Efficient Fine-Tuning of BERT with ELECTRA

Stanford CS224N Default Project

Akshay Gupta
Department of Computer Science
Stanford University
agupta23@stanford.edu

Vincent Huang
Department of Computer Science
Stanford University
hvincent@stanford.edu

Erik Rozi
Department of Computer Science
Stanford University
erikrozi@stanford.edu

Abstract

In our study, we investigated how various pretraining and finetuning processes improved a BERT model, focusing particularly on a minBERT variant. BERT is typically pretrained with a masked language modeling (MLM) approach. We introduced a novel pretraining step and novel training objective to enhance the pretraining process, as described by He et al. (2023), with its disentangled attention mechanism and ELECTRA-style pre-training.

Our novel training step incorporates ELECTRA-Style Pre-Training along with Gradient-Disentangled Embedding Sharing techniques from the DeBERTaV3 framework, which is a unique attempt to improve BERT model's learning process with these specific methods. Our primary goal is to increase the efficiency of the model's learning when training on smaller, domain-specific datasets, since traditional MLM pretraining is often inefficient for maximizing model performance on these smaller, specific datasets.

We also conducted experiments employing various fine-tuning strategies to further assess their impact on the model's performance, including gradient surgery for multi-task learning as in Yu et al. (2020), round-robin fine tuning, cosine embedding adjustments, and SMART Regularization. We evaluated our model on the NLU tasks of Sentiment Analysis, Paraphrase Detection, and Semantic Textual Similarity.

Overall, our approach yielded great performance benefits across Paraphrase Detection, Semantic Analysis, and Semantic Textual Similarity tasks on domain-specific datasets. Specifically, MLM pretraining, ELECTRA-style pretraining, and Gradient-Disentangled Embedding Sharing resulted in slight improvements in the final model. Concatenating sentences with a [SEP] token for similarity tasks, gradient surgery techniques, cosine annealing, Adam weight regularization, and round-robin training significantly enhanced model performance. Our results suggest that these approaches are promising for improving BERT model performance on smaller, domain-specific datasets and tasks.

1 Key Information to include

Our mentor is David Lim. We have no external collaborators nor are we sharing projects. All group members worked equally on coding through pair-coding for various parts of the project, and worked equally on the paper and poster. We are using one late day on this report. Erik Rozi is supplying 3 late days for the group to use.

2 Introduction

Pretrained language encoder models, especially the BERT architecture, have revolutionized the approach to a wide range of NLP tasks, but traditional pretraining techniques like masked language modeling (MLM) often face challenges when applied to smaller, domain-specific datasets. Specifically, trying to use a traditional, vanilla fine-tuning process for NLU tasks on specific datasets such as Semantic Textual Similarity, Sentiment Analysis, and Paraphrase Detection yields results with much room for improvement. *(See the minBERT w/ vanilla finetuning performance on Table 1: Performance comparison of models*

Our study addresses these weaknesses by introducing a new step to the BERT pretraining process and a more robust approach to fine-tuning that significantly improves model performance on multiple Natural Language Understanding (NLU) tasks. To enhance the model's learning efficiency for specialized datasets and to improve performance on NLU tasks, we leverage the ELECTRA-Style Pre-Training and Gradient-Disentangled Embedding Sharing techniques from the DeBERTaV3 framework, as in the work of He et al. (2023). To optimize the learning rate of the fine-tuning process, we also experimented with changing the approach to MLM in pre-training with equal batch sizes across diverse datasets, and leveraged cosine annealing with Adam weight regularization to improve multi-task performance.

To refine the fine-tuning process and build on vanilla fine-tuning, we used round-robin fine-tuning (RRFT). Our goal with RRFT was to enhance model robustness and generalization across the 3 tasks of Paraphrase Detection, Semantic Analysis, and Semantic Textual Similarity (STS). Another step we took to improve performance was the inclusion of cosine embeddings (CE) in our loss function, enabling the model to consider geometric relationships between embeddings for correlation measures during fine-tuning for improved performance on tasks related to text similarity. We also implemented gradient surgery for multi-task learning as in Yu et al. (2020) to fine-tune the model with more stability and precision, and to avoid issues related to vanishing or exploding gradients to improve convergence.

