國立臺灣大學電機資訊學院資訊工程學系
碩士論文
Department of Computer Science and Information Engineering
College of Electrical Engineering and Computer Science
National Taiwan University
Master Thesis

預訓練對於醫療影像的探討
Rethinking Pre-training in Medical Imaging

張友誠
Yu-Cheng Chang

指導教授：徐宏民博士
Advisor: Winston Hsu, Ph.D.

中華民國 109 年 7 月
July, 2020

doi:10.6342/NTU202002487

國立臺灣大學碩士學位論文
口試委員會審定書

預訓練對於醫療影像的探討

Rethinking pre-training in medical imaging

本論文係張友誠君（學號 R07922058）在國立臺灣大學資訊工程學系完成之碩士學位論文，於民國 109 年 7 月 27 日承下列考試委員審查通過及口試及格，特此證明

口試委員：

（指導教授）

陳文進　　葉梅珍

蘇東弘　　李綱

系主任　莊永裕

誌謝

回顧這兩年的碩士生涯，很高興能夠加入這個實驗室 (CMLab-MiRA)，在這個研究氛圍濃厚的環境下，各種有趣又具挑戰性的研究議題正如火如荼的進行。首先，最感謝的人莫過於我的指導教授徐宏民教授 (Prof. Winston Hsu)，每周的團隊討論中，從研究議題的鎖定、問題釐清、架構設計到論文寫作上，親力親爲的提供協助，並且不斷地鼓勵我秉持深入鑽研的初衷，在醫療影像處理的領域上持續精進，才能有這兩年來的研究成果。此外，也感謝實驗室提供充沛的資源，讓我能夠專心在研究上而無其餘後顧之憂。

除此之外，我很幸運能夠遇到實驗室的同儕：智遠、旻昇、岳承、哲宇、雅量、宸晞、與晟、俞安以及昱翔，大家總是不吝分享自己的研究，讓我有機會學習各式各樣的研究議題，並激發我一些研究上的想法。特別感謝研究夥伴智遠，完整參與我碩班兩年的研究課題，包含挑選研究主題、方法實作、實驗的設計與驗證、到最後論文寫作並投稿，每次的討論總是有意想不到的收穫，一步步奠定了穩固的根基，才能有最終的研究成果。由衷地感謝一起奮鬥的各個實驗室夥伴，無論在學業或生活上對於我的幫助。

最後，感謝家人給我的支持，至始至終鼓勵並且相信我；感謝京樺陪我練習口試報告，並且提供了我很多研究上的想法；感謝東毅一起參與研究，並共同發表論文到國際會議上。兩年的生涯即將劃下句點，回頭審視這段時間的種種，心中滿是懷念與感謝，在此獻上我衷心的感謝給所有曾經幫助過我的人。

doi:10.6342/NTU202002487

摘要

模型預訓練 (Pre-training) 在電腦視覺及自然語言處理的領域中被廣泛地使用，透過巨量的訓練資料，模型可以習得泛用的影像或文字的特徵提取能力，進一步幫助網路模型有更好的表現。然而，鮮少的文獻深入探討是否有哪些因素會影響模型預訓練在醫療影像領域的效果，因此本篇研究著重探討預訓練的本質，透過大量的分析闡明預訓練對於醫療影像處理的限制，且進一步提出新穎的解決方法。

透過實驗分析驗證，問題複雜度與模型的預訓練資料型態 (Modality) 皆對於預訓練的效果有著明顯的影響。我們也分析了批量標準化 (Batch normalization) 當中的縮放參數 (Scaling term γ)，並且建立一套高效率的模型能力評估方法。除此之外，我們更進一步提出神經網路煉金術 (Network Alchemy)，有系統性的激發模型的潛能，以妥善利用模型所有的可學習參數。大量的實驗結果說明在各式各樣的實驗設定中，我們的方法均能夠提升模型的表現，展示其泛化性與穩定性。

關鍵字： 深度學習、醫療影像、預訓練

doi:10.6342/NTU202002487

Abstract

Pre-training is a well-developed technique to extract general feature representations from abundant data. However, the factors affecting how pre-training works in medical imaging is rarely studied. In this work, we fully explore the essence of pre-training in medical imaging and provide comprehensive analysis. We conclude that both the target task complexity and the pre-trained data modality have considerable impact on the effectiveness of pre-training in medical imaging. In addition, we analyze the trainable parameter γ in batch normalization (BatchNorm) and establish an original standard to efficiently assess the effectiveness of pre-trained weights. We further propose the *Network Alchemy* to stimulate the considerable potential of the network and fully utilize model parameters in fine-tuning stage. Extensive experimental results exhibit the robustness and the generalization ability of our proposed methodology in various experimental scenarios.

Keywords: Deep Learning, Medical Imaging, Pre-training

doi:10.6342/NTU202002487

Contents

doi:10.6342/NTU202002487

doi:10.6342/NTU202002487

List of Figures

doi:10.6342/NTU202002487

List of Tables

doi:10.6342/NTU202002487

doi:10.6342/NTU202002487

Chapter 1

Introduction

Pre-training on large-scale data (*e.g.*, ImageNet) has achieved great success because the network could learn transferable feature representations and further improve the target task performance. A well-established practice is initializing the network with ImageNet pre-trained weights and then fine-tuning on downstream tasks, such as object detection [21, 20], image segmentation [27, 23], and action recognition [36, 10]. In addition, the development in natural language processing (NLP) has also made remarkable progress as the success of language representations pre-training [16, 40, 14, 33, 9]. Due to the considerable achievement of pre-training in computer vision and NLP, how to adapt it to the medical domain becomes a central topic [37, 35, 11, 43]. Recently, Chen *et al*.confirm the effectiveness of the pre-trained network with the established medical dataset which contains various image modalities, organs, and objective tasks [11]. Zhou *et al*.pre-train the network in a self-supervised learning fashion and show the pre-trained network outperforms the one trained from scratch in many medical imaging tasks [43]. However, these works do not pay much attention on studying the limitations of pre-training in medical imaging. We wonder whether pre-training will always give significant boost to the target task performance and whether the characteristics of pre-trained dataset have an influence on the representation learning. Therefore, we conduct comprehensive pilot studies to reveal the essence of pre-training under various experimental settings. Based on the pilot studies, we come to two conclusions: (i) pre-training is helpful for achieving better performance when the downstream task is challenging; (ii) the weights pre-trained from the

