Knowledge Distillation on Cross-Modal Adversarial Reprogramming for Data-Limited Attribute Inference

Quan Li
Pennsylvania State Univeristy
University Park, PA, USA
qbl5082@psu.edu

Lingwei Chen*
Wright State University
Dayton, OH, USA
lingwei.chen@wright.edu

Shixiong Jing
Pennsylvania State Univeristy
University Park, PA, USA
svj5489@psu.edu

Dinghao Wu*
Pennsylvania State University
University Park, PA, USA
dinghao@psu.edu

ABSTRACT

Social media generates a rich source of text data with intrinsic user attributes (e.g., age, gender), where different parties benefit from disclosing them. Attribute inference can be cast as a text classification problem, which, however, suffers from labeled data scarcity. To address this challenge, we propose a data-limited learning model to distill knowledge on adversarial reprogramming of a visual transformer (ViT) for attribute inferences. Not only does this novel cross-modal model transfers the powerful learning capability from ViT, but also leverages unlabeled texts to reduce the demand on labeled data. Experiments on social media datasets demonstrate the state-of-the-art performance of our model on data-limited attribute inferences.

CCS CONCEPTS

• **Computing methodologies → Natural language processing**; • **Information systems** → *Document representation*.

KEYWORDS

Attribute Inference, Adversarial Reprogramming, Data-limited Learning, Knowledge Distillation.

ACM Reference Format:
Quan Li, Lingwei Chen, Shixiong Jing, and Dinghao Wu. 2023. Knowledge Distillation on Cross-Modal Adversarial Reprogramming for Data-Limited Attribute Inference. In *Companion Proceedings of the ACM Web Conference 2023 (WWW '23 Companion), April 30-May 4, 2023, Austin, TX, USA*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3543873.3587313

1 INTRODUCTION

Social media has drastically changed our everyday lives, which allows us to effortlessly post personal ideas for social engagements [14]. This generates a mass of text data reserving basic yet rich user information, which, more importantly, often implies intrinsic user attributes, such as age, gender, location, and political view. Different parties have been thus attracted to reveal user attributes from their text data, either conscientiously (e.g., for assessing risks and analyzing social behaviors [17, 32]) or opportunistically (e.g., for promoting advertisements and tracking users [14, 33]).

In this paper, we put aside the intents of user attribute inferences, and focus on the investigation of how we can generalize the attribute inference model into a more challenging setting. Due to privacy concerns, most social media websites and apps limit the access to some personal information [16]; thus, user attribute labels, especially for those private attributes, may only be available on few texts. In other words, when we cast user attribute inference on social media as a text classification problem, we face the challenge that the model needs to have the ability to learn from limited text examples.

To address data-limited learning challenge, meta-learning has been proposed to leverage a distribution of tasks to learn a shared initialization that adapts to new task [6, 27, 29]. The classes for meta-training and meta-testing are disjoint, but the data are typically obtained from the same domain [13], which, however, is infeasible in many real-world settings such as social media attribute inference, since annotations for any class are difficult. This leads data-limited attribute inferences to training a single model. Posterior inference using message passing with neural networks over graphs [7, 15, 19, 28, 31] is a promising paradigm of this kind for label propagation, but high memory consumption is in need, especially for large graphs. Another alternative is transfer learning on pretrained models [22, 23], where the prevalent fine-tuning strategy still requires a good amount of labeled samples to yield good results [30].

Adversarial reprogramming [5] shares the same objective as transfer learning that repurposes a neural network pretrained in a source domain to perform a target-domain task, with an additional advantage: only a universal perturbation is learned to the input data, while the pretrained model architecture and parameters keep untouched. Hence it takes less labeled samples for training and copes better with labeled data scarcity issue [1, 2]. To this end, in this paper, we introduce a novel adversarial reprogramming model for our data-limited attribute inferences. As domains and tasks can be completely different in reprogramming, we propose to adversarially reprogram a visual transformer (ViT) [4] for text classification task. The motivations behind this model choice are: (1) ViT enables transformer to capture contexts among patches, but does not necessarily work as language models (e.g., BERT [3]) that are pretrained using masked

*Corresponding authors.

WWW '23 Companion, April 30-May 4, 2023, Austin, TX, USA
© 2023 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9419-2/23/04.
https://doi.org/10.1145/3543873.3587313

language modeling to introduce extra training layers for output transformation [8], and (2) ViT has been undergoing vibrant study, which delivers better learning capability than traditional ImageNet models [9, 12] to extract more expressive patterns from subtle embeddings.