Finally, we honed our approach by incorporating advanced techniques such as the concatenation of sentences with separator tokens to enrich the model's contextual understanding, and the use of mean squared error on cosine similarity as a loss function to refine semantic understanding.

Overall, our approach yielded great performance benefits across Paraphrase Detection, Semantic Analysis, and Semantic Textual Similarity tasks on domain-specific datasets *(see our final model in Table 1)*.

We found that MLM pretraining and ELECTRA-style pretraining, combined with Gradient-Disentangled Embedding Sharing resulted in very slight improvements in the final model, with minimal improvement observed in paraphrase detection tasks, suggesting that potential benefits may require a larger volume of training examples to materialize. We also found that employing cosine similarity with separate embeddings proved substantially less effective than a simpler approach of concatenating sentences with a [SEP] token. Notably, gradient surgery techniques, alongside cosine annealing and Adam weight regularization, demonstrated efficacy in reducing overfitting, thereby enhancing model performance. The round-robin training strategy emerged as advantageous over single-dataset training, with an extension to scaled batch sizes offering further improvements at the cost of extended training durations.

3 Related Work

Our work on refining pre-trained language models, builds on the work done for DeBERTaV3, He et al. (2023), which proposes using a generator and discriminator similar to Clark et al. (2020). The generator is trained with MLM and is used to generate ambiguous tokens to replace masked tokens in the input sequence. This input sequence is then fed to the discriminator, which classifies if a corresponding token is either an original token or a token replaced by the generator. This specific objection is referred to as Replaced Token Detection. With its innovative use of replaced token detection (RTD) and gradient-disentangled embedding sharing (GDES), He et al. (2023) enhances the pre-training efficiency and effectiveness of the DeBERTa model by integrating the strengths of ELECTRA-style pre-training (Clark et al., 2020) and introducing a novel gradient-disentangled embedding sharing technique. This method addresses the inherent inefficiencies and conflicts present

in traditional embedding sharing mechanisms used in language models, which often result in a "tug-of-war" between the training objectives of the discriminator and generator. By solving these problems, DeBERTaV3 aims to set new benchmarks in natural language understanding tasks, demonstrating significant improvements over its predecessors. The motivation behind this work is to push the boundaries of what pre-trained language models can achieve by optimizing their training process and enhancing their ability to understand and process natural language, thereby opening new avenues for research and application in the field of natural language processing.

By building on the revolutionary groundwork laid by BERT and ELECTRA, DeBERTaV3 combines the concept of masked language modeling (MLM) with the efficiency of RTD. Where BERT introduced the transformative power of MLM, enabling models to grasp context and meaning by predicting masked tokens, ELECTRA refined this approach with the RTD task. This methodology has a generator model substitute tokens in the input text, which are then identified by a discriminator model, thus turning every token into a potential learning opportunity and pushing the model to achieve a deeper comprehension of language.

The adoption of RTD in DeBERTaV3, as detailed by He et al. (2023), enhanced model sample efficiency and its ability to understand the nuanced semantics of language. This method has proven to be exceptionally effective across various NLU benchmarks, showcasing DeBERTaV3's adaptability and power.

Equally important is DeBERTaV3's introduction of GDES, a solution to the longstanding inefficiencies and conflicts arising from traditional embedding sharing practices. By allowing the generator and discriminator to learn independently, GDES eliminates the counterproductive "tug-of-war" effect, enabling more streamlined training and the production of superior token embeddings. This breakthrough addresses one of the key challenges in embedding sharing, providing valuable insights for future pre-training strategies.

The evaluation of DeBERTaV3, demonstrated through its stellar performance on benchmarks like GLUE and XNLI, attests to its superiority over previous models. Such achievements not only highlight DeBERTaV3's innovative features but also underscore its role in setting new standards for NLU tasks.

Our approach to improving fine-tuning with gradient surgery for multi-task learning largely drew from the work of Yu et al. (2020), which addresses the problem of conflicting gradients in multi-task learning by implementing techniques such as Gradient Sign Dropout (GSD) and Projected Gradient Descent (PGD). These methods strategically manipulate gradients to diminish conflicts, thereby enhancing the model's ability to learn cohesively from multiple tasks and preventing improvements in one task from detrimentally affecting another, ultimately leading to a more efficient and balanced learning process across tasks.