doi:10.6342/NTU202002487

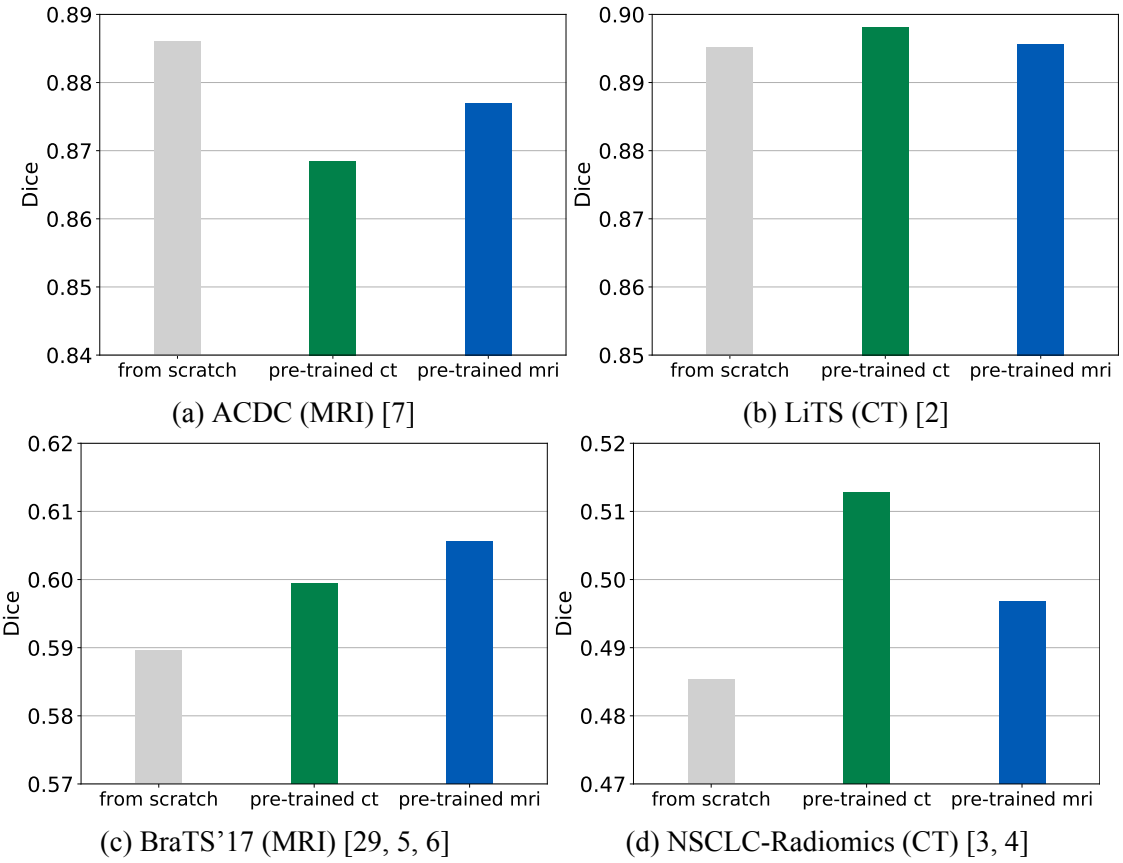
same modality as the downstream data surpass that trained from different modality.

In reality, the diversity of medical imaging increases with the rapid evolution of hardware and scanning protocol in the radiology and pathology [18, 38, 22]. Therefore, collecting the data in the same modality for taking full advantage of pre-training is impractical. However, we wonder whether we could still benefit from pre-training for the complicated target task even the network pre-trained from other modality data. Once the aforementioned issue has been tackled, we can leverage publicly accessible pre-trained weights straightforwardly. In recent, the trainable per-feature coefficient γ in batch normalization (BatchNorm) is intensively studied in [19]. Their findings inspire us to carry out the in-depth analysis of γ and offer noteworthy insights into the correlation between the power of pre-trained weights and the distribution of whole γ in the network. Based on these observations, we propose the *Network Alchemy* to guide the network to make full use of all trainable parameters. Besides, we further establish an original standard to measure the effectiveness of pre-trained weights, which can provide the pilot estimate to decrease the cost of pre-trained networks selection. We evaluate the proposed network alchemy on multiple medical imaging tasks. The experimental results demonstrate that pre-trained weights from other modality can be properly refined and achieve convincing performance on the downstream task.

