공학석사 학위논문

두부 계측 엑스선 영상을 이용한 교정 진단에서 거리 학습과 자기 지도 학습의 영향

Effects of Metric Learning and Self-Supervised Learning on Orthodontic Diagnosis using Cephalogram

울산대학교 대학원

의 과 학 과

김 성 철

두부 계측 엑스선 영상을 이용한 교정 진단에서 거리 학습과 자기 지도 학습의 영향

지도교수 김 남 국

이 논문을 공학석사 학위 논문으로 제출함

2021 년 08 월

울산대학교 대학원

의 과 학 과

김 성 철

김성철의 공학석사학위 논문을 인준함

심사위원 김 윤 지 (인)

심사위원 김 남 국

심사위원 윤 지 혜

울 산 대 학 교 대 학 원

2021 년 08 월

감사의 글

약 4 년전, 인공지능을 처음으로 공부하면서 의공학도로서 의료인공지능 분야의 한 획을 긋고 싶다는 꿈을 가졌습니다. 꿈을 실현하기위해 MI2RL 에서 시작한 석사과정이 벌써 2 년이 지나 도착점을 바라보고 있습니다. 이곳에서 쌓은 지식과 수 많은 경험은 꿈을 향한 새로운 시작의 소중한 양분이 될 것이라고 자부합니다.

2 년의 석사 과정 동안 강하지만 부드러운 조언으로 연구자로서의 정체성을 확립할 수 있도록 도와주신 김남국 교수님, 연구실에 처음 들어왔을 때 사수로서 큰 길잡이가 되어주신 윤지혜 학자님께 진심으로 감사의 인사를 드립니다. 두 분의 말씀을 항상 귀담아 들으며 연구를 바라보는 시각의 수준을 높일 수 있었고, 해외 학회 참여를 통해 현재에 안주하지않고 새로운 지식을 끊임없이 습득하는 연구자로 성장할 수 있었습니다.

연구실 생활동안 서로에게 긍정적인 영향을 주고 받으며 함께 성장할 수 있었던 김인환 선생님, 장령우 선생님, 조성만 선생님, 연구를 함께 진행하며 항상 좋은 피드백을 주신 이경화 선생님, 김민규 박사님, 저의 뒤를 이어 연구실에서 석사과정을 시작한 의공학부 후배 경성구 선생님, 이정욱 선생님, 졸업을 함께 준비하며 힘이 되어주신 김민지 선생님, 그리고 함께 일했던 조경진 선생님, 김준식 선생님과 MI2RL 모든 선생님들께 진심으로 감사의 인사를 드립니다. 저의 잠재력을 알아봐주시고 좋은 기회를 많이 제안해주시는 서울대병원 김영곤 교수님께도 고개 숙여 감사의 인사를 드립니다.

바쁘신 와중에도 졸업 심사위원을 맡아주신 서울아산병원 김윤지 교수님께 진심으로 감사의 인사를 드립니다. 덴티스트리 과제를 진행하며 좋은 연구 성과를 낼 수 있도록 최고의 인사이트를 제공해주신 서울대치과병원 백승학 교수님, 긴 시간동안 주기적으로 소통하며 좋은 연구 결과를 위해 힘써주신 임선진 선생님께도 고개 숙여 감사의 인사를 드립니다.

마지막으로 항상 저를 믿고 응원해주시고 지지해주시는 부모님과 사랑하는 동생 영선이, 그리고.. 8 년이라는 오랜 시간을 넘어 평생 서로의 옆을 지켜줄 든든한 동반자 나경이에게 감사의 인사와 사랑한다는 말을 전합니다.

Abstract

Deep learning has been applied to various fields, showing remarkable performance improvement. However, when published studies are applied to subtasks derived from common tasks or different domains, the performance improvement is often not as significant as expected or declines. To solve these problems, two types of representation learning, a study that skillfully handles the features obtained from the model, have been actively studied: *metric learning* and *self-supervised learning*.

In this study, two experiments were conducted to check how representation learning, especially metric learning and self-supervised learning, affects medical images using cephalogram: 'orthodontic diagnosis with cephalogram' and 'the effect of self-supervised learning on orthodontic diagnosis'.

In the first study, three orthodontic diagnoses were conducted: anteroposterior skeletal discrepancies (APSD: Class I, Class II, and Class III), vertical skeletal discrepancies (VSD: normo-divergent, hyper-divergent, and hypo-divergent), and vertical dental discrepancies (VDD: normal overbite, open bite, and deep bite). To avoid ‘the gray zone’ where individual diagnoses are overlapped, ArcFace was added to the existing model. Also, Group Normalization was used for stable training with small data instead of Batch Normalization. As a result, the proposed model has consistently shown good performance in internal validation and external validation.

In the second study, pretext task was conducted using SimSiam, one of the self-supervised learning models, and downstream task was conducted in APSD. For comparison, randomly initialized weights and weights pre-trained on ImageNet dataset were used. As a result of the linear evaluation and fine tuning, SimSiam showed better performance in full and low data regimes and did not induce overfitting compared to training from Scratch and ImageNet.

Both studies confirmed that metric learning and self-supervised learning in medical images could improve performance, extract discriminative features, and train models that are robust to data distribution and the number of data. In the future medical image artificial intelligence, research that incorporates representation learning should be conducted rather than simply evaluating performance by learning by model.

Abbreviations

AI (Artificial Intelligence)

SNUDH (Seoul National University Hospital)

KADH (Kooalldam Dental Hospital)

AJUDH (Ajou University Dental Hospital)

AMC (Asan Medical Center)

CNUDH (Chonnam National University Dental Hospital)

CSUDH (Chosun University Dental Hospital)

EUMC (Ewha University Medical Center)

KHUDH (Kyung Hee University Dental Hospital)

KNUDH (Kyungpook National University Dental Hospital)

WKUDH (Wonkwang University Dental Hospital)

ANB (A point-Nasion-B point angle)

FMA (Frankfort mandibular plane angle)

FHR (Jarabak's posterior/anterior facial height ratio)

GT (Ground truth)

SD (Standard deviation)

BN (Batch Normalization)

GN (Group Normalization)

GAP (Global Average Pooling)

Contents

Contents of Tables

Contents of Figures

Introduction

Backgrounds

Among artificial intelligence (AI) technologies, deep learning has been applied to various fields, such as computer vision, natural language processing, reinforcement learning, recommendation system, and database, showing remarkable performance improvement. The performance of general tasks, such as classification, detection, segmentation, automatic summarization, machine translation, and question answering, seems to have been converged. Researches that have achieved 1-2% performance improvement by modifying specific parts of the existing algorithm are dominantly compared to studies that achieved a considerable performance improvement as in 2-3 years ago.

However, when the same study is applied to subtasks derived from these tasks or different domains, the performance improvement is often not as significant as expected or declines. One of the various reasons to be the main thing is that the model structure implemented for general tasks is not suitable for subtasks or other domains. To solve this problem, two types of representation learning, a study that skillfully handles the features obtained from the model, have been actively studied: *metric learning*, which usually is based on the distance between each feature, and *self-supervised learning*, which is used to obtain data-specific features. These researches have in common that they cluster the given data into specific rules. After these processes, it has the advantage of improving performance and helping researchers visually check the feature distribution.

These methods described above are essential elements in the medical domain. Unlike natural images, the information of the image can be decided from only a few pixels to most pixels of the image, the pixel intensity is distributed from 8 bits to 16 bits, and related prior or posterior is often complexly intertwined with the data. In collecting, the data distribution can also vary significantly due to various conditions, such as the specificity of the institute collected and the period collected. Therefore, it can be said that the research dealing with feature distribution is essential in the medical domain.

Metric learning for discriminative features

Metric learning, based on the distance between each feature, has been mainly used for face

recognition. Most metric learning studies, such as SphereFace [29], Additive Margin Softmax [30], and ArcFace [1], have been adapted softmax and used angle distance mainly on a hypersphere. Also, margin loss, which maximizes inter-class margin and minimizes intra-class variation, has been used in metric learning. For example, ArcFace [1] changes individual weights and embedding features in softmax to 1 by L2 normalization. Then, the product of the weights and features could be regarded as a radius of the hypersphere and softmax could be controlled by just an angle between the weights and features. Additionally, an additive angular margin penalty was added to the angle. Through these steps, the features could be induced to be more discriminative.

These metric learning methods usually have yielded good performance in face recognition which has long-tail datasets, but they are difficult to optimize the model. The process of fine-tuning hyperparameters and model architecture is needed to ensure stable training.

Self-supervised learning

Labeling data usually needs lots of time and cost. Strong label, such as bounding box for detection and pixel-level label for segmentation, requires more time and cost than weak label. Specialized fields, such as the medical domain, also require more time and cost because experts are needed. Researches about efficient usage of data have been presented to reduce this time and cost: 'weakly-supervised learning' that generates strong label using weak label only, 'semi-supervised learning' that trains the model with labeled data and unlabeled data, and 'self-supervised learning' that extracts data-specific features using unlabeled data only.