Overall, we aimed to leverage elements from the training of DeBERTaV3, which improved efficiency and effectiveness through RTD and GDES,to improve the default BERT model. We were also partially motivated by the success of Sun et al. (2019), which outlines how additional pretraining with target domain data can help improve model performance.

4 Approach

Our project began with the BERT model as a starting point, from which we sought to significantly enhance performance through strategic modifications. Our first major enhancement was the adoption of ELECTRA-style pre-training, a choice motivated by the potential for more nuanced language comprehension. This method contrasts with BERT's masked word prediction by teaching the model to differentiate between correct and intentionally altered words, a technique we believed would fine-tune its sensitivity to language's subtler aspects.

Building on this foundation, we introduced a round-robin training strategy, cycling through different datasets—specifically, SST for sentiment analysis, Quora, and SemEval for paraphrase detection and semantic similarity—in each training iteration. This method aimed to holistically optimize the model across various tasks by blending the losses from each dataset, thereby bolstering its generalization across these distinct tasks. To further refine our model's understanding of text, we experimented with cosine embedding loss, particularly for enhancing performance on semantic textual similarity and paraphrase detection tasks. Following the implementation of round-robin training, we

further augmented our model's learning capabilities by incorporating SMART (Sharpness-Aware Minimization for Efficiently Training Neural Networks) Regularization. This technique is designed to improve generalization by directly minimizing the sharpness of the loss landscape, encouraging the model to converge to flatter minima. By integrating SMART Regularization, we aimed to enhance the model's robustness and stability, particularly in handling the nuanced expressions of sentiment and complex paraphrase scenarios identified as challenging.

With these enhancements in place, we proceeded to fine-tune our evolved model. This phase was critical for validating the improvements, allowing us to measure the model's enhanced capabilities across our targeted tasks and offering a direct comparison to its baseline performance. Notably, we also incorporated gradient surgery techniques to optimize the fine-tuning process further. This approach was expected to mitigate training conflicts and promote more efficient learning.

Our approach, beginning with foundational BERT pre-training modifications and progressing through iterative fine-tuning experiments, was designed to comprehensively assess the impact of each strategic enhancement on the model's performance. By integrating ELECTRA-style learning, round-robin training, cosine embedding adjustments, and gradient surgery Yu et al. (2020), we aimed to create a more adept model for domain-specific datasets. Combined, these improvements allowed us to train a model that performed much better than baseline for sentiment analysis based on the Stanford Sentiment Treebank (SST), paraphrase detection, and Semantic Textual Similarity (STS) performance metrics. Our final model leveraged the foundational strengths of BERT and DeBERTaV3, pushing the boundaries of what our minBERT model could achieve in these three core Natural Language Understanding (NLU) tasks.

5 Experiments

5.1 Data

For our improved model, we used the Stanford Sentiment Treebank (SST) dataset,[1] which is comprised of 1,855 single sentences from movie reviews and 215,154 unique phrases. Additionally, we used the Quora dataset, which consists of 400,000 question pairs with labels for paraphrase detection. [2] Finally, we used the SemEval STS Benchmark Dataset, which is made up of 8,628 sentence pairs scaled from 0-5, for STS.[3] Our round robin training method trains the model on each specific dataset to enhance its performance on the corresponding tasks of sentiment analysis, paraphrase detection, and STS, without intermixing the datasets across these distinct tasks.

5.2 Evaluation method

We evaluated our improved model on sentiment analysis (SST), paraphrase detection and semantic text similarity (STS). For paraphrase detection, we use accuracy as our evaluation metric. For semantic text similarity, we use the Pearson correlation of the true similarity values against the predicted similarity values as our evaluation metric. For sentiment analysis, we use accuracy as our evaluation metric. To compare the various approaches we used to train our model, we prioritized approaches that maximized the combined score of all three of these NLU tasks, since we cared about the model's overall performance on multi-task NLU. Additionally, we ensured that none of the individual scores were too low, since we wanted to train a model that was capable of robust performance on all three tasks.

