Camouflaged Chinese Spam Content Detection with Semi-supervised Generative Active Learning

Zhuoren Jiang[1]*, **Zhe Gao**[2]*, **Yu Duan**[2], **Yangyang Kang**[2],
Changlong Sun[2], **Qiong Zhang**[2], **Xiaozhong Liu**[3]†
[1]School of Public Affairs, Zhejiang University, Hangzhou, China
[2]Alibaba Group, Hangzhou & Sunnyvale, China & USA
[3]School of Informatics, Computing and Engineering, IUB, Bloomington, USA
jiangzhuoren@zju.edu.cn,gaozhe.gz@alibaba-inc.com
{derrick.dy,yangyang.kangyy,qz.zhang}@alibaba-inc.com,
changlong.scl@taobao.com,liu237@indiana.edu

Abstract

We propose a **S**emi-supervI**s**ed **G**e**N**erative **A**ctive **L**earning (**SIGNAL**) model to address the imbalance, efficiency, and text camouflage problems of Chinese text spam detection task. A "self-diversity" criterion is proposed for measuring the "worthiness" of a candidate for annotation. A semi-supervised variational autoencoder with masked attention learning approach and a character variation graph-enhanced augmentation procedure are proposed for data augmentation. The preliminary experiment demonstrates the proposed SIGNAL model is not only sensitive to spam sample selection, but also can improve the performance of a series of conventional active learning models for Chinese spam detection task. To the best of our knowledge, this is the first work to integrate active learning and semi-supervised generative learning for text spam detection.

1 Introduction

The recent successes of learning-based models all share the same prerequisite: a decent labeled training dataset is available for a given task (Jiang et al., 2019b; Arora and Agarwal, 2007). However, the annotating process can be "*a tedious, laborious, and time consuming task for humans*" (Sharma et al., 2015). To achieve high task performance with low labeling cost, (pool-based) active learning (Cohn et al., 1996) algorithms are proposed to select the most representative and informative sample to be labeled by human oracles (Druck et al., 2009). Although effective in general, in Chinese text spam detection context, the following reasons make the active learning a challenging task:

Imbalance: in reality, the ratio of spam samples to normal ones is very imbalanced. For instance, in North America, "*much less than 1% of SMS messages were spam*" (Almeida et al., 2013). As a result, the active learning model should be more sensitive to spam samples. The general active learning methods, e.g., (Lewis and Gale, 1994; Li and Guo, 2013; Roth and Small, 2006), can hardly address this problem. **Efficiency**: when competing with anti-spam models, spammers are constantly creating new forms for spam texts (Xie et al., 2012; Jiang et al., 2019a). The amount of unlabeled samples is huge and keeps increasing. Classical diversity-based approach (Brinker, 2003; Xu et al., 2003), which iteratively compares each unlabeled sample with each labeled sample to select the most "diverse" ones for annotating, will perform poorly as its computational complexity is $O(n^2)$. An efficient-oriented active learning algorithm is needed. **Camouflage**[1]: Chinese character has glyph and phonetic variations (Norman, 1988), e.g., "账 (account)" and "帐(curtain)" have the similar structure and pronunciation. Spammers can take advantage of this characteristic to escape from the detection algorithms (Jindal and Liu, 2007; Jiang et al., 2019a). It is important to propose a novel active learning model that can predict the new Chinese character variation patterns not appearing in the labeled dataset.

To address these challenges, we propose a novel solution, **S**emi-supervI**s**ed **G**e**N**erative **A**ctive **L**earning (**SIGNAL**) model to naturally integrate active learning and semi-supervised generative learning into a unified framework. SIGNAL is

*These two authors contributed equally to this research.
†Corresponding author

[1]"Camouflaged text spam" refers to the intentional mutation of Chinese character to escape from the spam detection algorithms. The variation-based spam text is purposely created and highly camouflaged for machine learning algorithms. Typos of normal text is not spam.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 3080–3085
July 5 - 10, 2020.

inspired by a simple yet powerful observation in computer vision domain (Zhou et al., 2017) : the patches generated from the same image share the same label, and are naturally expected to have similar predictions by the classifier. Hence, the diversity of predictions of patches can successfully measure the "power" of a candidate image in elevating the performance of the current classifier. Similarly, in this study, a set of semantically similar texts for each candidate sample is automatically generated through data augmentation. We hypothesize that: *the diversity of predictions of augmented texts is a useful indicator to predict the boost ability of a candidate text sample for the performance of the classifier.* We define this strategy as a **"self-diversity"** based active learning strategy.