Our main contributions can be summarized as follows:

- We conduct detailed pilot studies and conclude that both the target task complexity and the pre-trained data modality have considerable impact on the effectiveness of pre-training in medical imaging.
- We thoroughly explore the scaling parameter γ in BatchNorm and highlight that it can efficiently assess the capability of pre-trained weights.
- We propose the novel network alchemy to properly adapt pre-trained weights to boost the performance in the fine-tuning stage.
- Extensive experimental results demonstrate the generalization ability and the robustness of the proposed methodology.

doi:10.6342/NTU202002487



(a) ACDC (MRI) [7]

(b) LiTS (CT) [2]

(c) BraTS'17 (MRI) [29, 5, 6]

(d) NSCLC-Radiomics (CT) [3, 4]

Figure 1.1: **Analysis of the effectiveness of pre-training.** We make the comparison between different network initialization. We use the pre-trained weights from large-scale CT and MRI datasets [41, 1], denoted as 'pre-trained ct' and 'pre-trained mri'. The results demonstrate that the pre-training technique does not always give significant boost to the performance of target tasks. Besides, applying the weights pre-trained from the data in the same modality significantly improves the performance of downstream tasks compared with the weights from other modality.

doi:10.6342/NTU202002487

Chapter 2

Related work

2.1 Pre-training

Pre-training model with the large-scale dataset (*e.g.*, ImageNet) and then fine-tuning it on the downstream task has become a common practice in computer vision field. By exploiting abundant data, the pre-trained network can learn informative representations and transfer the knowledge to the target task. It has enabled promising results on object detection [21, 20], image segmentation [27, 23], and action recognition [36, 10]. However, supervised pre-training such as ImageNet requires massive manual labels, which is impractical in practice. Therefore, self-supervised learning receives increased attention as it can learn representations from the data itself without explicit manual supervision [17, 42, 30, 31, 12]. Through reasonable pre-text tasks such as colorizing a grayscale image [42], solving jigsaw puzzles [30], and contrastive learning [12], the network can be guided to learn meaningful visual representations. On the other hand, BERT [16] and all the following landmark breakthroughs [40, 14, 33, 9] have achieved great success in the field of Natural Language Processing (NLP) by constructing unsupervised tasks to learn the context in corpus.

doi:10.6342/NTU202002487

2.2 Pre-training in medical imaging

In view of the significant achievement in pre-training, there are some researches attempt to adapt the well-established paradigm to medical imaging domain [37, 35, 11, 43]. Lately, Chen *et al*.establish a large-scale 3D medical dataset consisting of multiple public datasets with various modalities and pre-train a model called MedicalNet [11]. Experimental results demonstrate that it can facilitate training convergence and boost the performance in multiple tasks. However, collecting numerous data with annotation is extremely time-consuming and labor-intensive especially in medical field since the labels need to be annotated by domain experts. As a result, Zhou *et al*.design appropriate pre-text tasks for medical imaging and build a set of pre-trained networks called ModelsGenesis in the manner of self-supervised learning [43]. These self-taught models outperform those trained from scratch on multiple downstream datasets. However, the aforementioned studies do not deeply explore the essence of pre-training in medical imaging. We therefore carry out a comprehensive analysis and extensive experiments to reveal the characteristics of pre-training.

2.3 Batch normalization

Batch normalization (BatchNorm) [25] has been widely adopted in neural network architecture as it can stabilize the training process and facilitate the convergence. BatchNorm performs the z-norm on the features with the mean and variance computed within a batch. Besides, the trainable coefficient γ and β parameters were introduced to ensure that the transformation can represent the identity mapping. Since BatchNorm was proposed, multiple works were seeking to explore the essence of it and trying to understand the reason for its effectiveness [34, 8, 28, 39, 19]. In particular, the expressive power of γ and β was studied in [19]. By freezing all other parameters at initialization and training only γ and β, the sufficiently deep models have impressive performance on ImageNet [15]. In addition, γ naturally learns to disable substantial random features. This demonstrated that the affine parameters in BatchNorm enable the network to have the ability of selecting appropriate

doi:10.6342/NTU202002487

features. We further extend this observation and develop a methodology to evaluate the capability of the pre-trained weights. Moreover, we propose the BNScale regularization to fully and evenly utilize the kernels in the fine-tuning stage.

doi:10.6342/NTU202002487

Chapter 3

Methodology

3.1 Pilot study

In this section, we empirically examine the hypotheses we suggest on four downstream datasets, namely ACDC [7], LiTS [2], BraTS'17 [29, 5, 6], and NSCLC-Radiomics [3, 4]. Besides, we pre-train the networks on FastMRI [41] and LIDC-IDRI [1], whose modality is CT and MRI, respectively. The detailed description of used datasets is shown in Tab. 3.1. Since the downstream tasks are segmentation problems, we adopt the Dice coefficient to evaluate the network performance in the following experiments.

Table 3.1: **Dataset description.** Detail information of used datasets in this work.

Dataset	Modality	Purpose	Task	Total subjects (Train:Valid:Test)
FastMRI [41]	MRI	pre-training	-	3769 (3769 : - : -)
LIDC-IDRI [1]	CT	pre-training	-	534 (534 : - : -)
ACDC [7]	MRI	fine-tuning	heart segmentation	180 (120 : 30 : 50)
LiTS [2]	CT	fine-tuning	liver segmentation	131 (84 : 21 : 26)
BraTS'17 [29, 5, 6]	MRI	fine-tuning	brain tumor segmentation	285 (171 : 57 : 57)
NSCLC-Radiomics [3, 4]	CT	fine-tuning	lung tumor segmentation	288 (169 : 84 : 35)