Among them, self-supervised learning, which is used in this study, has been shown remarkable results. Self-supervised learning can be divided into two stages: 'pretext task' that is set to train the model to extract data-specific features and 'downstream task' that trains the model to apply real task and evaluates performance.

Early pretext task researches were conducted to find semantic features without any label in the image, like Exemplar [34], Context Prediction [35], Jigsaw Puzzle [36], Colorization [37], and Rotation [38]. Starting with CPC ([39], [40]), contrastive learning, which trains the model for the instance discrimination using positive and negative pairs, has been introduced, such as MoCo ([41], [42], [43]), SimCLR ([44], [45]), and SwAV [46]. Also, the researches about

pixel-level contrastive tasks, such as DenseCL [47], PixPro [48], SCRL [49], VADeR [50], and DetCon [51], have been studied to effectively obtain improved performance of pixel-level prediction, like detection or segmentation. However, negative pairs in contrastive learning cause a large batch size for training the model. Recently, researches that do not use negative pairs, such as BYOL [52], SimSiam [53], and Barlow Twins [54], have been presented.

For downstream task, four tasks usually have been used to evaluate the encoder trained by pretext task: linear evaluation, fine-tuning, transfer learning, and retrieval. Linear evaluation only trains the fully-connected layer with the backbone network fixed. Fine-tuning trains all networks, including the backbone network. Transfer learning in self-supervised learning is used to train other domains. For example, if the backbone network is trained with ImageNet dataset, it is tested in COCO dataset. Last, retrieval is used by calculating recall among clustering, such as K-Nearest Neighbor.

Objectives

In this study, two experiments were conducted to check how representation learning, especially metric learning and self-supervised learning, affects medical images using cephalogram: 'orthodontic diagnosis with cephalogram' and 'the effect of self-supervised learning on orthodontic diagnosis'.

In the first study, three orthodontic diagnoses were conducted: anteroposterior skeletal discrepancies (APSD: Class I, Class II, and Class III), vertical skeletal discrepancies (VSD: normo-divergent, hyper-divergent, and hypo-divergent), and vertical dental discrepancies (VDD: normal overbite, open bite, and deep bite). All of the diagnoses can be classified using the cephalometric parameters obtained from the cephalometric landmarks, and 'the gray zone' between individual diagnoses was formed due to the continuous parameters. Also, there are there was a class imbalance problem due to the data collected from 10 dental institutes. To solve these problems, ArcFace [1], which induces discriminative features, was added to the existing model. In addition, Group Normalization [2], which shows good performance regardless of batch size, was used instead of Batch Normalization [3] for small batch training to use a small number of training data effectively. As a result, the proposed model has consistently shown good performance in internal validation and external validation.

In the second study, pretext task was conducted using SimSiam [5], one of the self-supervised learning models, and downstream task was conducted in APSD. For comparison, randomly initialized weights and weights pre-trained on ImageNet dataset were used. As a result of the linear evaluation and fine-tuning, SimSiam showed better performance in full and low data regimes and did not induce overfitting compared to training from weights pre-trained on ImageNet dataset.

Both studies confirmed that metric learning and self-supervised learning in medical images could improve performance, extract discriminative features, and train models that are robust to data distribution and the number of data. Therefore, in the future medical image AI, research that incorporates representation learning should be conducted rather than simply evaluating performance by learning by model.

Orthodontic Diagnosis using Metric Learning

1. Dataset

1.1. Data description

Figure 1 shows the flowchart of a dataset and experimental setup. The cephalogram dataset for orthodontic diagnosis from 10 the Department of Orthodontics in 10 multi-centers, including Seoul National University Hospital (SNUDH), Kooalldam Dental Hospital (KADH), Ajou University Dental Hospital (AJUDH), Asan Medical Center (AMC), Chonnam National University Dental Hospital (CNUDH), Chosun University Dental Hospital (CSUDH), Ewha University Medical Center (EUMC), Kyung Hee University Dental Hospital (KHUDH), Kyungpook National University Dental Hospital (KNUDH), and Wonkwang University Dental Hospital (WKUDH) in South Korea was constructed. The data of Korean adult patients who underwent orthodontic treatment and/or orthognathic surgery between 2013 and 2020, except patients who were in childhood and adolescent period and had mixed dentition, was collected. All datasets were strictly anonymized before utilized. Table 1 shows the information on the product, radiation exposure condition, sensor, and image condition of the cephalometric radiograph system in each center, which showed a diversity of conditions.

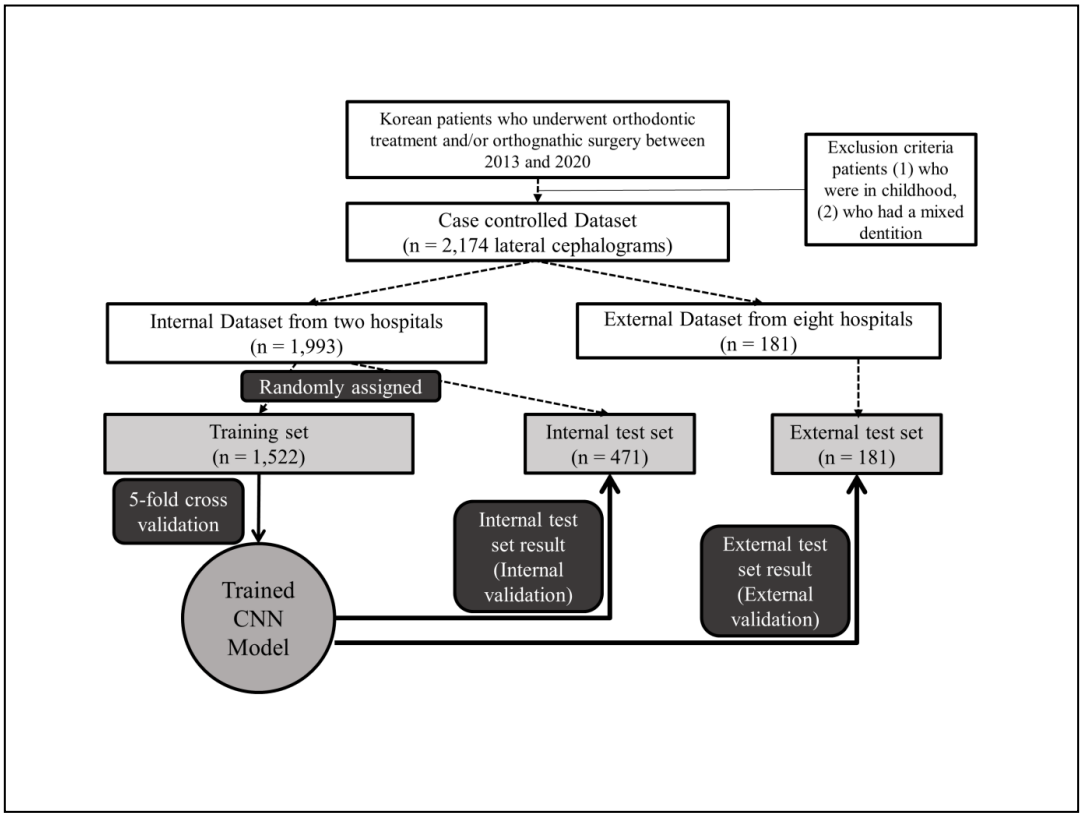

Figure 1. Flowchart of dataset and experimental setup.

Table 1. Information on the product, radiation exposure condition, sensor, and image condition of the cephalometric radiograph system in 10 multi-centers.