Algorithmically, unsupervised generative models, such as variational autoencoder (Kingma and Welling, 2013), only learn to generate similar texts without considering the labeling information. Therefore, we utilize a Semi-supervised Variational AutoEncoder (S-VAE) (Kingma et al., 2014) to automatically generate semantically similar texts for each candidate sample, while trying to keep the label-consistency. To enable S-VAE to gain the ability of perceiving the sensitive positions of the candidate sample, we enrich the human annotation feedback. The annotator is required to provide not only a label for the candidate but also a rationale (critical terms in the candidate) (Sharma et al., 2015) for the chosen spam label. Based on the human-annotated rationales, we introduce a pseudo-mask distribution P_m to guide the attention learning in S-VAE. A character variation graph-enhanced augmentation procedure is then applied to integrate the Chinese character variation knowledge and simulate the glyph and phonetic variation mutations in further data augmentation.

Compared with conventional active learning, SIGNAL offers three advantages: (1) SIGNAL is more sensitive to seek the spam samples[2]. (2) SIGNAL does not need to compare with the labeled samples, which reduces its computational complexity to $O(N)$. (3) SIGNAL considers the heterogeneous variation knowledge of Chinese characters for spam detection.

The major contributions of this paper can be summarized as follows:

1. We propose a SIGNAL model, in the context of Chinese text spam detection, to address the imbalance, efficiency, and text camouflage problems. To the best of our knowledge, this is the first work to integrate active learning and semi-supervised generative learning for text spam detection task.

2. The preliminary experiments on the Chinese SMS dataset demonstrate the efficacy and potential of SIGNAL for Chinese spam detection. A series of conventional active learning models can be improved after merging the SIGNAL model.

3. While focusing on the Chinese spam detection task in this study; theoretically, SIGNAL has a great potential to be applied in other NLP tasks. It can mitigate the **data-hungry problem** by cutting the labeling cost.

2 SIGNAL Model

Figure 1 depicts the proposed SIGNAL framework[3]. It starts with a small set of labeled samples, a large set of unlabeled samples, and an initial classifier trained on the labeled samples. The goal of SIGNAL is to seek "salient" samples from the pool of unlabeled samples for annotation. Then the classifier can be continuously improved by incrementally enlarging the training set with newly annotated samples. The pseudocode of SIGNAL is described as Algorithm 1.

Self-Diversity Based Active Learning. As aforementioned, in SIGNAL, we develop a "self-diversity" criterion for active candidate selection. Formally, for a candidate sample x_i, a set of augmented texts $AT_i = \left\{at_i^1, at_i^2, \cdots, at_i^j \cdots, at_i^M\right\}$ is generated. The self-diversity SD_i of x_i can be defined as:

$$SD_i = \frac{\sum_M^{j=1}\left(p_i^j - \bar{p_i}\right)^2}{M} \quad (1)$$

p_i^j is the prediction of the current classifier for augmented text at_i^j; $\bar{p_i}$ is the arithmetic mean of all predictions for AT_i; M is the total number of augmented texts. SD suggests the "worthiness" of a candidate for annotation. A large SD indicates that the current classifier's prediction for the target candidate is unstable. With a slight mutation, the prediction will change drastically. Such a candidate is worthy of annotation. This criterion has the potential to locate the vital samples and also to reduce the computational complexity. Furthermore,

[2]More detailed information can be found in the experiment section.