Specifically, given a pretrained ViT and a host image randomly selected from ImageNet, our cross-modal model proceeds by embedding words from each text in order into patches of the host image, and learning a universal perturbation to be added to all inputs, such that the outputs of the ViT can be mapped to the final inference results regarding a specific attribute. These operations are simple and computationally inexpensive. To further leverage unlabeled texts to improve data-limited performance, knowledge distillation is devised to optimize the adversarial reprogramming model for attribute inferences. The overview of our model is illustrated in Figure 1.

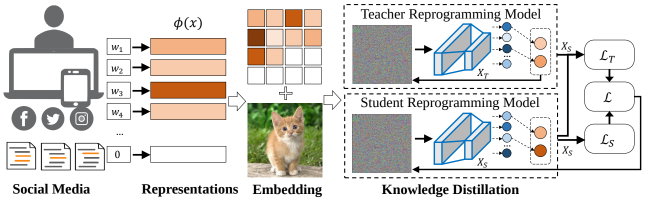

Figure 1: The overview of our proposed model.

2 NOTATIONS AND PROBLEM DEFINITION

Data-limited attribute inference. Without loss of generality, social media text data is presented as $\mathcal{X} = \{(x_i, y_i)\}_{i=1}^{m} \cup \{x_i\}_{i=1}^{n}$ with $m+n$ sample texts. m of them have attribute labels and $m \ll n$. Each labeled text has a ground truth $y \in \mathcal{Y}$ for a specific attribute (e.g., $\mathcal{Y} = \{0\text{:male}, 1\text{:female}\}$ regarding gender). We map discrete text data into d-dimensional feature vectors $\phi : \mathcal{X} \rightarrow \mathbf{X} \subseteq \mathbb{R}^{(m+n)\times d}$. A text classification model $f : \mathbf{X} \rightarrow \mathbf{Y}$ can thus leverage few labeled texts and large unlabeled texts to infer the attribute label of a given text $\mathbf{x}$ as $y^* = \operatorname{argmax}_{y \in \mathcal{Y}} f_y(\mathbf{x})$, where $f_y(\mathbf{x})$ is the confidence score of predicting $\mathbf{x}$ as attribute label y.

Adversarial reprogramming. To repurpose a pretrained model for a new task, adversarial reprogramming relies on nonlinear interactions of the input and the perturbation [2]. A visual transformer $v(\cdot)$ of nonlinear deep structure can satisfy this requirement. We define $\mathbf{x}$ as texts, $f(\mathbf{x})$ a text classification model, $\tilde{\mathbf{X}}$ images, and $v(\tilde{\mathbf{X}})$ ViT. The input transformation $h_v(\cdot; \boldsymbol{\theta})$ comprises embedding $\mathbf{x}$ to a host image $\tilde{\mathbf{X}}$, and learning a universal perturbation $\boldsymbol{\theta}$ added to $\tilde{\mathbf{X}}$ such that $\tilde{\mathbf{X}} = h_v(\mathbf{x}; \boldsymbol{\theta})$; the output transformation $h_f(\cdot)$ maps ImageNet classes to attribute classes such that $f(\mathbf{x}) = h_f(v(\tilde{\mathbf{X}}))$. During model optimization, only the perturbation $\boldsymbol{\theta}$ is trainable.

3 PROPOSED MODEL

3.1 Word Representations

ViT splits an image into patches and takes their linear embedding sequence as an input to a transformer [4]. In other words, the patches of an image are processed in the same way as words of a text in ViT. To proceed with input transformation for adversarial reprogramming of ViT, the first step is to initialize each word w_i of a text x into a feature vector $\mathbf{w}_i \in \mathbb{R}^d$, and each of them can be used for conversion from word to image patch. Pretrained word embeddings are easily accessible, such as GloVe [21], counter-fitting [20], and BERT [3]. As ViT further elaborates transformer to attend patches, learn context-aware information among them, and derive higher-level representations for classifications, here we can simply use any of these embeddings for initialization.

3.2 Input Transformation

Input transformation is to convert a sequence of words from each text $\mathbf{x} = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_l\}$ to image data $\tilde{\mathbf{X}} = \{\tilde{\mathbf{P}}_1, \tilde{\mathbf{P}}_2, \ldots, \tilde{\mathbf{P}}_l\}$ for ViT classification (l is sequence length), which includes image construction and perturbation formulation.