		SNUDH	KADH	AJUDH	AMC	CNUDH	CSUDH	EUMC	KHUDH	KNUDH	WKUDH
Product	Company	Asahi	Vatech	Planmeca	Carestream	Instrumentarium	Planmeca	Asahi	Asahi	Asahi	Planmeca
	Model	CX-90SP-II	Uni3D NC	Proline XC	CS9300	OrthoCeph OC 100	Proline XC	Ortho stage (Auto III N CM)	CX-90SP	CX-90SP-II	Promax
Radiation exposure condition	Kvp	76 kVp	85 kVp	68 kVp	80 kVp	85 kVp	80 kVp	75 kVp	70 kVp	70 kVp	Female 72 kVp, Male 74 kVp
	mA	80 mA	10mA	7 mA	12 mA	12 mA	12 mA	15 mA	15 mA	80 mA	10 mA
	sec	0.32 sec	0.9 sec	2.3 sec	0.63 sec	1.6 sec	1.8 sec	1 sec	0.3–0.35 sec	0.32 sec	1.87 sec
Sensor	image sensor	Cassette (CR system)	CCD sensor	CCD sensor	CCD sensor	Cassette (CR system)	Cassette (CR system)	Cassette (CR system)	Cassette (CR system)	Cassette (CR system)	CCD sensor
	sensor size	10 x 12 (inch)	30 x 25 (cm)	10.6 x 8.85 (inch)	30 x 30 (cm)	10 x 12 (inch)	8 x 10 (inch)	8 x 12 (inch)	10 x 12 (inch)	11 x 14 (inch)	27 x 30 (cm)
Image	Image size (pixel x pixel)	2000 x 2510 / 2010 x 1670	2360 x 1880	1039 x 1200	2045 x 2272 / 1012 x 2020	2500 x 2048	2392 x 1792 / various	2510 x 2000	2500 x 2048	1950 x 2460 / 2108 x 1752	1818 x 2272
	Actual resolution (mm/pixel)	0.150 / 0.100	0.110	0.250	0.132 / 0.145	0.115	0.100	0.100	0.110	0.100	0.132
Lateral cephalogram images used in this study (number)		1,129	864	22	21	20	30	26	23	19	20

1.2. Diagnosis criteria description

For the orthodontic diagnosis, three orthodontic diagnoses were used: anteroposterior skeletal discrepancies (APSD: Class I, Class II, and Class III), vertical skeletal discrepancies (VSD: normo-divergent, hyper-divergent, and hypo-divergent), and vertical dental discrepancies (VDD: normal overbite, open bite, and deep bite). The cephalometric parameters, which can be obtained through cephalometric landmarks, were used to classify each diagnosis: A point-Nasion-B point angle (ANB) [6] for APSD, Frankfort mandibular plane angle (FMA) [7] and Jarabak's posterior/anterior facial height ratio (FHR) [8] for VSD, and overbite for VDD. A single orthodontic specialist detected the cephalometric landmarks, such as A point, Nasion, B point, Orbitale, Porion, Gonion, Menton, Sella, Mandible 1 crown, Mandible 6 distal, Maxilla 1 crown, and Mandible 6 crown and calculated these cephalometric parameters using the V-Ceph 8.0 program (Osstem, Seoul, Korea). After these processes, the specialist reexamined the calculated parameters by his opinion.

As a golden standard, all of the cephalograms were classified into the three diagnosis groups by the specialist as follows: (1) For classification of APSD, the ANB value was defined within one standard deviation (SD) from the ethnic norm and sex as skeletal Class I; over one SD as skeletal Class II; and under one SD as skeletal Class III. (2) For classification of VSD, FMA and FHR were combined for training of the VSD. First, the FMA and FHR values were normalized by using the SD values. Second, the FHR values were flipped due to an opposite sign compared to the FMA values. Third, the values of FMA and flipped FHR were added because each is regarded as having equal weights. Fourth, the mean and SD value was obtained to classify into three groups. Then, the values were defined within one SD from the mean as a normo-divergent; over one SD as a hyper-divergent; and under one SD as a hypo-divergent. And (3) For classification of VDD, the overbite value was defined between 0 mm and 3 mm as a normal overbite, over 3 mm as a deep bite, and under 0 mm as an open bite. Table 2 shows classification criteria for the APSD, VSD, and VDD for orthodontic analysis. [9]

Table 2. Classification criteria for the anteroposterior skeletal discrepancy (APSD), vertical skeletal discrepancy (VSD), and vertical dental discrepancy (VDD) for orthodontic analysis.

Sex	APSD		VSD				VDD	
	ANB		FMA		FHR		overbite	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Female	2.4	1.8	24.2	4.6	65	9	1.5	1.5
Male	1.78	2.02	26.78	1.79	66.37	5.07		

2. Model Architecture

In this study, there are three key points to perform orthodontic diagnosis: backbone network, metric learning for discriminative features, and normalization layer.

2.1. Backbone network

Many backbone networks had achieved the state-of-the-art in ImageNet datasets. The early models, such as LeNet [10], AlexNet [11], VGG [12], Inception [13], ResNet [14], DenseNet [15], and SENet [16], were made by generalizing particular network instantiations and design principles and applying them to numerous settings. Recently, the models based on neural architecture search (NAS) ([17], [18]) have shown better performance than earlier's, but there are no network design principles. EfficientNet [19], based on NAS and adjusted using compound scaling, has shown good performance and suggested optimized network design using model scaling. RegNet [20], which quantified the quality of a design space by sampling a set of models from that design space and characterizing the resulting model error distribution, has also shown state-of-the-art performance and presented a new network design paradigm.

To select the most effective backbone network, some backbone networks were trained, not including RegNet, with pre-trained weights for ImageNet dataset and default settings. As a result, DenseNet-169 [15] was determined as our backbone network because it showed the best performance with our datasets.

2.2. Metric learning for discriminative features

As shown in Figure 2, all datasets for orthodontic diagnosis have overlapping parts near the red lines pointing to each diagnosis criteria. It means these parts are challenging to clarify in two parts. In this study, ArcFace [1] was added to the last convolutional layer of the backbone in parallel with softmax layer during training to overcome this problem. For stable training, dropout and regularization were excluded, unlike [1]. After training, ArcFace head was removed and inference was implemented using only softmax layer like basic classifiers using the backbones.

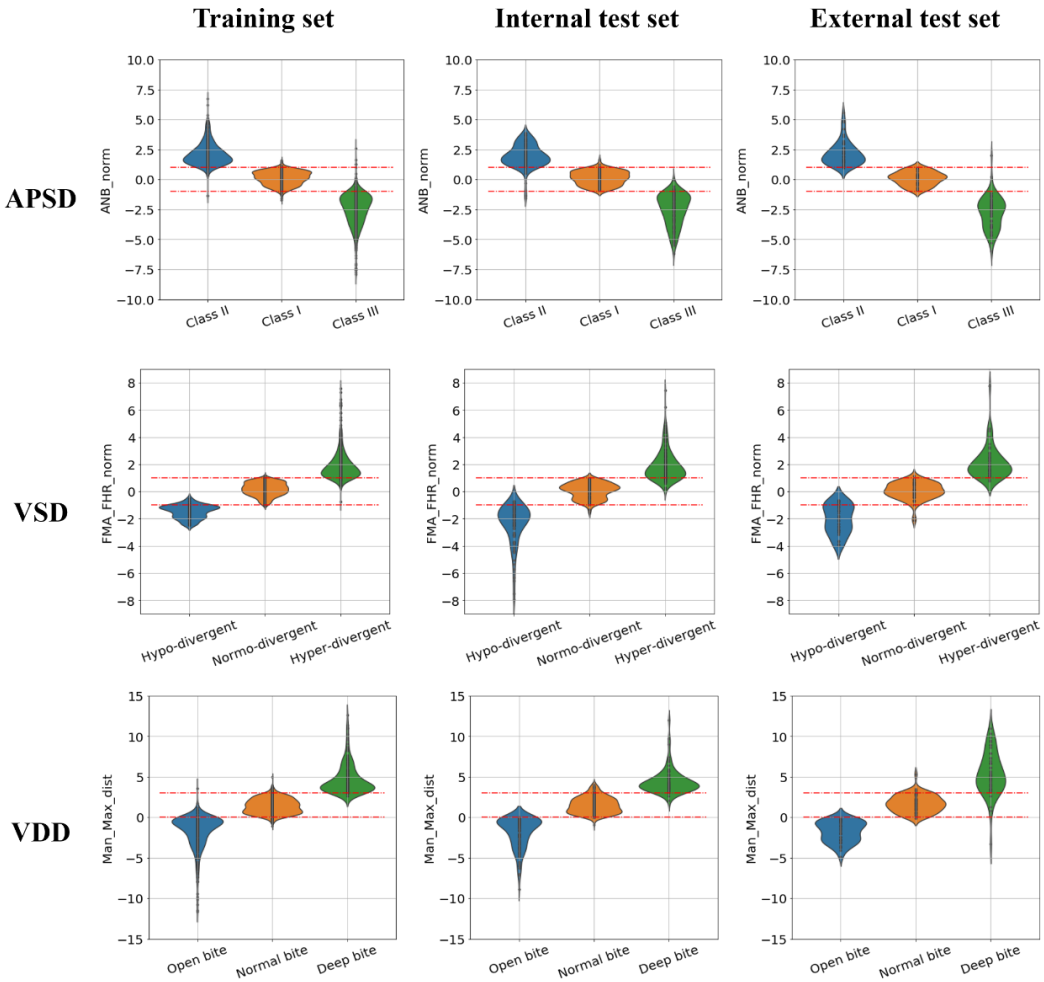

Figure 2. The cephalometric parameters' distribution of the APSD, VSD, and VDD per each dataset. Red lines in the APSD and VSD indicate one SD of the normal classes addressed in Table 2. Red lines in the VDD indicate the boundary values, which were 0 mm and 3 mm.

