benchmark creators

We provide some simple, nonexhaustive suggestions for researchers to address external validity concerns arising during benchmark creation and model evaluation.

Guidelines for benchmarks. Benchmark creators and the community writ large are responsible for rigour in data curation, annotation, and metric selection.

a. *Check for spurious correlations.* Benchmark creators should apply domain-specific knowledge to check for spurious correlations in their datasets (e.g., for text datasets, syntactic heuristics or negation operators (Niven & Kao, 2019; McCoy et al., 2019)).

b. *Compare against existing benchmarks.* Researchers should compare performance of state-of-the-art (SOTA) methods on their benchmark and existing benchmarks if possible. As an anti-example, three of four major scene text detection datasets failed to meet this bar: MSRA-TD500 (2012) (Yao et al., 2012) compared up to seven SOTA methods on two additional datasets, but COCO-Text (2016) (Veit et al., 2016), Total Text (2017) (Ch'ng & Chan, 2017) and SCUT-CTW1500 (2017) (Liu et al., 2017) only tested on their own datasets.

c. *Baseline simple methods.* In addition to baselining SOTA methods, benchmark creators should ensure that the learning problem is not unintentionally easy by baselining random, linear, or classical methods where applicable and checking assumptions about the task (e.g. verify that it is not possible to achieve a higher performance by ignoring parts of the input or dataset).

Guidelines for model evaluation. As much as possible, algorithm developers should strive to evaluate on multiple learning problems, i.e., multiple datasets and metrics. Ideally, the learning problems cover a diverse range of possible instantiations of the targeted task. (We comment further on data diversity in Section 4). Algorithm developers should be careful to scope claims about model performance based on the supplied evidence, avoiding over-general claims with lack of support. As authors have a clear self-interest in making strong statements about their contributions, incentive alignment must be provided by reviewers.

We invite further discussion on improving external validity. As more methodological checks are introduced (NeurIPS 2021 Paper Checklist, ICML 2020 Reproducibility Checklist, inter alia) managing the process burden on authors becomes a greater challenge.

4 Open questions in external validity

Providing recommendations on how to alleviate external validity failures is limited without further research into external validity. We outline two framing questions for external validity, algorithm transfer and learning problem transfer, then describe specific open problems. In general, the open questions surround either the relationships between different learning problems and tasks (**I, II, III, IV, VIII**) or about properties of the data and how they affect model performance (**V, VI, VII**).

Algorithm transfer poses a familiar question: is the relative performance of a model stable from one learning problem to another? A new method which exploits spurious correlations in a dataset will exhibit weak algorithm transfer, whereas methods like residual connections (He et al., 2016) consistently improve performance across many learning problems. In contrast to algorithm transfer, *learning problem transfer* is about the trajectory of progress of all models evaluated on a learning problem A versus the full performance history of those same models adapted to another learning problem B. Consider ImageNet (Russakovsky et al., 2015) - as models improved over time on that benchmark, they also improved over time on other learning problems (Kornblith et al., 2019; Recht et al., 2019).

I. Comparing learning problem transfer. Comparing model performance over time between learning problems would allow us to identify patterns which dictate when learning problem transfer occurs, helping researchers to build better benchmarks. Researchers converge to a smaller standard set of benchmarks over time, leading to a partial picture of model performance across all the learning problems introduced. How do extant learning problems transfer between each other? For example, (Kornblith et al., 2019) explicitly compare performance improvements on ImageNet to sixteen other image classification datasets. Comparing learning problem transfer between problems allows us to answer counterfactuals such as: "could model development have progressed faster with a different set of standard evaluation benchmarks ((Liu et al., 2021))?" and "is learning problem transfer from benchmark X to other benchmarks plateauing (i.e., the correlation of improvement on benchmark X with improvement on other benchmarks is decreasing over time)?"

II. Formalising similarity between datasets. Researchers often hold intuition about how well an algorithm will transfer between learning problems based on a vague notion of how similar or different two datasets are. Formalising how datasets are similar or different is worth further investigation. Statistical measures of distribution differences (Deng & Zheng, 2021; Borgwardt et al., 2006; Ben-David et al., 2006; Ganin et al., 2016; Glorot et al., 2011) are not the same as knowing that sim-to-real transfer is usually harder than switching geographic locales, or that news articles are more similar to social media posts than medical journals. A promising direction is to use trained models to describe differences in distributions (e.g. language models for text (Zhong et al., 2022). Such measures could be applied to systematically identify large, unlabelled datasets (inexpensive) to use for pre-training prior to labelled datasets (expensive). Another application is improving synthetic data generation as a substitute for data annotation.