[3]https://github.com/Giruvegan/generative-camouflaged-spam-detector

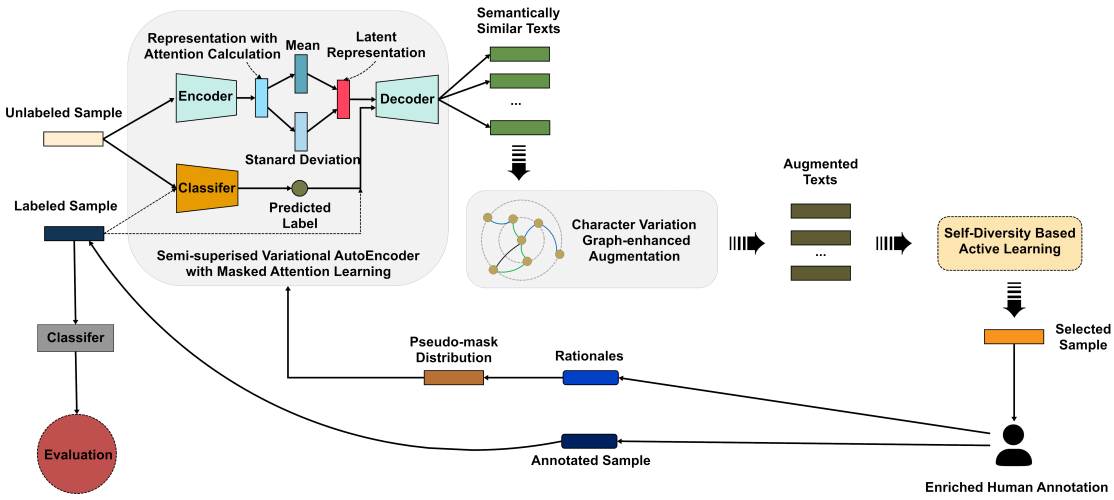

Figure 1: An Illustration of "SIGNAL" Framework

in the context of Chinese text spam detection, spam candidate has a greater possibility to gain a larger SD. For instance, if the spam candidate mutates at the critical positions, the label of the augmented text is likely to change. On the contrary, normal candidates are less likely to be affected by this situation.

S-VAE with Masked Attention Learning. As shown in Figure 1, we utilize S-VAE with masked attention learning to generate similar texts at the semantic level. In this study, with annotated rationales $\mathbb{R}$ (a set of critical terms), a pseudo-mask distribution P_m is generated for each candidate sample. For i_{th} term t_i of the candidate sample, the pseudo-mask probability Pr_i can be calculated as:

$$Pr_i = \frac{\rho \mathbb{I}_{\mathbb{R}}(t_i)}{\Delta} \tag{2}$$

where $\mathbb{I}_{\mathbb{R}}(t_i)$ is an indicator function to determine whether t_i belongs to $\mathbb{R}$; Δ is used for normalization; ρ is the weight to ensure the critical terms will have less attention, in other words, it can have a greater possibility to be "masked" during the generative process.

Following (Kingma et al., 2014), the generative semi-supervised model with masked attention learning can be defined as:

$$\begin{aligned} &Pr(y) = \mathrm{Cat}(y|\pi); \\ &Pr(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I}); \\ &Pr_\omega(\mathbf{x}'|f_r(\mathbf{x})) = f_a(\mathbf{x}'; f_r(\mathbf{x}), \omega); \\ &Pr_\theta(\mathbf{x}'|y, z) = f(\mathbf{x}'; y, z, \theta) \end{aligned} \tag{3}$$

where $\mathbf{x}$ is a sample (labeled or unlabeled); $f_r(\mathbf{x})$ is a matrix generated by a non-linear transformation of $\mathbf{x}$. $\mathbf{x}'$ is a representation of $f_r(\mathbf{x})$ with an attention calculation, $\mathbf{x}' = \sum \omega_i f_r(\mathbf{x})_i$; ω denotes the attention distribution, $\omega_i = \mathrm{softmax}(f_c(f_r(\mathbf{x}))_i$, which is scalar; f_c is an single-dimensional non-linear transformation; $\mathrm{Cat}(y|\pi)$ is the multinomial distribution, if x is unlabeled, the class labels y are treated as latent variables; z is the latent variable; θ denotes the parameters of a non-linear transformation. Labeled samples can be used to train a classifier that predicts class labels y. During the inference process, we can predict the missing class for an unlabeled sample from the inferred posterior distribution $Pr_\theta(y|\mathbf{x}')$.