5.3 Experimental details

In our research, we focused on enhancing the learning process of a BERT-derived model, herein referred to as minBERT, specifically tailored for small, domain-specific datasets. Our methodology involved a series of incremental experiments designed to assess the impact of various fine-tuning strategies on the model's performance. The following outlines our systematic approach to refining minBERT's training regimen, presented with a balance between formal discourse and detailed explanation.

[1] `https://nlp.stanford.edu/software/lex-parser.shtml`

[2] `https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs.`

[3] https://aclanthology.org/S13-1004/

The initial phase of our experimentation served as a benchmark, comparing the performance of minBERT following standard fine-tuning practices against a baseline model initialized with random weights. This comparison established a foundational understanding of the benefits derived from conventional fine-tuning techniques.

Subsequently, we introduced round-robin fine-tuning (RRFT) to our training methodology. RRFT deviates from traditional batch processing by cyclically presenting data from multiple datasets to the model. This approach is hypothesized to foster a more robust generalization capability by exposing the model to a broader spectrum of linguistic patterns and contexts.

Building upon the RRFT framework, we integrated cosine embeddings (CE) into the fine-tuning process. Cosine embeddings evaluate the semantic similarity between text samples, offering a nuanced metric for the model to optimize. This addition aimed to enhance minBERT's ability to discern and represent the semantic proximity of textual data.

Further modifications to our fine-tuning strategy involved the application of a binary threshold for correlation based on cosine similarity scores. Specifically, we trained the model to recognize and categorize text pairs exceeding a predefined similarity threshold. This experiment sought to refine the model's precision in identifying semantically related texts. Additionally, we explored the utilization of mean squared error (MSE) on cosine similarity scores as an alternative loss function, directing the model's learning focus towards minimizing discrepancies in semantic similarity assessments.

An innovative approach we tested was concatenating pairs of sentences, separated by a [SEP] token, and evaluating whether this contiguous presentation of textual data could improve the model's understanding of inter-sentential relationships.

To address potential inefficiencies in the learning process, we employed gradient surgery techniques in conjunction with RRFT. Gradient surgery involves the strategic manipulation of gradient flow during backpropagation to prevent detrimental interference and promote more effective learning. This technique was anticipated to enhance model convergence and performance.

As an extension of our pre-training enhancements, we subjected minBERT to masked language modeling (MLM) for 10 epochs across all datasets, maintaining uniform batch sizes. This pre-training phase aimed to instill a deep foundational comprehension of language in the model, preparing it for the subsequent fine-tuning stages.

Our culminating experiments combined gradient surgery with two additional optimization strategies: cosine annealing and Adam weight regularization. Cosine annealing adjusts the learning rate following a cosine curve, potentially improving the model's adaptability over training epochs. Adam weight regularization introduces a penalty term to the optimization process, mitigating the risk of overfitting by discouraging excessive weight magnitudes.

Each experimental iteration was meticulously designed to isolate the effects of the introduced variables, thereby enabling a clear evaluation of their individual and collective contributions to enhancing minBERT's training efficacy. This comprehensive approach not only facilitated a deeper understanding of effective training strategies for domain-specific datasets but also provided valuable insights into the optimization of BERT-derived models for specialized applications.

5.4 Results

Model (GS = Gradient Surgery, CA = Cosine Annealing)	Paraphrase Detection (Accuracy)	Semantic Text Similarity (Pearson Correlation)	Sentiment Analysis (Accuracy)
Random	0.50	0	0.20
minBERT w/ vanilla finetuning	0.527	0.490	0.030
minBERT w/ round-robin fine tuning	0.523	0.721	0.368
Round robin (two separate embeddings, simple MSE loss)	0.721	0.368	0.523
minBERT w/ RRFT + cosine embeddings	0.523	0.719	0.392
Round robin (PyTorch cosine similarity loss, binary labels)	0.720	0.464	0.516
Round robin (cosine similarity, MSE loss)	0.729	0.787	0.487
Gradient surgery + SMART (lambda=5)	0.817	0.877	0.502
Round robin ([SEP] token concatenation)	0.836	0.873	0.498
Gradient surgery (equal batch sizes, MLM)	0.838	0.868	0.513
Gradient surgery (equal batch sizes)	0.837	0.864	0.520
Gradient surgery (equal batch sizes, MLM)	0.843	0.872	0.511
Gradient surgery (equal batch sizes, RTD + GDES)	0.849	0.867	0.510
Gradient surgery + SMART (lambda=2)	0.836	0.877	0.529
GS + CA + Adam (1e-2, batch 32 for paraphrase 2 for others)	0.874	0.877	0.516
GS + CA + Adam (1e-4)	0.875	0.870	0.527
GS + CA + Adam (1e-2, batch 32 for paraphrase 1/2 for others)	0.874	0.874	0.524
Final Submitted Model (Test Scores)	0.876	0.869	0.526