Image construction. If $\mathbf{w}_i$ is derived from BERT ($d = 768$), it can be reshaped directly into $\tilde{\mathbf{P}}_i \in \mathbb{R}^{16\times16\times3}$, which satisfies the patch setting of pretrained ViTs. If we use GloVe or counter-fitting embedding ($d = 300$), $\mathbf{w}_i$ needs to be zero-padded before converting to $\tilde{\mathbf{P}}_i$. Afterwards, these patches are embedded to a host image in order from top left to bottom right. The maximum number of patches or words that can be fitted into the host image is decided by its width W and height H: $L = (W/16) \times (H/16)$. When $l > L$, we remove words with indices larger than L from the sequence.

Perturbation formulation. The perturbation to be added to all new image inputs can be defined as $\tilde{\boldsymbol{\theta}} = \epsilon \cdot \tanh(\boldsymbol{\theta}) \in \mathbb{R}^{H\times W\times3}$, where $\tanh(\cdot)$ bounds the perturbation to be in $(-1, 1)$, and ϵ is a hyperparameter to govern the magnitude of the perturbation. Accordingly, the input transformation function $h_v(\mathbf{x}; \boldsymbol{\theta})$ can be finalized as:

$$\tilde{\mathbf{X}} = h_v(\mathbf{x}; \boldsymbol{\theta}) = \operatorname{clip}(\tilde{\mathbf{X}} + \epsilon \cdot \tanh(\boldsymbol{\theta})) \tag{1}$$

where clip$(\cdot)$ performs per-pixel clipping of the image to limit each pixel value to $[-1, 1]$. As input transformation only involves matrix additions to a host image, it is significantly time-efficient.

3.3 Output Transformation

Output transformation is to map ImageNet classes back to attribute classes to derive inference results. ViT outputs $\tilde{y} \in \{0, 1, \cdots, 999\}$, while attribute inference model outputs $y \in \mathcal{Y}$. Given an attribute to infer, $|\mathcal{Y}|$ is generally smaller than $1,000$. Hence we leverage a simple hard coded mapping method to build output transformation function $h_f(\cdot)$, which assigns $|\mathcal{Y}|$ random class outputs of ViT to predict individual attribute classes. Let $\mathbf{z} = v(\tilde{\mathbf{X}})$, and $h_f(v(\tilde{\mathbf{X}}))$ can be specified as:

$$f(\mathbf{x}) = h_f(v(\tilde{\mathbf{X}})) = \langle \mathbf{z}_{i_1}, \mathbf{z}_{i_2}, \ldots, \mathbf{z}_{i_{|\mathcal{Y}|}} \rangle \tag{2}$$

This mapping is non-parametric that can avoid extra training effort.

3.4 Knowledge Distillation for Optimization

ViT provides powerful learning capability, while reprogramming ViT alone delivers promising inference results. To further leverage unlabeled texts to improve data-limited performance, we devise knowledge distillation [10] to reinforce model optimization. We divide the labeled texts into teacher texts $\mathcal{X}_T$ and student texts $\mathcal{X}_S$. A teacher model is first trained on $\mathcal{X}_T$, and then used to perform inference on $\mathcal{X}_S$. The knowledge distilled by the teacher is defined

Table 1: Comparing statistics of the two datasets

Dataset	Attribute	#Post	#Class	#Vocabulary
Twitter	Gender	13,926	2	21k
Blog	Gender, Age	25,176	2	30k

as inference probability for text $\mathbf{x}_S \in \mathcal{X}_S$:

$$p(\mathbf{x}_S|\mathcal{X}_T) = \frac{\exp\left(f_y(\mathbf{x}_S)/\tau\right)}{\sum_{y \in \mathcal{Y}} \exp\left(f_y(\mathbf{x}_S)/\tau\right)} \tag{3}$$

where τ is distillation temperature. Similarly, a student model is trained on $\mathcal{X}_S$ by generating inference probability $p(\mathbf{x}_S|\mathcal{X}_S)$ and computing the loss between predictions and hard attribute labels:

$$\mathcal{L}_S = -\frac{1}{|\mathcal{X}_S|} \sum_{\mathbf{x}_S \in \mathcal{X}_S} y \log p(\mathbf{x}_S|\mathcal{X}_S) \tag{4}$$