3.1.1 Is pre-training always useful in medical imaging?

Pre-training seems to be omnipotent for tackling various tasks in medical domain [37, 35, 11, 43]. However, is pre-training always useful in medical imaging? As shown in Figure 1.1, there is a significant improvement on BraTS'17 and NSCLC-Radiomics for over 1% Dice score when the modality of pre-trained dataset is the same as that of downstream

doi:10.6342/NTU202002487

dataset. On the contrary, the performance on LiTS just slightly enhances and the performance on ACDC even decreases. We conclude that the complexity of downstream task is critical to the effectiveness of pre-training. A possible reason is that the network trained from scratch is good enough on the simple tasks compared with the difficult ones, so the improvement is not obvious. Besides, the task complexity is inversely correlated with the performance of networks trained from scratch on a specific task. For example, as the Dice score on LiTS is higher than the one on NSCLC-Radiomics by a huge margin, the task on NSCLC-Radiomics is much more complicated than the one on LiTS. In addition, the task of BraTS'17 and NSCLC-Radiomics is tumor segmentation while the one of ACDC and LiTS is organ segmentation. Tumors have a variety of size and position while organs are almost located in the particular region, which is straightforward to explain why the task complexity of BraTS'17 and NSCLC-Radiomics is greater than the one of ACDC and LiTS.

3.1.2 Does the modality of pre-trained datasets matter?

There are numerous modalities in medical imaging, such as CT, MRI, X-ray, and PET. To the best of our knowledge, there is no research in medical imaging domain analyzes how the modality of pre-trained data will impact on the performance of downstream tasks. As depicted in Fig 1.1, applying the pre-trained weights from the same modality can significantly improve the performance of target tasks compared with the ones from different modality. For instance, the network pre-trained from MRI outperforms that pre-trained from CT in the BraTS'17 [29, 5, 6] challenge since it is a brain MRI dataset. It is worth pointing out that the phenomenon can be consistently observed even in the easier organ segmentation tasks. We conclude that which modality is adopted to pre-train the network plays a important role in the improvement of the downstream tasks. It is intuitive that the pre-trained weights from the same modality will perform well as the model has understood the data characteristics in advance.

doi:10.6342/NTU202002487

3.2 Proposed approach

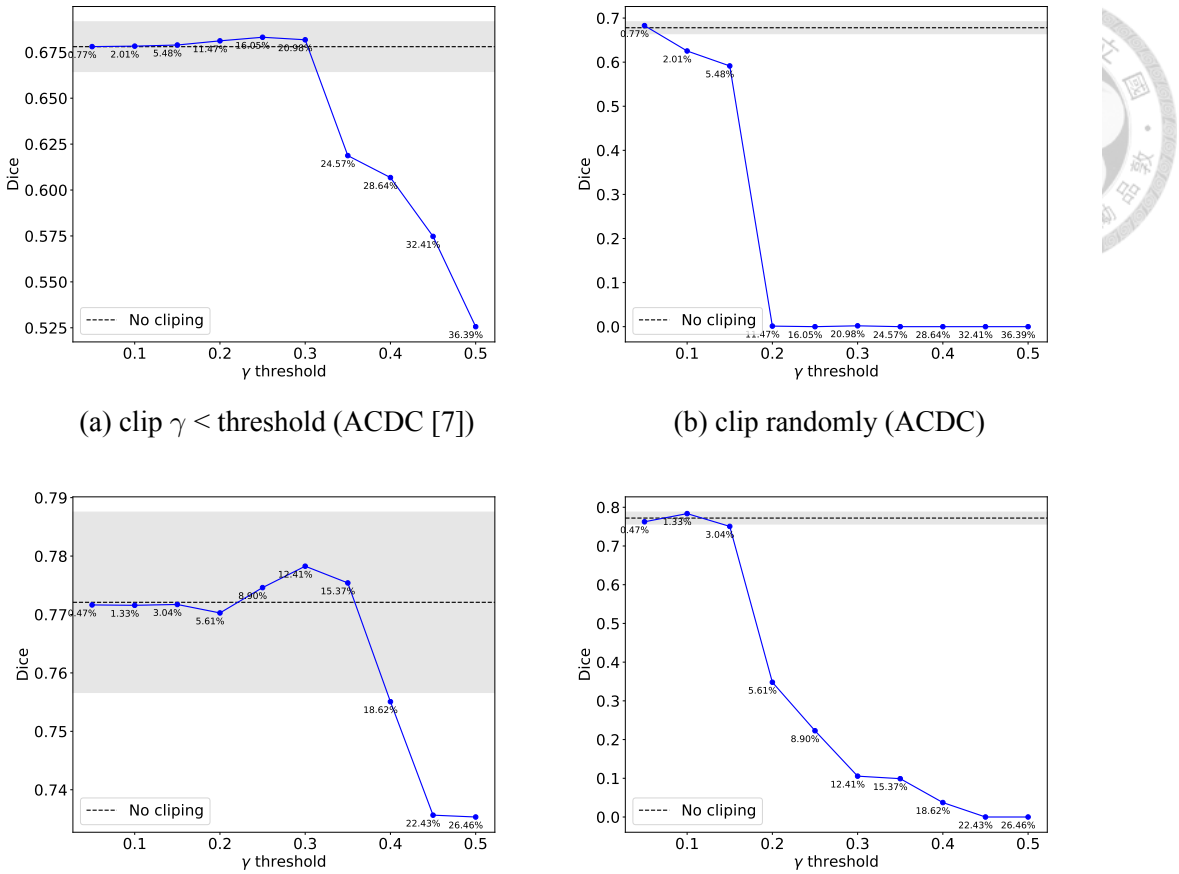
In this section, we first provide detailed studies on BatchNorm parameter γ in Sec. 3.2.1. Based on the analyses, we propose the *network alchemy* composed of *Identification, Modification, and Maximization* to appropriately refine the pre-trained weights to be competent at the target tasks. We elaborate the proposed approach in Sec. 3.2.2.