2.3. Normalization layer

Large batch size is usually used to obtain an accurate estimate of the gradient and finish training in a few hours. However, large batch training needs high computation cost and can cause low generalization performance when small size dataset is used. [21] and [22] said that using small batch size in training can improve training stability and generalization performance.

Batch Normalization (BN) [3], which is known to relieve internal covariance shift through batch-wise normalization and re-normalization and is usually used in almost backbone network, causes inaccurate batch statistics estimation with small batch size. Because of this issue, Group Normalization (GN) [2], which has consistent performance regardless of batch size, was presented.

In this study, small batch training regime was used to utilize small size dataset and improve training stability and generalization performance, and GN was used for this regime instead of BN.

2.4. Summary

Figure 3 shows that the model was set according to each state. Because the criterion of APSD and VSD is affected by patient's sex, the model which diagnoses APSD or VSD was concatenated with a one-hot vector of sex after global average pooling (GAP) layer and the model of VDD was not.

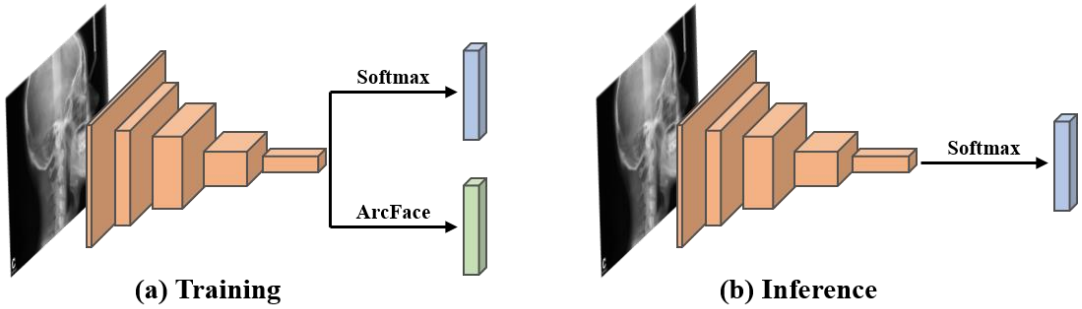

Figure 3. Diagrams of the model architecture. (a) During training, ArcFace head was added to the last convolutional layer of the backbone in parallel with softmax layer. (b) After training, ArcFace head was removed and inference was implemented using only softmax layer.

3. Experiments

3.1. Data augmentation

Many studies have used data augmentation techniques to solve the lack of data numbers in the real world. The techniques can expand the variance of data, making the model robust even when data is scarce. In addition to basic techniques such as rotation, crop, and flip, advanced techniques such as AutoAugment [23], RandAugment [24], and CTAugment [25] have recently been used for natural images. However, detailed techniques should be considered when the dataset is in the medical domain. In this study, the augmentation techniques which preserve essential features of cephalograms were used. Affine transformations, such as rotation, cropping, and shifting, were performed while maintaining the aspect ratio of each image. Especially, cropping which preserves orthodontic landmarks was used. The techniques that change the pixel level of cephalograms, such as contrast limited adaptive histogram equalization, gamma correction, and gaussian noise, were applied after the affine transformations.

3.2. Experimental details

This orthodontic diagnosis study was progressed with two questions: “Does metric learning really show better performance and have more discriminative features than other backbone networks?” and “Does metric learning show better performance in other tasks and external validation?”.

For the first question, a small dataset (n=905) was constructed from nine hospitals (SNUDH, AJUDH, AMC, CNUDH, CSUDH, EUMC, KHUDH, KNUDH, and WKUDH). The dataset which belongs to the whole dataset described below was divided for training (n=580), validation (n=146), and test (n=179). Figure 4 shows the distribution of this small dataset. To evaluate our proposed method, the first experiment was implemented to compare the

performance of three options: DenseNet-169, DenseNet-169 + GN, and DenseNet-169 + GN + ArcFace. The evaluation was conducted for diagnosing APSD only, and the environment was as described below except for 5-fold cross-validation.

For the second question, as shown in Table 3, 1,993 lateral cephalogram images from two hospitals (SNUDH and KADH) were used for trainset (n=1,522) and internal testset (n=471), and 181 images from eight other hospitals (AJUDH, AMC, CNUDH, CSUDH, EUMC, KHUDH, KNUDH, and WKUDH) were used as the external testset.

The image size was set to 224 x 224, which was the default setting of DenseNet-169, and the number of groups of GN was 2. ArcFace head was set to the default settings of [1] except for dropout and regularization for training stability. 5-fold cross-validation was used in the trainset and each result was described as the mean and SD. Training for APSD, VSD, and VDD was performed with a golden standard determined by a single orthodontic specialist and not the cephalometric parameters, including ANB, FMA, FHR, and overbite. The training condition was the fixed seed and 4 batch size during 100 epochs. The learning rate was 0.0001 with warm-up for the first 5 epochs and decay at 50 epochs by multiplying it by 0.1. Cross-entropy was used as an objective function.

After training was completed, one-step classification was performed with the internal testset and external testset set to validate the performance of the constructed model. The results for the internal testset and external testset were compared with golden standard diagnosis data.

First, our proposed model was evaluated to be suitable for this orthodontic diagnosis task in 4.1. Then, the model was evaluated in an extended environment, which consisted of the whole dataset, other orthodontic diagnoses (VSD and VDD), and internal and external evaluation in 4.2, 4.3, and 4.4.

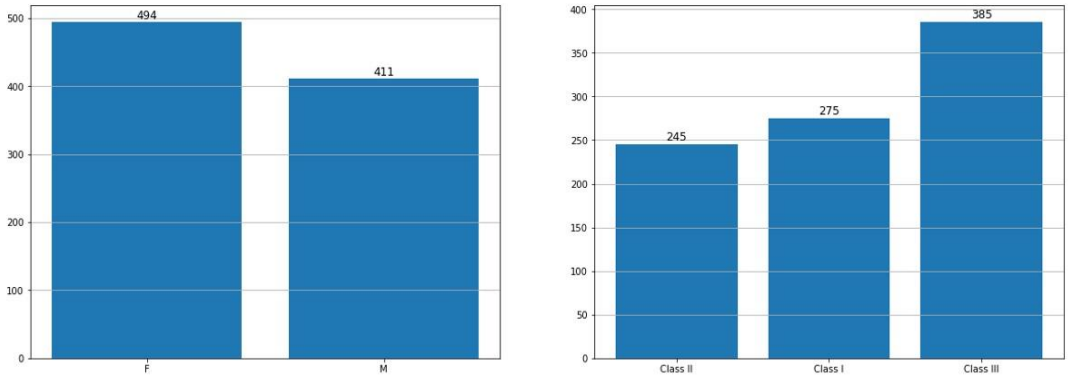

Figure 4. Distribution of small dataset to evaluate our proposed model. (a) gender (female and male) and (b) the label of APSD (Class II, Class I, and Class III).

Table 3. Distribution of classification groups in each diagnosis for human golden standard in trainset, internal testset, and external testset.

		Training set			Internal test set			External test set									sum	
		SNUDH	KADH	sum	SNUDH	KADH	sum	AJUDH	AMC	EUMC	CNUDH	CSUDH	KHUDH	KNUDH	WKUDH	sum	Int + Ext test sets	Total
APSD	Class I	238	323	561 (36.9%)	122	40	162 (34.4%)	8	6	5	4	7	11	7	2	50 (27.6%)	212 (32.5%)	773 (35.6%)
	Class II	183	263	446 (29.3%)	112	44	156 (33.1%)	8	8	11	8	13	4	4	6	62 (34.3%)	218 (33.4%)	664 (30.5%)
	Class III	359	156	515 (33.8%)	115	38	153 (32.5%)	6	7	10	8	10	8	8	12	69 (38.1%)	222 (34.0%)	737 (33.9%)
	Sum	780	742	1,522	349	122	471	22	21	26	20	30	23	19	20	181	652	2,174
VSD	Normo-divergent	331	389	720 (47.3%)	146	50	196 (41.6%)	10	6	7	9	17	10	7	7	73 (40.3%)	270 (41.4%)	989 (45.5%)
	Hyper-divergent	314	241	555 (36.5%)	135	40	175 (37.2%)	5	9	12	6	3	7	8	6	56 (30.9%)	231 (35.4%)	786 (36.2%)
	Hypo-divergent	135	112	247 (16.2%)	68	32	100 (21.2%)	7	6	7	5	10	6	4	7	52 (28.7%)	151 (23.2%)	399 (18.4%)
	Sum	780	742	1,522	349	122	471	22	21	26	20	30	23	19	20	181	652	2,174
VDD	Normal overbite	440	493	933 (61.3%)	196	53	249 (52.9%)	11	11	10	8	9	10	10	10	79 (43.6%)	328 (50.3%)	1,261 (58.0%)
	Open bite	209	194	403 (26.5%)	99	41	140 (29.7%)	4	7	9	5	9	8	4	5	51 (28.2%)	191 (29.3%)	594 (27.3%)
	Deep bite	131	55	186 (12.2%)	54	28	82 (17.4%)	7	3	7	7	12	5	5	5	51 (28.2%)	133 (20.4%)	319 (14.7%)
	Sum	780	742	1,522	349	122	471	22	21	26	20	30	23	19	20	181	652	2,174

3.3. Analysis methods

Three methods were used to analyze experimental results: Receiver Operating Characteristic (ROC) analysis [31], t-Stochastic Neighbor Embedding (t-SNE) [32], and Gradient-weighted Class Activation Mapping (Grad-CAM) [33].