The loss function of S-VAE with masked attention learning is defined as:

$$\mathcal{L} = \mathcal{L}_{\mathcal{S}-\mathcal{VAE}} + \alpha \mathcal{D}_{\mathcal{KL}}(P_{att}||P_m) \tag{4}$$

where $\mathcal{L}_{\mathcal{S}-\mathcal{VAE}}$ is the loss of original S-VAE (Kingma et al., 2014); $\mathcal{D}_{\mathcal{KL}}(P_m||P_{att})$ is the KL divergence of the attention distribution P_{att} from the pseudo-mask distribution P_m.

Character Variation Graph-enhanced Augmentation. In this study, a random-walk based graph-enhanced augmentation procedure is used for integrating the Chinese character variation knowledge and simulating the glyph and phonetic variation mutations. A Chinese character variation graph G (Jiang et al., 2019a) is utilized. $G = (C, R)$. C denotes the Chinese character (vertex) set. R denotes the variation relation (edge) set, and edge weight is the similarity of two characters given the target relation (variation) type. For critical positions in a piece of text, we adopt a random walk based graph exploration to predict the possible Chinese character variation patterns. For

Algorithm 1 Semi-supervised Generative Active Learning

Self-Diversity Based Active Learning (Labeled set: $\mathbb{L}$, Unlabeled set: $\mathbb{U} = \{x_1, \cdots, x_N\}$, Initial Classifier: $C_t, t = 0$, Chinese Character Variation Graph: G, Annotated Rationales: $\mathbb{R}$)
$\mathbb{R} = \varnothing$
repeat
  for all $x_i \in \mathbb{U}$ **do**
    With $\mathbb{R}$, generate a pseudo-mask distribution P_m^i using Eq.2
    SS_i = **S-VAE**(x_i, P_m^i)
    AT_i = **GraphAugmentation**(SS_i, G, P_m^i)
    With AT_i and C_t, calculate SD_i using Eq.1
  end for
  Select top K unlabeled samples $\mathbb{Q}$ from U
  Get $\hat{\mathbb{L}}$ and $\hat{\mathbb{R}}$ from enriched human annotation
  $\mathbb{L} \leftarrow \mathbb{L} \bigcap \hat{\mathbb{L}}, \mathbb{R} \leftarrow \mathbb{R} \bigcap \hat{\mathbb{R}}, \mathbb{U} \leftarrow \mathbb{U} / \mathbb{Q}$
  $t++, C_t \leftarrow Train(\mathbb{L}, C_{t-1})$
until Convergence
return $C_t, \mathbb{L}$

GraphAugmentation(Similar text set: SS, Chinese Character Variation Graph: G, pseudo-masked distribution P_m,)
$AT = \varnothing$
for all $ss_j \in SS$ **do**
  Probabilistically generate a position list POS with P_m
  for all $pos_k \in POS$ **do**
    Get the character Ch_{pos_k} at position pos_k
    $Ch_o \leftarrow Ch_{pos_k}$
    Randomly generate a walking step $T_p \in (0, T]$
    Ch_n = RandomWalk(Ch_o, T_p, G)
    $Ch_{pos_k} \leftarrow Ch_n$
  end for
  Append ss_j to AT
end for
return AT

more detailed information on this procedure, please refer to Algorithm 1.

3 Preliminary Experiment

Dataset and Experiment Setting. A Chinese SMS dataset[4] was used for the experiment. There were 48,896 testing samples, including 23,891 spam samples and 25,005 normal samples. The size of the active learning sample set was 48884, including 23,891 spam samples and 24,993 normal samples. 200 samples were randomly selected as the initial labeled set. The remaining samples were used as an unlabeled sample pool. For each iteration, 100 samples were selected by different active learning models. The iterative active learning process repeated 10 times. For evaluation, a single-layer CNN classifier was trained on the labeled samples. **Uncertainty** (Lewis and Gale, 1994), **Margin** (Roth and Small, 2006), and **Entropy** (Li and Guo, 2013) were chosen as baseline models. Similar baseline-settings can be found in (Zhou et al., 2017; Huang et al., 2018; Yoo and Kweon, 2019).