Table 1: Performance comparison of models, including final submitted model test scores

We explored various pretraining techniques and their impact on our final model. We found that Masked Language Modeling (MLM) pretraining offered a very slight improvement. We decided to incorporate it into the final model, although we suspect this enhancement might be attributable to random variation rather than a concrete benefit from the technique itself.

Our experiments with ELECTRA-style pretraining, specifically the Replacement Token Detection (RTD) with Gradient-Disentangled Embedding Sharing, did not yield significant improvements. This outcome aligns with our observations from MLM pretraining. However, we saw a marginal benefit in paraphrase detection tasks, suggesting that such techniques may offer some advantage given more training examples for certain NLU tasks.

Interestingly, our investigations into embedding strategies revealed that using separate embeddings for computing cosine similarity was markedly less effective than simply concatenating two sentences with a [SEP] token. This suggests the importance of analyzing texts in a shared context to enhance the model's comprehension of their relational dynamics. By concatenating the sentences with a [SEP] token, the approach leverages the model's ability to evaluate and encode the semantic interplay and connections between multiple texts in a way that is more nuanced than a simple cosine similarity score.

On the optimization front, gradient surgery emerged as a beneficial technique. Furthermore, employing cosine annealing alongside Adam weight regularization appeared to significantly enhance performance. We attribute these improvements to a reduction in overfitting, highlighting the importance of thoughtful regularization strategies.

Lastly, our experiments with dataset training strategies revealed the effectiveness of round-robin training. This approach, especially when combined with scaled batch sizes, outperformed training exclusively on a single dataset. Although the increased training time with scaled batch sizes presents a practical limitation, the benefits suggest that it could be a worthwhile investment for certain applications.

6 Analysis

Our model demonstrates remarkable adaptability to various linguistic tasks, as evident from its greatly improved scores on SST, STS, and paraphrase detection tasks we tested it against. The model's robustness, enhanced through round-robin fine-tuning, is evident in its ability to handle a variety of linguistic patterns in examples we tested across tasks, and through its vastly superior performance on SST, STS, and Paraphrase Detection compared to our baseline models. However, this robustness is tested when confronted with high variability in language use, such as idiomatic expressions, highly technical language, or domain-specific jargon. The model's performance was particularly bad when interpreting idioms or culturally specific expressions, suggesting that our methodologies may not fully capture the breadth of linguistic variations.

While the model shows proficiency in leveraging context for understanding, it faces challenges with texts requiring deeper semantic interpretation or reliance on broader contextual knowledge beyond the immediate text. This limitation was particularly evident in handling nuanced expressions of sentiment or complex paraphrase scenarios where contextual cues are subtle or indirect.

Consider our model's performance on the examples in the below table from the Stanford Sentiment Treebank, where the predicted sentiment varied widely from the actual sentiment.

Sentence	**Actual Sentiment**	**Predicted Sentiment**
Still , as a visual treat , the film is almost worth the price of admission .	4	1
It 's everything you do n't go to the movies for .	0	3
Hilariously inept and ridiculous .	3	0
This flick is about as cool and crowd-pleasing as a documentary can be .	4	1
It takes a certain kind of horror movie to quote the Talmud , and this ain't it .	0	3
We have n't seen such hilarity since Say It Is n't So .	4	1
No screen fantasy-adventure in recent memory has the show-stopping sights that this one does .	4	1
This riveting World War II moral suspense story could not be more timely .	1	3
Old-form moviemaking at its best .	4	2
No sophomore slump for director Sam Mendes , who expands his range with flair .	4	2