ROC curve is created by plotting the true positive rate (TPR, sensitivity) against the false positive rate (FPR, 1-specificity) at various threshold settings and shows the ability of a binary classifier when its discrimination threshold is varied. ROC analysis is related to cost/benefit analysis of decision making and provides tools to select optimal models through ROC curve. In this study, ROC analysis was used to evaluate the performance of our model using accuracy, area under the curve (AUC), sensitivity, and specificity. To calculate the sensitivity and specificity, we assumed binary classification tasks for each diagnosis. For example, APSD can be classified as one of three cases, such as Class I and the others (Class II and Class III), Class II and the others (Class I and Class III), and Class III and the others (Class I and Class II). Then, sensitivity and specificity for each case were calculated. Because 5-fold cross-validation was used in this study, the ROC curves of five models were obtained and applied using the mean and SD at each specificity.

t-SNE is a technique that visualizes high-dimensional data by giving each datapoint a location in a two or three-dimensional map. t-SNE uses a Student-t distribution to compute the similarity between two points in the low-dimensional space. In this study, t-SNE was used to check the feature distribution of trainset, internal testset, and external testset after GAP layer. For each diagnosis of APSD, VSD, and VDD, the labels of ground truth and prediction were set to check the distribution per each dataset.

Grad-CAM is a technique for producing visual explanations which uses the gradients of any target concept, flowing into the final convolutional layer to obtain a coarse localization map focusing important regions in the image for predicting the concept. In this study, Grad-CAM was used to confirm the regions where our model mainly focused on diagnosing APSD, VSD, and VDD.

4. Results

4.1. Evaluation of our proposed model

Figure 5 and Table 4 show that our proposed model (DenseNet-169 + GN + ArcFace) has better performance than other models (DenseNet-169 and DenseNet-169 + GN). Also, Figure 6 shows that our proposed model generates more discriminative features than others through the results of t-SNE. Through this sub-experiment, we could know that our proposed model, DenseNet-169 + GN + ArcFace, was optimized at orthodontic diagnosis task.

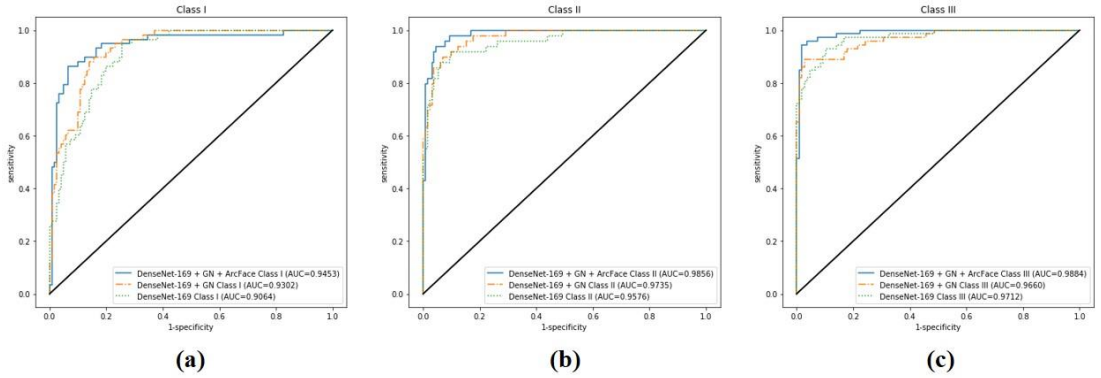

Figure 5. The results of the ROC curve with AUC per class of each model. (a) Class I, (b) Class II, and (c) Class III.

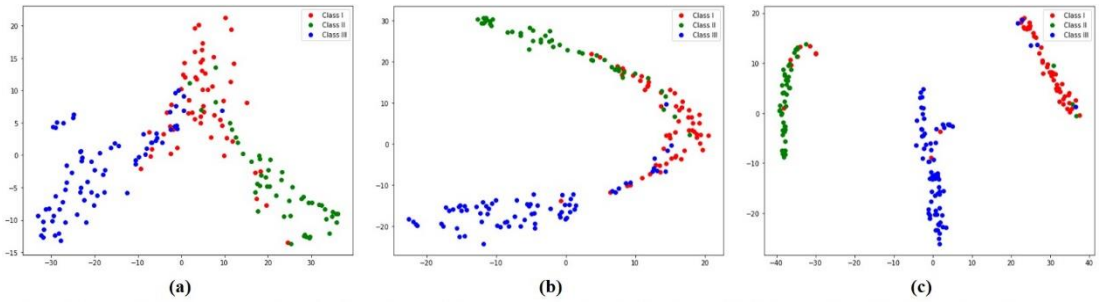

Figure 6. The results of t-SNE of the small dataset about each model. (a) DenseNet-169, (b) DenseNet-169 + GN, and (c) DenseNet-169 + GN + ArcFace.

Table 4. Accuracy, AUC, sensitivity, and specificity of each model in the first experiment.

	DenseNet-169			DenseNet-169 + GN			DenseNet-169 + GN + ArcFace		
	Class I	Class II	Class III	Class I	Class II	Class III	Class I	Class II	Class III
Accuracy	0.8268	0.9162	0.9106	0.8547	0.9274	0.9274	**0.8939**	**0.9330**	**0.9609**
AUC	0.9064	0.9576	0.9712	0.9302	0.9735	0.9660	**0.9453**	**0.9856**	**0.9884**
Sensitivity	0.7586	0.8571	0.8472	**0.8448**	0.8571	0.8611	0.8103	**0.9388**	**0.9306**
Specificity	0.8512	0.9462	0.9533	0.8595	**0.9538**	0.9720	**0.9338**	0.9308	**0.9813**

4.2. The accuracy and AUC in ROC analysis

Figure 7 and Figure 8 show the ROC curve in the internal testset and external testset for diagnosis of the APSD, VSD, and VDD, respectively. Table **5** shows the performance of our model in the internal testset and external testset using the binary ROC analysis.

4.2.1. The accuracy and AUC of the internal testset in ROC analysis

In APSD, Class III was highest (0.9372 and 0.9807), followed by Class II (0.8972 and 0.9533) and Class I (0.8488 and 0.9212). In VSD, hypo-divergent was highest (0.9346 and 0.9824), followed by hyper-divergent (0.9019 and 0.9730) and normo-divergent (0.8365 and 0.9186). In VDD, open bite was highest (0.8730 and 0.9475), followed by deep bite (0.8637 and 0.9286) and normal overbite (0.7376 and 0.8177).

In APSD and VSD, the total accuracy reached nearly 0.9 (0.8944 and 0.8910) and the total AUC exceeded 0.95 (0.9517 and 0.9580). However, VDD showed a relatively lower total accuracy (0.8248) and the total AUC (0.8979) than APSD and VSD.

4.2.2. The accuracy and AUC of the external testset in ROC analysis

In APSD, Class III was highest (0.9252 and 0.9930), followed by Class II (0.8796 and 0.9601) and Class I (0.8320 and 0.9042). In VDD, open bite was highest (0.8917 and 0.9626), followed by deep bite (0.8586 and 0.9238) and normal overbite (0.7591 and 0.8359). However, VSD

showed a different pattern between the accuracy and AUC. Although the accuracy was highest for hypo-divergent (0.9094), followed by hyper-divergent (0.9061) and normo-divergent (0.8309), the AUC was highest for hyper-divergent (0.9730), followed by hypo-divergent (0.9684) and normo-divergent (0.9157).

In APSD and VSD, total accuracy reached nearly 0.9 (0.8880 and 0.8821) and total AUC exceeded 0.95 (0.9524 and 0.9523). However, VDD showed a relatively lower total accuracy (0.8365) and total AUC (0.9074) than APSD and VSD.

4.2.3. Comparison of the AUC values between the internal testset and external testset in the binary ROC analysis

In APSD and VSD, Class III and open bite showed the highest AUC compared to other classifications (0.9807 and 0.9903 in the internal testset, 0.9475 and 0.9626 in the external testset). However, the VSD showed a different pattern. The internal testset showed the highest AUC for hypo-divergent pattern (0.9824), while the external testset showed the highest AUC for hyper-divergent pattern (0.9730). However, the difference in the AUC values was less than 0.01.