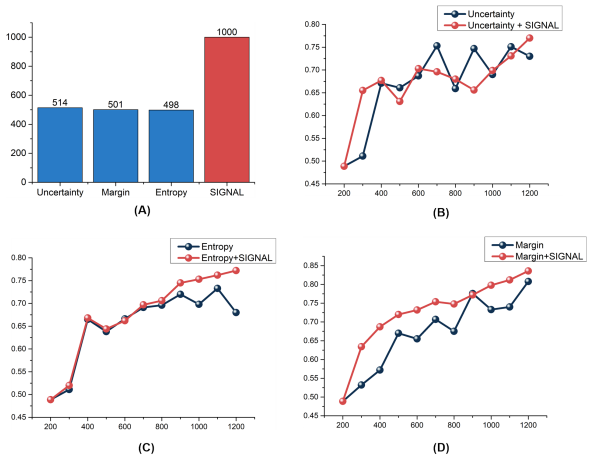

Figure 2: Preliminary Experiment Result: (A) the number of selected spam samples after 10 iteration of active learning; (B) the classifier performance (accuracy) comparison between "Uncertainty" and "Uncertainty merging SIGNAL"; (C) the classifier performance (accuracy) comparison between "Entropy" and "Entropy merging SIGNAL"; (D) the classifier performance (accuracy) comparison between "Margin" and "Margin merging SIGNAL"

In SIGNAL model, for S-VAE training[4], we chose "BiGRU+ Attention + MLP" as encoder structure, a "single-layer GRU" as decoder structure, and a "single-layer CNN+MLP" as classifier. For each candidate sample, 10 augmented texts is generated for "self-diversity" calculation.

Sensitivity of Spam Sample Selection. As shown in Figure 2 (A), compared with baseline models, SIGNAL can be more sensitive to spam samples. The selected spam samples from SIGNAL were significantly more than those from other baselines. This observation indicated the potential of SIGNAL for addressing the "imbalance" problem in Chinese text spam detection.

The Elevating "Power" of SIGNAL. As shown in Figure 2 (B), (C), and (D), after merging[5] SIGNAL, all baseline models had been improved to varying degrees. Especially for margin-based active learning (Roth and Small, 2006), SIGNAL can improve the performance in all active learning iterations. Averagely, by merging SIGNAL, **Margin** can be improved by 10% in the metric of the

[4]https://github.com/Giruvegan/generative-camouflaged-spam-detector

[5]In the preliminary experiment, we apply a simple yet effective merging strategy: in each iteration, the baseline model and SIGNAL model select 50 samples respectively.

Type	Text	Label	Variation
Original Sample	白色丝袜免费按流程领取,需要详情加微信l*****d Meaning: According to the process, white stockings are free of charge, friend me on WeChat if you want to know more l*****d.	Spam (Scam & Ads)	No Variation
Augmented Text 1	小妞光腿袜门槛最低的10元忧惠券，先注意抽奖奖 c****c Meaning: 10 yuan minimum coupon for chick's bare-leg socks, remember to draw first c****c	Spam (Scam & Ads)	Glyph Variation (忧←——优)
Augmented Text 2	不要忘了墙先加购，小的绝对谢谢这你的惠顾 c****c Meaning: Don't forget to put the item into the shopping cart first, thank you for your purchase c****c	Normal	Phonetic Variation (墙←——抢)

Figure 3: Case study: augmented texts from SIGNAL

classification performance.

Case Study. To gain a straightforward understanding of the generation quality of SIGNAL, we present two augmented texts in Figure 3. From these two cases, we have the following observations: (1) the augmented texts are semantically similar to the original sample. (2) Although the original sample has no variation character, the augmented texts can simulate the phonetic or glyph variation mutations. (3) If the critical terms in the original sample are replaced, the label of text can be different.