3.2.1 Analysis on γ in batch normalization

Impact of small γ on network output In [19], they have confirmed that small values of γ have little impact on the network output by explicitly clipping the parameters whose value is lower than a specified threshold to zero and evaluate the performance. Nonetheless, it may result from the amount of clipped γ rather than the value of γ. Therefore, we perform clipping γ under a specified threshold and randomly clipping with corresponding percentage of γ to verify this assumption. As shown in Fig 3.1, there is a slight drop in Dice score when the kernels with smaller γ value are disabled. In contrast, randomly closing equal percentage of kernels leads to a dramatic drop on the performance. Therefore, the number of clipped γ is not the main factor for affecting the performance and the kernels with smaller γ value do impact little on the network prediction.

Redundant kernels We have observed the correlation between γ values and the importance of the corresponding kernels in the previous analysis. More specifically, the γ value a well metric to reflect the importance of each kernel. Therefore, we can measure the percentage of redundant kernels in a network by freezing all the parameters except the parameters in BatchNorm and then analyzing the γ distribution. Fig. 3.2 expresses the γ distribution of networks with different initialization in four downstream tasks. As can be observed, the randomly initialized model has more γ values distributed in a lower range than pre-trained networks by a large margin. In addition, the network pre-trained from data whose modality is identical to the one of downstream data contains less redundant kernels for the current task. From the view of the effectiveness of network parameters, the results demonstrated in Fig 1.1 are reasonably explained.

doi:10.6342/NTU202002487



(a) clip $\gamma <$ threshold (ACDC [7])

(b) clip randomly (ACDC)

(c) clip $\gamma <$ threshold (LiTS [2])

(d) clip randomly (LiTS)

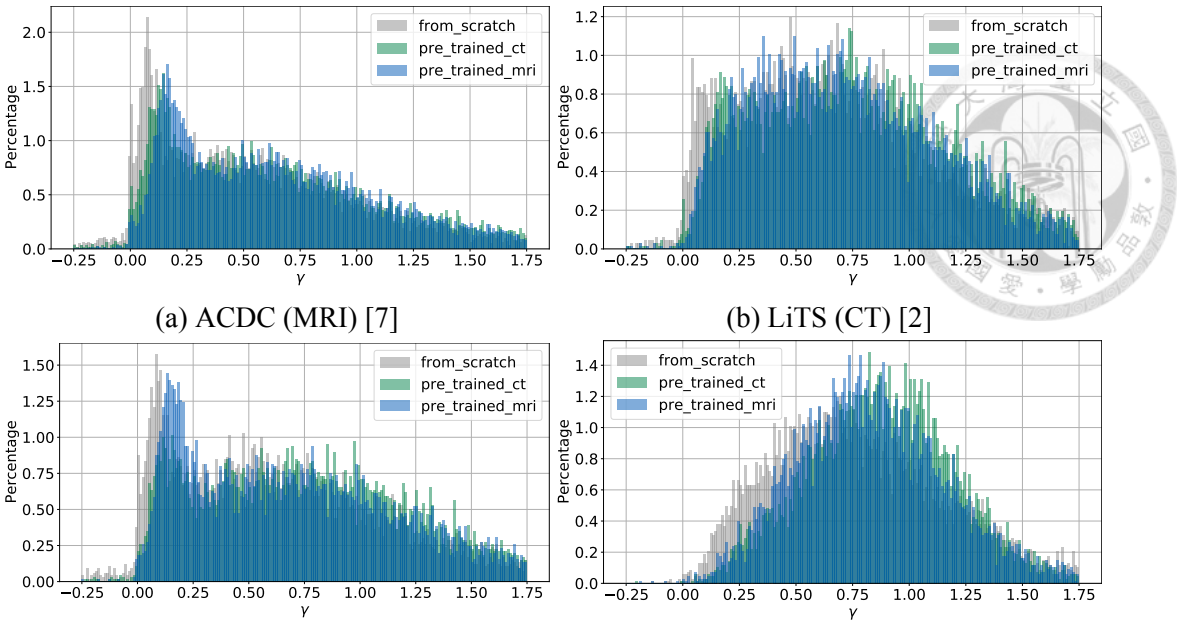
Figure 3.1: **Verification of small γ impact.** We compare clipping γ whose value is smaller than a specified threshold with randomly clipping. The number annotated on each point represents the percentage of γ equals to zero. The gray region covers the acceptable variation range of performance ($\pm 2\%$). It is obvious that there is only a slight drop in Dice coefficient when performing clipping on the smaller γ. On the contrary, randomly disabling equal percentage of kernels leads to a dramatic decrease on the performance.

3.2.2 Network Alchemy

We propose the network alchemy to separate the wheat from the chaff and make full use of the pre-trained weights. Let $\mathcal{L}$ denote the task-specific loss function in the fine-tuning stage. We elaborate the steps in the following sections.

Identification In this stage, we only train the BatchNorm with $\mathcal{L}$ and freeze other layers to leverage the γ distribution for judging which kernel is redundant. After reaching the convergence, the γ values are representative enough to be the importance of the corresponding kernels. Hence, we identify which kernel contributes little to the output predic-

doi:10.6342/NTU202002487



(a) ACDC (MRI) [7]

(b) LiTS (CT) [2]

(c) BraTS'17 (MRI) [29, 5, 6]

(d) NSCLC-Radiomics (CT) [3, 4]

Figure 3.2: γ **distribution.** We acquire the γ distribution by training only BatchNorm. As shown in the figures, the network pre-trained from data whose modality is identical to the one of downstream data contains less redundant kernels.

tion and further modify them.