III. Characterising relationships between learning problems and tasks. Learning problems exist independently of tasks. Researchers actively interpret performance on a learning problem as performance on relevant tasks: for example, results on SQuAD (Rajpurkar et al., 2016), a question-answering dataset, can be interpreted as measuring model capability in natural language understanding (as opposed to e.g. 3D point cloud classification), in reading comprehension (as opposed to machine translation), in English question-answering (as opposed to Mandarin question-answering), or Wikipedia span selection question-answering (as opposed to freeform Quora question-answering). From an evaluation perspective, there are usually at least one or two steps in the task hierarchy tree between a learning problem to a task of broader interest (e.g. SQuAD most directly measures English question-answering, which is derivative of language understanding). What makes one learning problem a better instantiation of a particular task than another learning problem?

IV. Formalising similarity between tasks. Architectures are increasingly shared between different tasks. The Transformer architecture (Vaswani et al., 2017) was initially benchmarked on machine translation, but has now been applied to speech recognition (Dong et al., 2018), image classification (Dosovitskiy et al., 2020), and even reinforcement learning (Chen et al., 2021). In other words, there is surprisingly strong learning problem transfer between machine translation and reinforcement learning, which seems to arise from some kind of underlying similarity between the two tasks; both can be posed as derivative subtasks of general sequence-to-sequence modelling. What are other such relationships between tasks? Is everything sequence-to-sequence modelling (i.e., we should build the best possible sequence-to-sequence model, then adapt it to derivative tasks)?

V. Comparing human- and computation-centric data difficulty metrics. Researchers have proposed two complementary strategies to quantify the difficulty of individual examples or classes of examples. One approach leverages information from human annotators to produce statistics like Krippendorff's α (Krippendorff, 1970), average human accuracy (Recht et al., 2019), human-model disagreement (Yang et al., 2019), or psychometric quantities such as from item response theory (Baker & Kim, 2004; Vania et al., 2021; Lalor et al., 2016). A second approach uses only information from model training or inference, such as the prediction depth (Baldock et al., 2021), consistency score (Jiang et al., 2020), model confidence and variability across training runs (Swayamdipta et al., 2020), and forgettability (Toneva et al., 2018). How do these two families of metrics relate to each other? Should we

treat harder versions of the same task as different tasks altogether? How do we determine the optimal difficulty for an evaluation dataset?

Further data scaling laws. Scaling laws (Kaplan et al., 2020) relate model performance with compute, dataset size, and parameter count. We suggest searching for similar phenomena with the qualities of data diversity and data quality.

VI. Data corpus diversity. Models are increasingly pre-trained on massive datasets blended from distinct corpora. For example, BERT, a language model (Devlin et al., 2019), was pre-trained on BooksCorpus (800M words) (Zhu et al., 2015) and English Wikipedia (2.5B words). What is the relationship between model performance and data diversity (and overall dataset size)? Is it possible to maintain model performance while decreasing total training dataset size by increasing the number of distinct data sources? This topic is closely related to formalising dataset similarity, above.

VII. Data quality. In an example of just how bespoke the treatment of data quality can be, GPT-3 (Brown et al., 2020), another language model, was pre-trained on five different datasets with sampling frequency hand-tuned based on subjective quality judgments (60% filtered Common Crawl (Raffel et al., 2019), 22% WebText2 (Radford et al., 2019), 8% Books1, 8% Books2, 3% Wikipedia). Similarly, pre-training datasets are frequently filtered or preprocessed to remove low quality data in an ad-hoc manner (Ng et al., 2019), which has been shown to improve training time and model outputs (Lee et al., 2021). Researchers have reported redundant examples or duplicates between the test and training set in datasets such as Common Crawl (Brown et al., 2020; Raffel et al., 2019) and CIFAR-100 (Barz & Denzler, 2020). What are more sophisticated quantifications of data quality, beyond the number of duplicate examples? What is the relationship between model performance and data quality? Does it ever become pointless to add extremely low quality data?

VIII. External validity and pretraining. Pretraining and finetuning add another wrinkle to each of the questions above. For example, how does the external validity of a pretraining dataset interact with the external validity of the finetuning dataset? How does diversity in the pretraining tasks affect algorithm transfer? When comparing algorithm transfer from one downstream learning problem to another, how do we account for shared pretraining datasets? It remains unclear how to choose the optimal pretraining task or learning problem for a given downstream task.

5 Conclusion

We argue for the importance of external validity to machine learning evaluation. Even if individual benchmarks are well-designed, absent external validity, conclusions are not guaranteed to apply to other learning problems or the broader tasks that the community are ultimately interested in. As external validity is still not well understood, we hope that the open problems in this paper will drive further progress in machine learning.