Meanwhile, the student can also learn the distilled knowledge from the teacher by optimizing the loss:

$$\mathcal{L}_T = -\frac{1}{|\mathcal{X}_S|} \sum_{\mathbf{x}_S \in \mathcal{X}_S} p(\mathbf{x}_S|\mathcal{X}_T) \log p(\mathbf{x}_S|\mathcal{X}_S) \tag{5}$$

$p(\mathbf{x}_S|\mathcal{X}_T)$ is predicted by teacher model on student texts, which are unlabeled data to the teacher. It can be considered soft attribute label with the same distribution as $p(\mathbf{x}_S|\mathcal{X}_S)$ from the student. This advances the model to learn from unlabeled texts. The final loss function can be formalized as:

$$\mathcal{L} = (1 - \lambda)\mathcal{L}_S + \lambda\mathcal{L}_T \tag{6}$$

where λ is a balance parameter to trade off $\mathcal{L}_S$ and $\mathcal{L}_T$. By minimizing $\mathcal{L}$ using gradient descent, the only trainable perturbation $\boldsymbol{\theta}$ can be easily derived.

4 EXPERIMENTAL RESULTS AND ANALYSIS

4.1 Experimental Setup

Datasets. We use two real-world social media datasets to perform experimental evaluations: Twitter dataset[1] and Blog dataset [25]. For the Twitter dataset, we filter out those with gender confidence score less than 0.5, and obtain 13,926 tweets with two genders (female and male). For Blog dataset, it consists 25,176 blogs with two attributes: (1) gender (female and male), and (2) age (teenagers (age between 13-18) and adults (age between 23-45)). The statistics of the dataset are shown in the Table 1.

Baselines. We choose single-training baselines for comparisons, including SBERT [24], text-graph classification models: TL-GNN [11], TextGCN [31], TextING [34], and HGAT [18], GNN-based few-shot models: FSGNN [7] and TPN [19], and ViT fine-tuned via transfer learning ViT$_{Tran}$. We also compare adversarial reprogramming of ViT with other backbones: ResNet-50 [9], DenseNet-161 [12], and Inception-V3 [26].

Parameter setting. We select 15 labeled instances per class as training data and randomly select 20% instances from the remaining as test data, and report mean accuracy of 3 runs for each task. We use vit_base_patch16_384, and set learning rate as 0.001, knowledge distillation temperature $\tau = 5$, balance parameter $\lambda = 0.3$, perturbation magnitude $\epsilon = 0.2$, host images size $384 \times 384 \times 3$, and word embedding as BERT. Evaluations are performed on 4 NVIDIA TITAN Xp GPUs with 12GB of RAM each.

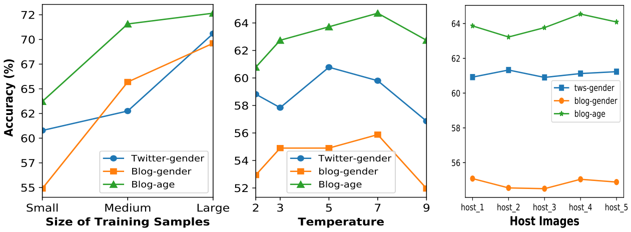

Figure 2: Evaluation on different model parameters.

Table 2: Comparisons of baselines with small data (accuracy %)

Model	Twitter-gender	Blog-gender	Blog-age
SBERT	51.20	50.93	54.59
TL-GNN	50.49	51.26	56.10
TextGCN	49.36	53.43	52.30
TextING	51.36	52.76	58.28
HGAT	52.41	51.69	58.56
FSGNN	57.37	53.61	62.10
TPN	55.20	52.17	53.20
ViT$_{Tran}$	51.98	51.48	57.67
DenseNet-161	55.88	52.94	60.78
ResNet-50	54.90	53.92	61.76
Inception-V3	59.80	53.58	62.74
Our Model (ViT)	**60.78**	**54.90**	**63.72**

4.2 Evaluation of Our Model

Effectiveness. We test our model with three training sizes: small (2×15), medium (2×100), and large (2×2500), and distillation temperatures $\tau \in \{2, 3, 5, 7, 9\}$ when data is small. As shown in Figure 2, our model achieves promising inference results when only 30 labeled texts are available: the inference accuracy is 60.78%, 54.90%, and 63.72% for three inference tasks respectively.