Modification After identifying the useless kernels, the next step is to properly adjust those kernels and make it competent at the current objective. We substitute kernels with Kaiming initialization [24] for the redundant kernels and then fine-tune the whole network with $\mathcal{L}$. Surprisingly, the preformance is substantially improved (Tab. 4.1). The possible reason may be that the well-studied Kaiming initialization [24] is the better initial parameters in comparison with the already identified redundant kernels.

Maximization Finally, we propose the BNScale regularization consisting of the mean and the standard deviation terms to further regularize the network parameters when fine-tuning. The regularization is formulated as:

$$\mathcal{R}_\gamma = -\alpha\, \mathbb{E}(\Gamma) + \beta\sqrt{Var(\Gamma)} \tag{3.1}$$

where Γ is the set of γ in BatchNorm of the whole network, $\mathbb{E}$ is the expectation operator, Var is the standard deviation symbol, and α, β are hyperparameters. Therefore, the loss

doi:10.6342/NTU202002487

function in this stage can be written as:

$$\mathcal{L}' = \mathcal{L} + \lambda \mathcal{R}_{\gamma} \tag{3.2}$$

where λ is a scalar to adjust the intensity of regularization.

We have verified that if the γ in BatchNorm are close to zero, the corresponding kernels are redundant. Hence, the mean term is tailored to expect the γ to be higher, which forces the corresponding kernels to improve their capability. On the other hand, the network naturally learns to disable some kernels, which makes the model too rely on other specified kernels [19]. To avoid this problem, the standard deviation term is designed to expect the variation of γ to be small, which compels the model to equally use every kernel. With these explicit constraints, the network will be guided to make full use of whole parameters and converge to better local minima.

To sum up, the proposed network alchemy elegantly unlocks the network potential and achieve better performance. Different from the conventional fine-tuning, we explicitly provide the regularization on the network parameters. This provides the clear guidance on pointing in the direction of the network optimization.

doi:10.6342/NTU202002487

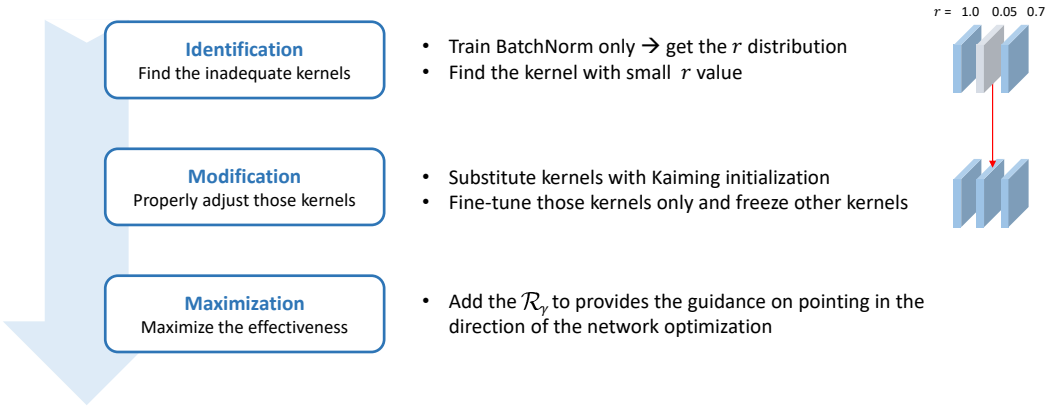


Figure 3.3: **Network Alchemy.** The proposed approach includes three steps: Identification, Modification, and Maximization. First, we identify redundant kernels with the help of γ distribution. Next, we properly modify these kernels to make the network more competent at the current task. Finally, we imply the additional regularization to provide the direct guidance to assist the network optimization.

doi:10.6342/NTU202002487

Chapter 4

Experiments

The used data in the experiments has been described in Sec. 3.1. We evaluate the effectiveness and the generalization ability of our proposed method only on BraTS'17(MRI) [29, 5, 6] and NSCLC-Radiomics [3, 4] as they are more complicated task than ACDC [7] and LiTS [2]. We also provide the comprehensive ablation study to demonstrate the effectiveness of our proposed components. In addition, we propose the original standard to rapidly estimate the performance of pre-trained weights, which can largely bring down the estimation cost.

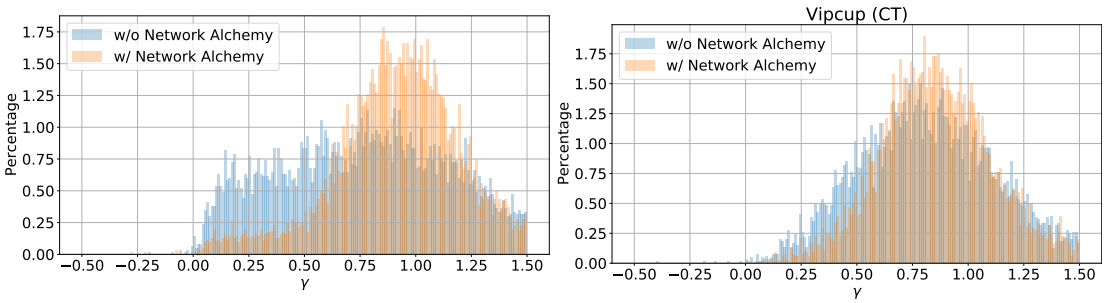
4.1 Implementation details

For model pre-training, we adopt 3D U-Net [13] as the backbone and follow the self-supervised pre-text tasks proposed in [43] to acquire the pre-trained weights. For the fine-tuning on downstream tasks, the data is resampled to the fixed image resolution and then augmented by random cropping. We empirically choose $\alpha = 0.1, \beta = 1$ in Eq. 3.1 and $\lambda = 0.0001$ in Eq. 3.2 We use Adam [26] optimizer with β_1 = 0.9, β_2 = 0.999, and ϵ = 10^{-8}, and the batch size is 16. The training is stopped once the loss curve is converged on the validation set. We do k-fold cross-validation to rigorously examine the generalization ability of our proposed method. We implement with Pytorch framework [32] and conduct the experiments on NVIDIA V100 GPUs.