References

Baker, F. B. and Kim, S.-H. *Item response theory: Parameter estimation techniques*. CRC press, 2004.

Baldock, R. J. N., Maennel, H., and Neyshabur, B. Deep learning through the lens of example difficulty. *CoRR*, abs/2106.09647, 2021. URL `https://arxiv.org/abs/2106.09647`.

Barz, B. and Denzler, J. Do we train on test data? purging cifar of near-duplicates. *Journal of Imaging*, 6(6):41, 2020.

Ben-David, S., Blitzer, J., Crammer, K., and Pereira, F. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19, 2006.

Borgwardt, K. M., Gretton, A., Rasch, M. J., Kriegel, H.-P., Schölkopf, B., and Smola, A. J. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, 2006.

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

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020. URL `https://arxiv.org/abs/2005.14165`.

Callison-Burch, C., Osborne, M., and Koehn, P. Re-evaluating the role of bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, 2006.

Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., Abbeel, P., Srinivas, A., and Mordatch, I. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34, 2021.

Ch'ng, C. K. and Chan, C. S. Total-text: A comprehensive dataset for scene text detection and recognition. In *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*, volume 1, pp. 935–942. IEEE, 2017.

DeGrave, A. J., Janizek, J. D., and Lee, S.-I. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, May 2021. doi: 10.1038/s42256-021-00338-7. URL `https://doi.org/10.1038/s42256-021-00338-7`.

Deng, W. and Zheng, L. Are labels always necessary for classifier accuracy evaluation? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15069–15078, 2021.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

Dong, L., Xu, S., and Xu, B. Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5884–5888. IEEE, 2018.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. URL `https://arxiv.org/abs/2010.11929`.

Edunov, S., Ott, M., Ranzato, M., and Auli, M. On the evaluation of machine translation systems trained with back-translation. *arXiv preprint arXiv:1908.05204*, 2019.

Fabbri, A. R., Kryscinski, W., McCann, B., Xiong, C., Socher, R., and Radev, D. R. Summeval: Re-evaluating summarization evaluation. *CoRR*, abs/2007.12626, 2020. URL `https://arxiv.org/abs/2007.12626`.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

Glorot, X., Bordes, A., and Bengio, Y. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *ICML*, 2011.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. 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.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. `https://arxiv.org/abs/1512.03385`.

Hill, F., Bordes, A., Chopra, S., and Weston, J. The goldilocks principle: Reading children's books with explicit memory representations. *arXiv preprint arXiv:1511.02301*, 2015.

Jiang, Z., Zhang, C., Talwar, K., and Mozer, M. C. Exploring the memorization-generalization continuum in deep learning. *CoRR*, abs/2002.03206, 2020. URL `https://arxiv.org/abs/2002.03206`.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *CoRR*, abs/2001.08361, 2020. URL `https://arxiv.org/abs/2001.08361`.

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

Kaushik, D., Hovy, E., and Lipton, Z. C. Learning the difference that makes a difference with counterfactually-augmented data. *arXiv preprint arXiv:1909.12434*, 2019.

Kornblith, S., Shlens, J., and Le, Q. V. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2661–2671, 2019.

Krippendorff, K. Bivariate agreement coefficients for reliability of data. *Sociological methodology*, 2:139–150, 1970.

Kryściński, W., Keskar, N. S., McCann, B., Xiong, C., and Socher, R. Neural text summarization: A critical evaluation. *arXiv preprint arXiv:1908.08960*, 2019.

Lalor, J. P., Wu, H., and Yu, H. Building an evaluation scale using item response theory. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2016, pp. 648. NIH Public Access, 2016.

Lee, K., Ippolito, D., Nystrom, A., Zhang, C., Eck, D., Callison-Burch, C., and Carlini, N. Deduplicating training data makes language models better. *arXiv preprint arXiv:2107.06499*, 2021.

Lewis, P., Oğuz, B., Rinott, R., Riedel, S., and Schwenk, H. Mlqa: Evaluating cross-lingual extractive question answering. *arXiv preprint arXiv:1910.07475*, 2019.

Liao, T. I., Taori, R., Raji, I. D., and Schmidt, L. Are We Learning Yet? A Meta Review of Evaluation Failures Across Machine Learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

Lin, C.-Y. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL `https://aclanthology.org/W04-1013`.

Liu, N. F., Lee, T., Jia, R., and Liang, P. Can small and synthetic benchmarks drive modeling innovation? a retrospective study of question answering modeling approaches. *arXiv preprint arXiv:2102.01065*, 2021.