Impact of training size, temperature and host image. As illustrated in Figure 2, larger data leads to better inference accuracy. Also, as many more labeled texts are used and performance is closer to the upper bound, our model seems yielding less data-limited learning advantage. For distillation temperature, when τ is enlarged, the accuracy first rises to a stable high level at $\tau = 5$, and then drops afterwards. When τ is small, the soft labels distilled from the teacher are significant for student model optimization; when τ is large, the distilled knowledge is ambiguous and in turn smooths the student's inference ability. As for host image, the evaluation results slightly vary in five host images (i.e., terrier, toucan, theater, cellphone, and trimaran), but the standard deviations of accuracy are less than 0.5, implying that our model is loosely coupled with host images.

4.3 Comparisons with Baselines

We compare our model with baselines listed in Section 4.1, which are trained on 30 samples. We can observe from Table 2 that our

[1] https://www.kaggle.com/crowdflower/twitter-user-gender-classification

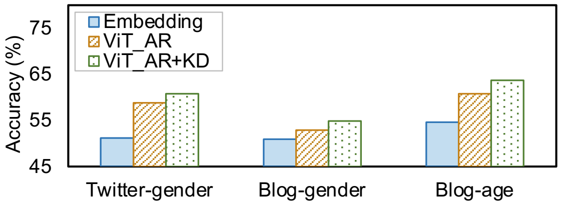

Figure 3: Evaluation on model components.

model outperforms non-reprogramming baselines with a large margin, which confirms that (1) text-graphs built on word co-occurrence barely learn from few labeled texts; (2) FSGNN with message passing on text graph addresses much better data scarcity issue while TPN significantly underperforms; (3) transfer learning still suffers from limited data to fine-tune the whole ViT (about 80M parameters) with accuracy as poor as SBERT; (4) our model benefits from only updating perturbation tensor and enables a better solution for data-limited attribute inference. Most of pretrained models adversarially reprogrammed yield better accuracy than other baselines, which indicates the feasibility of adversarial reprogramming. As ViT uses transformer to process image patches, offering powerful capability to process input texts, and thus our model performs better than other pretrained models.

4.4 Ablation Study

We also conduct ablation study to investigate the component contributions to our model performance. We formulate three models here: (1) Embedding: feed BERT representations to a shallow MLP; (2) ViT$_{AR}$: adversarial reprogramming of ViT; (3) ViT$_{AR+KD}$: the complete design of our model. As illustrated in Figure 3, we can see that reprogramming ViT delivers the greatest contribution to our model, which significantly improves the accuracy from embedding model by $(3.0, 8.6)\%$. Knowledge distillation is able to further advance the state-of-the-art performance to a higher level, which implies that this operation provides an additional advantage for data-limited learning.

5 CONCLUSION

In this paper, we generalize attribute inferences over social media text data into the more challenging yet more realistic setting with limited labels on texts, and design a novel model that distills knowledge on adversarial reprogramming of ViT to address this challenge. We conduct extensive experiments over two social media datasets to evaluate this model. The promising results validate its attribute inference effectiveness, and its feasibility to cope with data-limited learning in practice.

REFERENCES

[1] Lingwei Chen, Yujie Fan, and Yanfang Ye. 2021. Adversarial Reprogramming of Pretrained Neural Networks for Fraud Detection. In *CIKM*. 2935–2939.

[2] Lingwei Chen, Xiaoting Li, and Dinghao Wu. 2022. Adversarially Reprogramming Pretrained Neural Networks for Data-limited and Cost-efficient Malware Detection. In *SDM*. SIAM, 693–701.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT* (2019).

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR* (2021).

[5] Gamaleldin F Elsayed, Ian Goodfellow, and Jascha Sohl-Dickstein. 2018. Adversarial reprogramming of neural networks. *ICLR* (2018).

[6] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*. 1126–1135.

[7] Victor Garcia and Joan Bruna. 2018. Few-shot learning with graph neural networks. *International Conference on Learning Representations* (2018).

[8] Karen Hambardzumyan, Hrant Khachatrian, and Jonathan May. 2021. Warp: Word-level adversarial reprogramming. *ACL* (2021).

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*. 770–778.

[10] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* 2, 7 (2015).

[11] Lianzhe Huang, Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2019. Text level graph neural network for text classification. *EMNLP* (2019).

[12] Forrest Iandola, Matt Moskewicz, Sergey Karayev, Ross Girshick, Trevor Darrell, and Kurt Keutzer. 2014. Densenet: Implementing efficient convnet descriptor pyramids. *arXiv preprint arXiv:1404.1869* (2014).