doi:10.6342/NTU202002487

4.2 Quantitative evaluation

To verify the effectiveness of the proposed network alchemy, we initialize the network with the pre-trained weights from the modality different from the one of target dataset. For example, we adopt the model weights pre-trained from CT data as the initialization and fine-tune on BraTS'17 whose data modality is MRI. We report the quantitative results in Table 4.1 and depict the γ distribution in Fig 4.1. We can see that there are significant improvements in both datasets. In term of the γ distribution, there are much less parameters with small γ value after performing the network alchemy, which explains the showed improvements. Therefore, the proposed network alchemy is effective for elegantly increasing the network capability even pre-trained weights are from other modality. The experimental results demonstrate that pre-trained weights from other modality can be properly refined and achieve convincing performance on the downstream task.



(a) BraTS'17 [29, 5, 6]

(b) NSCLC-Radiomics [3, 4]

Figure 4.1: **Change of γ distribution after performing network alchemy.** As we can see, there are much less parameters with small γ value after performing the network alchemy.

4.3 Ablation study

As shown in Tab. 4.1, pre-trained network enables the promising result compared with the network trained from scratch. Besides, the pre-trained weights from the same modality can achieve better performance as mentioned in Sec. 3.1.2. Here we verify whether the proposed network alchemy will reach the performance comparable to the model pre-trained from the same modality. We first analyze Identification and Modification in network alchemy to confirm whether these two steps will really separate the wheat from the

doi:10.6342/NTU202002487

chaff and benefit the network. Then, we also provide the experiment to justify the effectiveness of BNScale regularization. We can see that it leads to the better results with only the help of Identification and Modification. Furthermore, Maximization guides the network to fully utilize the trainable parameters and increases the network capability. In summary, the model assisted with the network alchemy can fully exploit the capability to equal or even surpass to the network pre-trained from the data whose modality is identical to the target dataset modality.

Meanwhile, we additionally provide the investigation on the effectiveness of BNScale regularization. Surprisingly, as shown in Tab. 4.2, BNScale regularization brings the much improvement in most experiments even when training from scratch. The results strongly exhibit the robustness and the generalization of the proposed regularization.

Table 4.1: **Ablation study.** The red and the blue indicate the best and the second best performance, respectively. As mentioned in Sec. 3.1.2, *pre-trained (same modality)* will lead to better performance than *pre-trained (diff. modality)*. Therefore, *pre-trained (same modality)* can be seen as an upperbound that *pre-trained (diff. modality)* can achieve. We verify whether the proposed components in the network alchemy can properly adjust the pre-trained weights from different modality (*pre-trained (diff. modality)*). The result turns out that the network alchemy can effectively reduce the performance gap between *pre-trained (same modality)* and *pre-trained (diff. modality)*.

Pre-trained (same modality)	Pre-trained (diff. modality)	Network Alchemy			Dice↑ (BraTS'17)	Dice↑ (NSCLC-Radiomics)
		Identification	Modification	Maximization		
✓					0.6056	0.5129
					0.5896	0.4854
	✓				0.5995	0.4969
	✓	✓	✓		0.6031	0.5114
	✓	✓	✓	✓	0.6045	0.5163

Table 4.2: **Quantitative analysis of proposed BNScale regularization ($\mathcal{R}_\gamma$).** The results exhibit the effectiveness of the proposed BNScale regularization. It is worth noting that the BNScale regularization can significantly improve the performance even when training with randomly initial weights.

	BraTS'17 (MR) [29, 5, 6]			NSCLC-Radiomics (CT) [3, 4]		
	From scratch	Pre-trained CT	Pre-trained MRI	From scratch	Pre-trained CT	Pre-trained MRI
w/o $\mathcal{R}_\gamma$	0.5896	0.5995	**0.6056**	0.4854	0.5129	0.4969
w/ $\mathcal{R}_\gamma$	**0.5974**	**0.6036**	0.6040	**0.5108**	**0.5212**	**0.5151**

doi:10.6342/NTU202002487

Table 4.3: **Quantitative verification of the proposed evaluation standard.** There exists the positive correlation between the performance of training all modules and training only BatchNorm, which proves the rationality of the proposed evaluation procedure.

	BraTS'17 [29, 5, 6]			NSCLC-Radiomics [3, 4]		
	From scratch	Pre-trained CT	Pre-trained MRI	From scratch	Pre-trained CT	Pre-trained MRI
BN trainable	0.0542	0.2738	0.2885	0.1254	0.2809	0.2291
All trainable	0.5896	0.5995	0.6056	0.4854	0.5129	0.4969

4.4 Pre-trained network evaluation standard

Based on the characteristic of γ observed in Sec. 3.2.1, we introduce the original standard to efficiently assess the performance of the pre-trained network. By training only BatchNorm and freeze other parameters, the model can adjust the γ value to select useful and important kernels during the optimization process. After reaching the convergence, the network has learned how to properly perform the downstream task by just scaling and shifting the features from convolutional kernels rather than making any modification on kernel weights. As reported in Tab. 4.3, the positive correlation between the performance of training all modules and training only BatchNorm exists. Accordingly, we can train only BatchNorm to efficiently select the best one from amounts of pre-trained networks for the current target task. By doing so, it greatly diminishes the computational cost and the overall memory usage since only the parameters in BatchNorm need to be optimized.