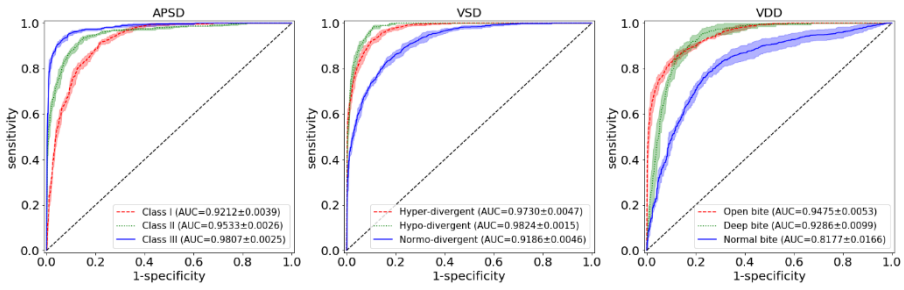

Figure 7. The results of the ROC curve in the internal testset from two hospitals for diagnosis of the APSD, VSD, and VDD.

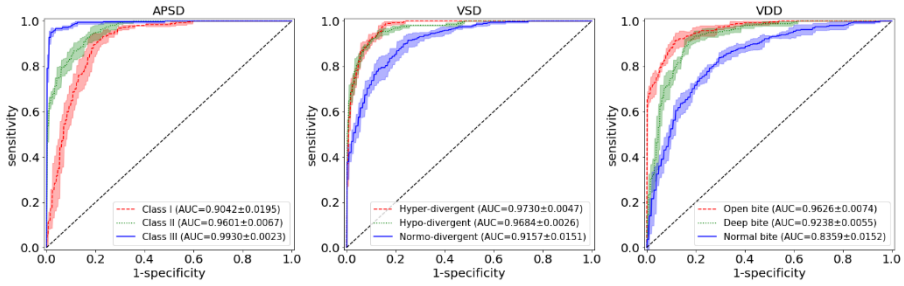

Figure 8. The results of the ROC curve in the external testset from other eight hospitals for diagnosis of the APSD, VSD, and VDD.

Table 5. Performance of our model for the diagnosis of the APSD, VSD, and VDD in the internal test set and external test set using the binary ROC analysis.

		Accuracy				AUC				Sensitivity				Specificity			
		Internal test set		External test set		Internal test set		External test set		Internal test set		External test set		Internal test set		External test set	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
APSD	Class I	0.8488	0.0103	0.8320	0.0230	0.9212	0.0038	0.9042	0.0195	0.7938	0.0328	0.7840	0.0297	0.8764	0.0186	0.8504	0.0273
	Class II	0.8972	0.0057	0.8796	0.0153	0.9533	0.0026	0.9601	0.0067	0.8192	0.0334	0.7226	0.0515	0.9359	0.0161	0.9613	0.0046
	Class III	0.9372	0.0063	0.9525	0.0108	0.9807	0.0025	0.9930	0.0023	0.9111	0.0225	0.9652	0.0079	0.9497	0.0086	0.9446	0.0160
	Mean	0.8944	0.0368	0.8880	0.0518	0.9517	0.0245	0.9524	0.0382	0.8414	0.0571	0.8239	0.1076	0.9206	0.0345	0.9188	0.0516
VSD	Normo-divergent	0.8365	0.0082	0.8309	0.0267	0.9186	0.0046	0.9157	0.0151	0.8235	0.0279	0.7699	0.0416	0.8458	0.0122	0.8722	0.0178
	Hyper-divergent	0.9019	0.0035	0.9061	0.0203	0.9730	0.0047	0.9730	0.0047	0.8149	0.0273	0.9143	0.0293	0.9534	0.0190	0.9024	0.0360
	Hypo-divergent	0.9346	0.0098	0.9094	0.0164	0.9824	0.0015	0.9684	0.0026	0.9000	0.0394	0.8000	0.0661	0.9445	0.0127	0.9535	0.0110
	Mean	0.8910	0.0413	0.8821	0.0410	0.9580	0.0283	0.9523	0.0273	0.8461	0.0478	0.8280	0.0757	0.9146	0.0505	0.9094	0.0398
VDD	Normal overbite	0.7376	0.0291	0.7591	0.0230	0.8177	0.0166	0.8359	0.0152	0.6530	0.0956	0.6582	0.0664	0.8288	0.0441	0.8373	0.0557
	Open bite	0.8730	0.0130	0.8917	0.0139	0.9475	0.0053	0.9626	0.0074	0.8371	0.0366	0.8275	0.0611	0.8882	0.0304	0.9262	0.0228
	Deep bite	0.8637	0.0270	0.8586	0.0127	0.9286	0.0099	0.9238	0.0055	0.8000	0.1100	0.8196	0.0836	0.8781	0.0530	0.8723	0.0457
	Mean	0.8248	0.0654	0.8365	0.0584	0.8979	0.0582	0.9074	0.0538	0.7634	0.1111	0.7684	0.1006	0.8651	0.0468	0.8786	0.0535

4.3. t-SNE of the APSD, VSD, and VDD per dataset

In Figure 9, the ground truth in the trainset, internal testset, and external testset showed that dots with different colors mixed irregularly in the classification cutoff areas between the normal group (Class I in the APSD, normo-divergent in the VSD, and normal overbite in the VDD) and the other two groups (Class II and III in the APSD, hyper-divergent and hypo-divergent in the VSD, and open bite and deep bite in the VDD).

However, the AI prediction did not show an overlapped area between the normal group and the other two groups in the trainset, internal testset, and external testset. Therefore, it indicated that our model succeeded in creating a well-separation between the three classification groups for each diagnosis, resulting in the consistent classification of the three groups.

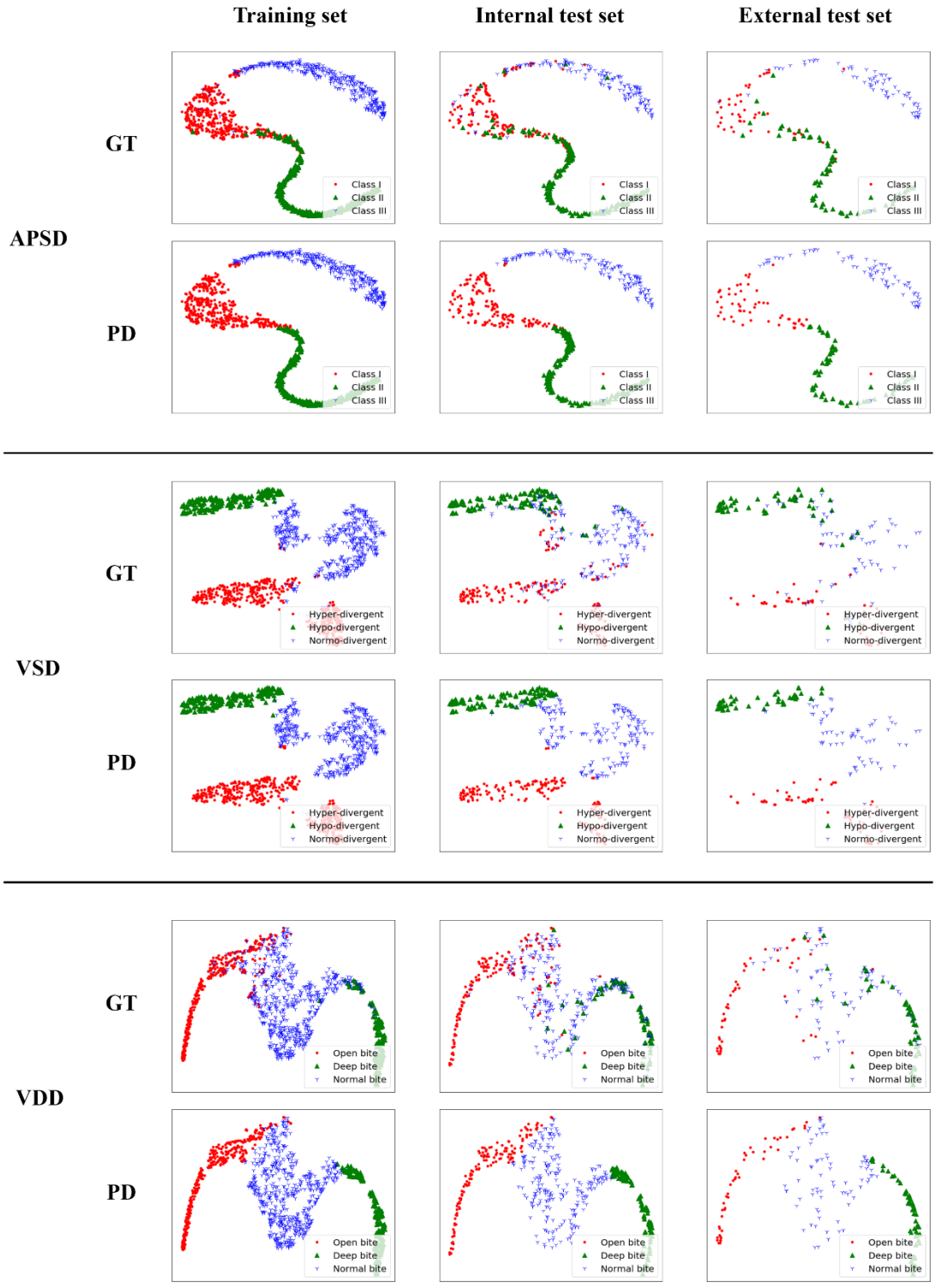

Figure 9. The results of t-SNE in the APSD, VSD, and VDD per each dataset. The labels of ground truth (GT) and prediction (PD) were set to check their distribution.