Table 2: Comparison of Actual vs. Predicted Sentiment for SST Sentiment Analysis

While the model demonstrates a certain level of proficiency in leveraging immediate context for understanding, it noticeably struggles with texts that demand a deeper semantic interpretation or that require reliance on broader contextual knowledge beyond the immediate text. This limitation becomes particularly evident in its handling of nuanced expressions of sentiment and complex paraphrase scenarios where contextual cues are subtle or indirect. For instance, sentences like "Hilariously inept and ridiculous." and "This riveting World War II moral suspense story could not be more timely." showcase the model's difficulty in interpreting irony, nuanced praise, or culturally and temporally situated critiques. Similarly, phrases that contain a positive sentiment embedded within a complex structure, such as "This flick is about as cool and crowd-pleasing as a documentary can be.", further highlight the model's challenges with texts where sentiment is conveyed through intricate comparisons or implicit references. These examples suggest an inability to fully grasp the subtleties and richness of more nuanced expressions.

Additionally, the model's misinterpretations extend to sentences that encapsulate sentiment through specific filmic or literary references, as seen in "It takes a certain kind of horror movie to quote the Talmud, and this ain't it." and "No sophomore slump for director Sam Mendes, who expands his range with flair." These sentences reflect the model's struggles with texts that necessitate an understanding

of specific cultural or domain knowledge, suggesting a superficial grasp of contexts that extend beyond the literal meaning of words. Moreover, the misjudgment of sentiments in sentences like "We haven't seen such hilarity since Say It Isn't So." indicates a broader issue with recognizing humor or evaluating comparative statements that require an appreciation of historical or genre-specific contexts.

Overall, the model's performance variability suggests areas where adaptability does not equate to optimal understanding or processing of complex language nuances. Enhancing the model's ability to understand such language would likely require training data that encompasses a wider range of cultural and idiomatic expressions and allows the model to generalize meanings from these corpuses. Additionally, despite improvements, our model was still performing poorly on examples that emphasized understanding indirect sentiment or nuanced semantic relationships, pointing to the need for methodologies that can imbue the model with a more nuanced understanding of subtle meanings. Also, the model struggled with paraphrase detection and semantic similarity under conditions of significant syntactical or structural diversity, revealing a critical area for improvement. Another potential research path could investigate the generalizability of our training approach by fine-tuning on a broader range of tasks and datasets.

Additional work could address these weaknesses that are present in the functions of our model by developing techniques that focus on advanced context understanding, finer-grained sentiment analysis, and enhanced handling of idiomatic language. Additionally, exploring the integration of cross-domain knowledge and further refining syntactical and structural understanding would greatly improve upon the structural weaknesses of our model.

7 Conclusion

Through the process of refining the minBERT model for small, domain-specific datasets, we discovered that both Masked Language Modeling (MLM) pretraining and ELECTRA-style pretraining with Gradient-Disentangled Embedding Sharing yielded only marginal improvements. This suggests that the observed benefits, particularly slight for MLM pretraining, might stem from random variation rather than substantial model enhancements. However, ELECTRA-style pretraining showed a slight advantage for paraphrase tasks, indicating that its effectiveness could increase with more extensive training examples. We also found using cosine similarity with separate embeddings to be significantly less effective than simply concatenating sentences with a [SEP] token, which materially improved our model's performance.

Techniques like gradient surgery, cosine annealing, and Adam weight regularization proved to be beneficial for reducing overfitting and enhancing BERT model performance on domain-specific datasets and tasks. Although RRFT increased training time, it also yielded significant performance improvements, with scaled batch sizes offering greater improvements.

Looking ahead, future work could explore additional pretraining techniques, expand the model's application to more languages and domains, and conduct an in-depth analysis of training dynamics. Addressing the challenges associated with scaled batch sizes in round-robin training could improve scalability and efficiency. Additionally, evaluating and mitigating biases within minBERT would ensure its fairness and applicability across diverse datasets, contributing to the broader field of natural language processing.

References

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings*, page 194–206, Berlin, Heidelberg. Springer-Verlag.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. 2020. Gradient surgery for multi-task learning.