doi:10.6342/NTU202002487

Chapter 5

Conclusion

In this work, we carry out the detailed analysis to explore the essence of pre-training in medical imaging and come to two conclusions. One is that the improvement yielded from pre-training is in inverse proportion to the downstream task complexity. The other is the more similar the modality of pre-trained data to the one of downstream data, the better performance we can achieve. Besides, we study the γ in batch normalization (BatchNorm) in depth and establish an efficient procedure to evaluate the effectiveness of existing pre-trained weights. Based on these observations, we propose network alchemy method to further exploit the ability of model parameters in fine-tuning stage. Quantitative results and the ablation study demonstrate that our proposed algorithm is effective to maximize the utility of existing pre-trained weights. In the future work, we are interested in extending the approach to natural images. We believe that our findings would be able to stand the tests even beyond the medical domain.

doi:10.6342/NTU202002487

Bibliography

[1] Data from lidc-idri.

[2] lits - liver tumor segmentation challenge.

[3] H. Aerts, E. Rios Velazquez, R. T. Leijenaar, C. Parmar, P. Grossmann, S. Carvalho, and P. Lambin. Data from nsclc-radiomics. the cancer imaging archive, 2015.

[4] H. J. Aerts, E. R. Velazquez, R. T. Leijenaar, C. Parmar, P. Grossmann, S. Carvalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature communications*, 2014.

[5] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, J. B. Freymann, K. Farahani, and C. Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4:170117, 2017.

[6] S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, R. T. Shinohara, C. Berger, S. M. Ha, M. Rozycki, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *arXiv preprint arXiv:1811.02629*, 2018.

[7] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. G. Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE transactions on medical imaging*, 37(11):2514–2525, 2018.

doi:10.6342/NTU202002487

[8] N. Bjorck, C. P. Gomes, B. Selman, and K. Q. Weinberger. Understanding batch normalization. In *Advances in Neural Information Processing Systems*, pages 7694–7705, 2018.

[9] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[10] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[11] S. Chen, K. Ma, and Y. Zheng. Med3d: Transfer learning for 3d medical image analysis. *arXiv preprint arXiv:1904.00625*, 2019.

[12] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

[13] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016.

[14] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019.

[15] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

doi:10.6342/NTU202002487

[17] C. Doersch, A. Gupta, and A. A. Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015.

[18] M. Eichenlaub, K. Astheimer, J. Minners, T. Blum, C. Restle, C. Maring, S. Schweitzer, U. Thiel, F.-J. Neumann, T. Arentz, et al. Evaluation of a new ultralow-dose radiation protocol for electrophysiological device implantation: A near-zero fluoroscopy approach for device implantation. *Heart Rhythm*, 17(1):90–97, 2020.

[19] J. Frankle, D. J. Schwab, and A. S. Morcos. Training batchnorm and only batchnorm: On the expressive power of random features in cnns. *arXiv preprint arXiv:2003.00152*, 2020.

[20] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[21] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[22] Y. Goltsev, N. Samusik, J. Kennedy-Darling, S. Bhate, M. Hale, G. Vazquez, S. Black, and G. P. Nolan. Deep profiling of mouse splenic architecture with codex multiplexed imaging. *Cell*, 174(4):968–981, 2018.

[23] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[24] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[25] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

doi:10.6342/NTU202002487

[26] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[27] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[28] P. Luo, X. Wang, W. Shao, and Z. Peng. Towards understanding regularization in batch normalization. *arXiv preprint arXiv:1809.00846*, 2018.

[29] B. Menze, A. Jakab, S. Bauer, J. Kalpathy-cramer, K. Farahani, J. Kirby, et al. The multimodal brain tumor image segmentation benchmark (brats). medical imaging. *IEEE Transactions on*, pages 1–32, 2014.

[30] M. Noroozi and P. Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pages 69–84. Springer, 2016.

[31] A. v. d. Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[32] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[33] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019.

[34] S. Santurkar, D. Tsipras, A. Ilyas, and A. Madry. How does batch normalization help optimization? In *Advances in Neural Information Processing Systems*, pages 2483–2493, 2018.

doi:10.6342/NTU202002487

[35] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers. Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging*, 35(5):1285–1298, 2016.

[36] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.

[37] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312, 2016.

[38] R. Werner, T. Sentker, F. Madesta, T. Gauer, and C. Hofmann. Intelligent 4d ct sequence scanning (i4dct): concept and performance evaluation. *Medical physics*, 46(8):3462–3474, 2019.

[39] G. Yang, J. Pennington, V. Rao, J. Sohl-Dickstein, and S. S. Schoenholz. A mean field theory of batch normalization. *arXiv preprint arXiv:1902.08129*, 2019.

[40] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5753–5763, 2019.

[41] J. Zbontar, F. Knoll, A. Sriram, M. J. Muckley, M. Bruno, A. Defazio, M. Parente, K. J. Geras, J. Katsnelson, H. Chandarana, et al. fastmri: An open dataset and benchmarks for accelerated mri. *arXiv preprint arXiv:1811.08839*, 2018.

[42] R. Zhang, P. Isola, and A. A. Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016.

[43] Z. Zhou, V. Sodha, M. M. R. Siddiquee, R. Feng, N. Tajbakhsh, M. B. Gotway, and J. Liang. Models genesis: Generic autodidactic models for 3d medical image analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 384–393. Springer, 2019.

doi:10.6342/NTU202002487