4.4. Grad-CAM for each diagnosis

In Figure 10, heatmaps showed differences in the location and size of the focus areas according to the three classification groups for each diagnosis. It indicated that our model could effectively use the information in the lateral cephalogram images.

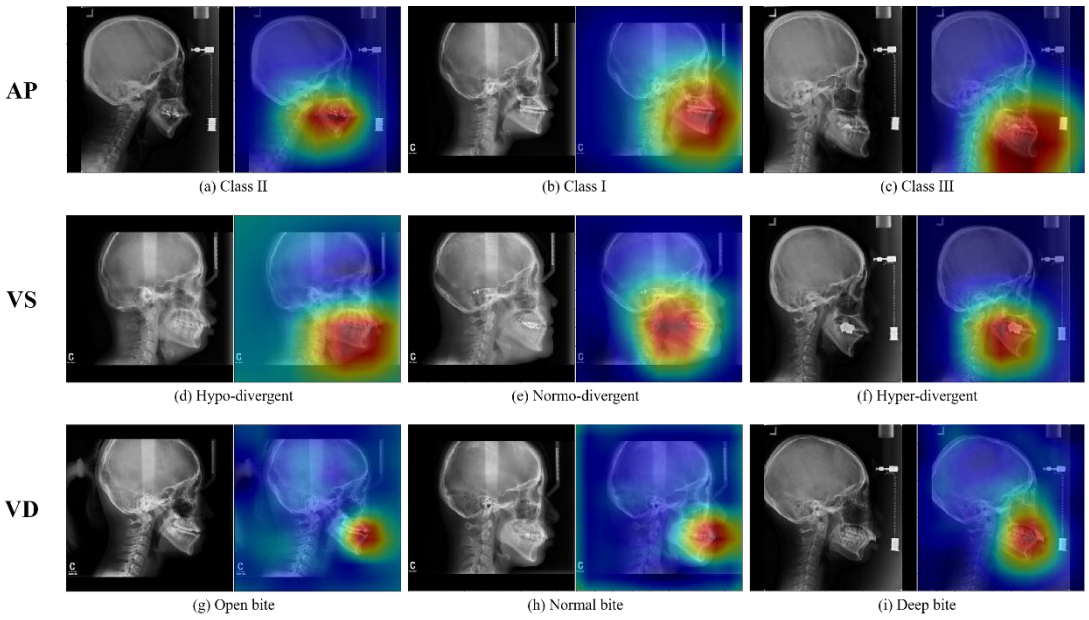

Figure 10. The results of Grad-CAM for the APSD, VSD, and VDD.

5. Discussions

The present study has some meaningful outcomes as follows: (1) Despite the different quality of lateral cephalograms from diverse condition of cephalometric radiograph system in 10 multi-centers (Table 1), clinically acceptable accuracy of diagnosis for APSD, VSD, and VDD was obtained; and (2) Since it was possible to give diagnosis for APSD, VSD, and VDD only with the input of lateral cephalograms, our model can be regarded as a general-purpose one-step orthodontic diagnosis model.

5.1. Clinical meaning of the comparison results between the internal and external

testsets in binary ROC analysis

In the results of binary ROC analysis, the amounts of difference in the mean AUC values between internal and external testsets were less than 0.01 (APSD, Δ=0.0007; VSD, Δ=-0.0057; VDD, Δ=0.0095) in Table 5. Since these amounts of difference in the AUC values for APSD, VSD, and VDD in binary ROC analysis were almost insignificant, the reliability of our model was well validated in the external testset. Our model can be regarded as a reliable general-purpose one-step orthodontic diagnosis model.

5.2. Difference in the AUC values of Class II and Class III groups in APSD and hyper- and hypo-divergent groups in VSD in binary ROC analysis

The hypo-divergent group showed a higher AUC score than the hyper-divergent group in the internal testset; while the hyper-divergent group showed the highest value for the AUC than the hypo-divergent group in the external testset (0.9824 vs. 0.9730 in the internal testset; 0.9684 vs. 0.9730 in the external testset) in Table 5. However, the Class III group showed higher AUC values than the Class II group both internal and external testsets (0.9807 vs. 0.9533 in the internal testset; 0.9930 vs. 0.9601 in the external testset) in Table 5. The reason might be a difference in the location and size of the focus areas in the diagnosis of VSD and APSD (i.e., relatively larger difference between Class II and Class III groups compared to those between the hyper-divergent and hypo-divergent groups in Figure 9). Further studies are necessary to investigate the reason why the Class III group showed a higher AUC score than the Class II group.

5.3. Summary

Through this study, our proposed model with metric learning has shown better performance than other baselines and consistent performance in three types of diagnosis using data with various qualities from multi-centers. The reason of these results is that discriminative features could be obtained due to training with metric learning, ArcFace, and sharp dividing lines between these diagnoses could be formed.

However, a question of whether data consisting of continuous variables, such as ANB, FMA, FHR, and overbite, should be clearly divided still remains in orthodontic diagnosis, as various

variables must be considered in addition to those used in this study. To effectively use a narrow range of label data and obtain meaningful features, new methods using unlabeled data or small labeled data, such as self-supervised learning and semi-supervised learning, should be studied.

Application of Self-Supervised Learning to Orthodontic Diagnosis

1. Dataset

The dataset, which is defined in orthodontic diagnosis task from 10 multi-centers (n=2,174), was used for labeled dataset. In addition, a dataset from KAD (n=15,833) was used for unlabeled dataset. Likewise, all of the unlabeled data were strictly anonymized before utilized.

2. Model Architecture

2.1. Self-supervised learning architecture to obtain data-specific features

Most architectures for pretext task shows good performance when they are trained with large batch size regime in large datasets, such as ImageNet, JFT-300M, and Instagram-1B dataset. In this study, however, there was a relatively small dataset with about 18K samples and large batch size could not be used due to poor generalization. Because of this problem, SimSiam [53], which has simple Siamese architecture and shows good performance even with small batch size, was used for pretext task with DenseNet-121 [15] for a backbone network.

2.2. Model settings for downstream task

In this study, linear evaluation and fine-tuning were used to evaluate our backbone network trained with SimSiam (SimSiam). Also, stress tests using different ratios of training data, such as 25%, 50%, 75%, and 100%, were conducted to evaluate how the data ratios affect models' performance. As a control, the backbone networks with pre-trained weights on ImageNet dataset (ImageNet) and randomly initialized weights (Scratch) were used to compare their performance.

3. Experiments

3.1. Experimental details for pretext task

Our SimSiam architecture was set to the default setting of [53]. Batch normalization was

applied to all convolutions of projection and prediction, not to prediction output. The dimension of the projection was 2048, the bottleneck dimension of the prediction was 512, and the output dimension was 2048.

In self-supervised learning, for data augmentation, random cropping, random flipping, color jittering, and gaussian blurring, which were usually used in SimCLR, are mainly used. However, since color jittering cannot be used in medical images, histogram equalization was used instead of color jittering and gaussian noise was added to increase the variance of the pixel distribution. Batch size was 64 for 100 epochs, Adam optimizer was used, and learning rate was 0.01 without warm-up and learning rate decay.

3.2. Experimental details for downstream task

Like pretext task, histogram equalization, gaussian blurring, and gaussian noise were used for data augmentation. For orthodontic diagnosis, only a specific area of the cephalogram is used and this area can be calculated by cephalometric landmarks set in 1.2. For this reason, random cropping was applied while maintaining this area and random flipping was excluded. Batch size was 4 for 50 epochs, Adam optimizer was used, and learning rate was 0.0001 without warm-up and learning rate decay. Cross-entropy was used as an objective function.

4. Results

4.1. Pretext task for pretrained weights optimized our cephalometric dataset

Figure 11 shows the results of pretext task using SimSiam. In Figure 11 (a), the training loss of SimSiam with our backbone network, DenseNet-121, converged to the optimal loss as the learning progressed. Figure 11 (b) shows that the SD of the projection outputs was distributed nearby the square root of the dimension of the projection.