Liu, Y., Jin, L., Zhang, S., and Zhang, S. Detecting curve text in the wild: New dataset and new solution. *CoRR*, abs/1712.02170, 2017. URL `http://arxiv.org/abs/1712.02170`.

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

Mille, S., Dhole, K. D., Mahamood, S., Perez-Beltrachini, L., Gangal, V., Kale, M., van Miltenburg, E., and Gehrmann, S. Automatic construction of evaluation suites for natural language generation datasets. *CoRR*, abs/2106.09069, 2021. URL `https://arxiv.org/abs/2106.09069`.

Miller, J., Krauth, K., Recht, B., and Schmidt, L. The effect of natural distribution shift on question answering models. In *International Conference on Machine Learning*, pp. 6905–6916. PMLR, 2020.

Narayan, S., Cohen, S. B., and Lapata, M. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *CoRR*, abs/1808.08745, 2018. URL http://arxiv.org/abs/1808.08745.

Ng, N., Yee, K., Baevski, A., Ott, M., Auli, M., and Edunov, S. Facebook fair's wmt19 news translation task submission. *arXiv preprint arXiv:1907.06616*, 2019.

Niven, T. and Kao, H.-Y. Probing neural network comprehension of natural language arguments. *arXiv preprint arXiv:1907.07355*, 2019.

Oakden-Rayner, L., Dunnmon, J., Carneiro, G., and Ré, C. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM conference on health, inference, and learning*, pp. 151–159, 2020.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://www.aclweb.org/anthology/P02-1040.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. Squad: 100, 000+ questions for machine comprehension of text. *CoRR*, abs/1606.05250, 2016. URL http://arxiv.org/abs/1606.05250.

Rajpurkar, P., Jia, R., and Liang, P. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*, 2018.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pp. 5389–5400. PMLR, 2019.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, December 2015. ISSN 0920-5691, 1573-1405. doi: 10.1007/s11263-015-0816-y.

Scialom, T., Dray, P., Lamprier, S., Piwowarski, B., and Staiano, J. MLSUM: the multilingual summarization corpus. *CoRR*, abs/2004.14900, 2020. URL https://arxiv.org/abs/2004.14900.

Shao, C. C., Liu, T., Lai, Y., Tseng, Y., and Tsai, S. Drcd: a chinese machine reading comprehension dataset. *arXiv preprint arXiv:1806.00920*, 2018.

Swayamdipta, S., Schwartz, R., Lourie, N., Wang, Y., Hajishirzi, H., Smith, N. A., and Choi, Y. Dataset cartography: Mapping and diagnosing datasets with training dynamics. *CoRR*, abs/2009.10795, 2020. URL https://arxiv.org/abs/2009.10795.

Toneva, M., Sordoni, A., Combes, R. T. d., Trischler, A., Bengio, Y., and Gordon, G. J. An empirical study of example forgetting during deep neural network learning. *arXiv preprint arXiv:1812.05159*, 2018.

Vania, C., Htut, P. M., Huang, W., Mungra, D., Pang, R. Y., Phang, J., Liu, H., Cho, K., and Bowman, S. R. Comparing test sets with item response theory. *arXiv preprint arXiv:2106.00840*, 2021.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need, 2017.

Veit, A., Matera, T., Neumann, L., Matas, J., and Belongie, S. Coco-text: Dataset and benchmark for text detection and recognition in natural images. *arXiv preprint arXiv:1601.07140*, 2016.

Weston, J., Bordes, A., Chopra, S., Rush, A. M., Van Merriënboer, B., Joulin, A., and Mikolov, T. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*, 2015.

Williams, A., Nangia, N., and Bowman, S. R. A broad-coverage challenge corpus for sentence understanding through inference. *CoRR*, abs/1704.05426, 2017. URL `http://arxiv.org/abs/1704.05426`.

Yang, Y., Agarwal, O., Tar, C., Wallace, B. C., and Nenkova, A. Predicting annotation difficulty to improve task routing and model performance for biomedical information extraction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 1471–1480, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1150. URL `https://aclanthology.org/N19-1150`.

Yao, C., Bai, X., Liu, W., Ma, Y., and Tu, Z. Detecting texts of arbitrary orientations in natural images. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 1083–1090. IEEE, 2012.

Zhang, M. and Toral, A. The effect of translationese in machine translation test sets. *arXiv preprint arXiv:1906.08069*, 2019.

Zhong, R., Snell, C., Klein, D., and Steinhardt, J. Summarizing differences between text distributions with natural language. *arXiv preprint arXiv:2201.12323*, 2022.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., and Fidler, S. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pp. 19–27, 2015.