4 Conclusion

In this paper, we propose a SIGNAL model for Chinese text spam detection. SIGNAL integrates active learning and semi-supervised generative learning into a unified framework. As an exploration study for this newly proposed problem, the preliminary results have revealed the potential of SIGNAL to address the critical problems in the proposed task. For instance, Figure 2 (A) proves that SIGNAL can be more sensitive to spam samples (*Imbalance Challenge*); case study (Figure 3) shows the generation capacity of SIGNAL to simulate the phonetic or glyph variation mutations (*Camouflage Challenge*); comparing to classical diversity-based approach, we integrate self-diversity based active learning and generative learning which can greatly reduce the computational complexity ($O(N) \rightarrow O(N)$, *Efficiency Challenge*).

In the future, we plan to enable the glyph and phonetic variation detection by integrating the variation graph representation learning, which may improve SIGNAL's performance.

Acknowledgments

We are thankful to the anonymous reviewers for their helpful comments. This work is supported by Alibaba Group through Alibaba Research Fellowship Program, the National Natural Science Foundation of China (61876003), Guangdong Basic and Applied Basic Research Foundation (2019A1515010837), and the Opening Project of State Key Laboratory of Digital Publishing Technology (cndplab-2020-Z001).

References

Tiago Almeida, José María Gómez Hidalgo, and Tiago Pasqualini Silva. 2013. Towards sms spam filtering: Results under a new dataset. *International Journal of Information Security Science*, 2(1):1–18.

Shilpa Arora and Sachin Agarwal. 2007. Active learning for natural language processing. *Language Technologies Institute School of Computer Science Carnegie Mellon University*.

Klaus Brinker. 2003. Incorporating diversity in active learning with support vector machines. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, pages 59–66.

David A Cohn, Zoubin Ghahramani, and Michael I Jordan. 1996. Active learning with statistical models. *Journal of artificial intelligence research*, 4:129–145.

Gregory Druck, Burr Settles, and Andrew McCallum. 2009. Active learning by labeling features. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume*

1-Volume 1, pages 81–90. Association for Computational Linguistics.

Sheng-Jun Huang, Jia-Wei Zhao, and Zhao-Yang Liu. 2018. Cost-effective training of deep cnns with active model adaptation. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1580–1588.

Zhuoren Jiang, Zhe Gao, Guoxiu He, Yangyang Kang, Changlong Sun, Qiong Zhang, Luo Si, and Xiaozhong Liu. 2019a. Detect camouflaged spam content via stoneskipping: Graph and text joint embedding for chinese character variation representation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6188–6197.

Zhuoren Jiang, Jian Wang, Lujun Zhao, Changlong Sun, Yao Lu, and Xiaozhong Liu. 2019b. Cross-domain aspect category transfer and detection via traceable heterogeneous graph representation learning. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 289–298. ACM.

Nitin Jindal and Bing Liu. 2007. Review spam detection. In *Proceedings of the 16th international conference on World Wide Web*, pages 1189–1190. ACM.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. 2014. Semi-supervised learning with deep generative models. In *Advances in neural information processing systems*, pages 3581–3589.

David D Lewis and William A Gale. 1994. A sequential algorithm for training text classifiers. In *SIGIR'94*, pages 3–12. Springer.

Xin Li and Yuhong Guo. 2013. Adaptive active learning for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 859–866.

Jerry Norman. 1988. *Chinese*. Cambridge University Press.

Dan Roth and Kevin Small. 2006. Active learning with perceptron for structured output. In *ICML Workshop on Learning in Structured Output Spaces*.

Manali Sharma, Di Zhuang, and Mustafa Bilgic. 2015. Active learning with rationales for text classification. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 441–451.

Sihong Xie, Guan Wang, Shuyang Lin, and Philip S Yu. 2012. Review spam detection via temporal pattern discovery. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 823–831. ACM.

Zhao Xu, Kai Yu, Volker Tresp, Xiaowei Xu, and Jizhi Wang. 2003. Representative sampling for text classification using support vector machines. In *European conference on information retrieval*, pages 393–407. Springer.

Donggeun Yoo and In So Kweon. 2019. Learning loss for active learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 93–102.

Zongwei Zhou, Jae Shin, Lei Zhang, Suryakanth Gurudu, Michael Gotway, and Jianming Liang. 2017. Fine-tuning convolutional neural networks for biomedical image analysis: actively and incrementally. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7340–7351.