[13] Ashraful Islam, Chun-Fu Richard Chen, Rameswar Panda, Leonid Karlinsky, Rogerio Feris, and Richard J Radke. 2021. Dynamic distillation network for cross-domain few-shot recognition with unlabeled data. *NeurIPS* (2021).

[14] Jinyuan Jia and Neil Zhenqiang Gong. 2018. Attriguard: A practical defense against attribute inference attacks via adversarial machine learning. In *27th USENIX Security Symposium (USENIX Security 18)*. 513–529.

[15] Quan Li, Xiaoting Li, Lingwei Chen, and Dinghao Wu. 2022. Distilling Knowledge on Text Graph for Social Media Attribute Inference. In *SIGIR*. 2024–2028.

[16] Xiaoting Li, Lingwei Chen, and Dinghao Wu. 2021. Turning Attacks into Protection: Social Media Privacy Protection Using Adversarial Attacks. In *SDM*. 208–216.

[17] Chung-Ying Lin. 2020. Social reaction toward the 2019 novel coronavirus (COVID-19). *Social Health and Behavior* 3, 1 (2020), 1.

[18] Hu Linmei, Tianchi Yang, Chuan Shi, Houye Ji, and Xiaoli Li. 2019. Heterogeneous graph attention networks for semi-supervised short text classification. In *EMNLP-IJCNLP*. 4821–4830.

[19] Yanbin Liu, Juho Lee, Minseop Park, Saehoon Kim, Eunho Yang, Sung Ju Hwang, and Yi Yang. 2018. Learning to propagate labels: Transductive propagation network for few-shot learning. *arXiv preprint arXiv:1805.10002* (2018).

[20] Nikola Mrkšić, Diarmuid O Séaghdha, Blaise Thomson, Milica Gašić, Lina Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. Counter-fitting word vectors to linguistic constraints. *NAACL* (2016).

[21] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*. 1532–1543.

[22] Matthew E Peters, Sebastian Ruder, and Noah A Smith. 2019. To tune or not to tune? adapting pretrained representations to diverse tasks. *arXiv preprint arXiv:1903.05987* (2019).

[23] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. (2018).

[24] Nils Reimers. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *ACL*. Association for Computational Linguistics.

[25] Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. 2006. Effects of age and gender on blogging.. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, Vol. 6. 199–205.

[26] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *CVPR*. 2818–2826.

[27] Risto Vuorio, Shao-Hua Sun, Hexiang Hu, and Joseph J Lim. 2019. Multimodal model-agnostic meta-learning via task-aware modulation. *Advances in Neural Information Processing Systems* 32 (2019).

[28] Yaqing Wang, Song Wang, Quanming Yao, and Dejing Dou. 2021. Hierarchical Heterogeneous Graph Representation Learning for Short Text Classification. *arXiv preprint arXiv:2111.00180* (2021).

[29] Huaxiu Yao, Ying Wei, Junzhou Huang, and Zhenhui Li. 2019. Hierarchically structured meta-learning. In *International Conference on Machine Learning*. PMLR, 7045–7054.

[30] Huaxiu Yao, Ying Wei, Long-Kai Huang, Ding Xue, Junzhou Huang, and Zhenhui Jessie Li. 2021. Functionally Regionalized Knowledge Transfer for Low-resource Drug Discovery. *NeurIPS* 34 (2021).

[31] Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Graph convolutional networks for text classification. In *AAAI*, Vol. 33. 7370–7377.

[32] Yanfang Ye, Shifu Hou, Yujie Fan, Yiyue Qian, Yiming Zhang, Shiyu Sun, Qian Peng, and Kenneth Laparo. 2020. α-Satellite: An AI-driven System and Benchmark Datasets for Hierarchical Community-level Risk Assessment to Help Combat COVID-19. *arXiv preprint arXiv:2003.12232* (2020).

[33] Sixie Yu, Yevgeniy Vorobeychik, and Scott Alfeld. 2018. Adversarial classification on social networks. In *International Conference on Autonomous Agents and MultiAgent Systems*. 211–219.

[34] Yufeng Zhang, Xueli Yu, Zeyu Cui, Shu Wu, Zhongzhen Wen, and Liang Wang. 2020. Every document owns its structure: Inductive text classification via graph neural networks. *arXiv preprint arXiv:2004.13826* (2020).