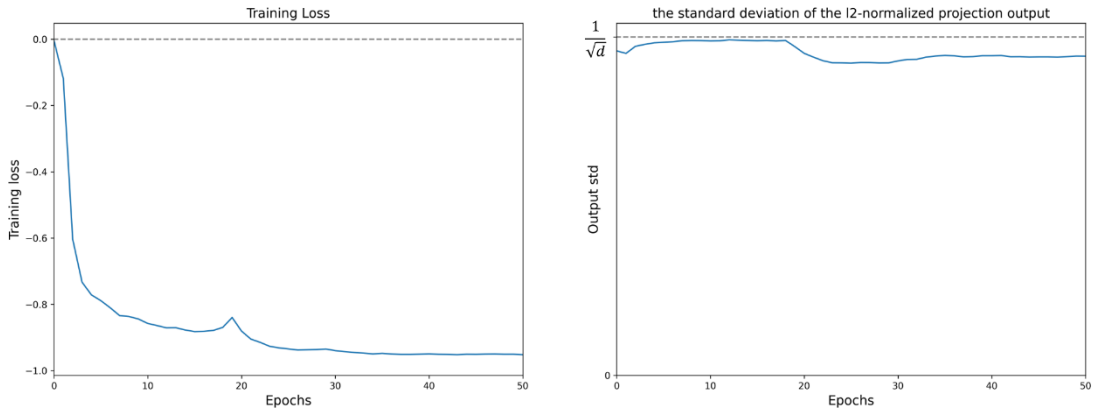

Figure 11. The results of pretext task using SimSiam. (a) training loss and (b) the SD of the projection outputs.

4.2. Downstream task to our cephalometric dataset using pre-trained weights

Figure 12 shows the results of validation accuracy comparison of models with weights of SimSiam, ImageNet, and Scratch and models trained with frozen encoder or not for different dataset ratios. Since the training of the backbone network with frozen randomly initialized weights did not progress at all, the result of this case was not included in our results.

As shown in Figure 12, the models which were trained with pre-trained weights of SimSiam (fine-tuning) showed better accuracy (0.7862, 0.8125, and 0.8618) than ImageNet (0.7730, 0.8026, and 0.8454) and Scratch (0.6020, 0.7796, and 0.7796) when the dataset ratios were 25%, 50%, and 100%, respectively. However, when the dataset ratio was 75%, ImageNet showed better accuracy (0.8520) than SimSiam (0.8289) and Scratch (0.8191). In 50% and 75%, epochs of the best performance of ImageNet were 9 and 11, respectively. In contrast, epochs of the best performance of SimSiam were more than 25 epoch in all dataset regimes.

When the backbone network was trained with frozen pre-trained weights (linear evaluation), SimSiam showed better accuracy (0.4046, 0.4375, 0.4638, and 0.4605) and ImageNet (0.3882, 0.4342, 0.4539, and 0.4079).

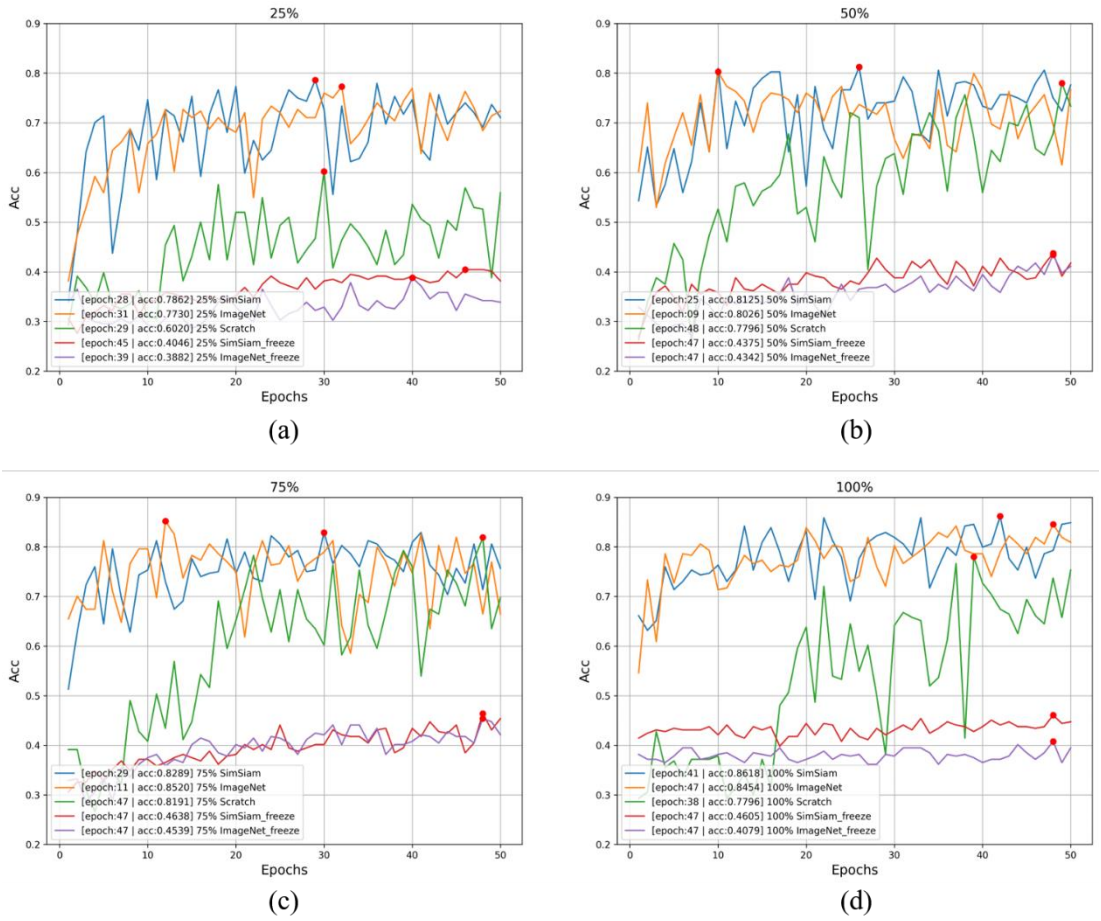

Figure 12. The results of validation accuracy comparison of models with weights of SimSiam, ImageNet, and Scratch and models trained with frozen encoder or not for different dataset ratios. (a) 25%, (b) 50%, (c) 75%, and (d) 100%.

5. Discussions

5.1. Discussion for pretext task using SimSiam

In Figure 11, the training loss converged to the optimal loss as the learning progressed and the SD of projection outputs was distributed nearby the square root of the dimension of the projection. If the projection outputs have a zero-mean isotropic Gaussian distribution, the SD of the L2-normalized projection outputs is $\frac{1}{\sqrt{d}}$, which d is the dimension of the projection. For these results, it can be known that our SimSiam was trained to extract data-specific features without collapsing.

5.2. Discussion for downstream task

Figure 12 shows the results of validation accuracy comparison of models with weights of SimSiam, ImageNet, and Scratch and models trained with frozen encoder (linear evaluation) or not (fine-tuning) for different dataset ratios. In linear evaluation, SimSiam showed better performance than ImageNet in all of data ratios. It means that SimSiam trained with 18K cephalograms only was more optimized to APSD than ImageNet, which have been regarded as well-initialized weights to extract good features of images because it was trained 1.2M large natural image dataset. In other words, ImageNet could not overcome the difference between the natural images and the medical images and the difference between instance discrimination and orthodontic diagnosis.

In fine-tuning, most ratio cases showed that SimSiam had better performance than ImageNet and Scratch, except 75%. However, in the 50% and 75% cases, the best performances of ImageNet were in the early stages of training and it seemed to be overfitting soon. It means that ImageNet can be an initialization point that can cause overfitting and should be used carefully.

5.3. Summary

Through this study, it was shown that even relatively smaller datasets than ImageNet can be used to train well-trained models. However, SimSiam architecture cannot be regarded as optimal architecture of pretext task for orthodontic diagnosis due to comparison results with ImageNet. Data augmentation, which is mainly used to obtain distinguishable features between positive pairs for instance discrimination, including SimSiam, cannot be a useful tool for non-instance discrimination, such as detection, segmentation, or tasks solving own problem. Especially, most tasks in the medical domain usually include various priors, like both skeletal deformity and spatial information have to be considered for orthodontic diagnosis simultaneously. Therefore, task-specific pretext task, which can extract intrinsic features for the task needs to be solved, like [47], [48], [49], [50], [51], should be considered delicately.

Conclusion

In this research, two studies were conducted to evaluate representation learning, especially

metric learning and self-supervised learning, in the medical domain. In the first study, metric learning was used to predict orthodontic diagnosis, resulting in improved performance and discriminative features of each data sample. In the second study, self-supervised learning was used to set pre-trained weights optimized for orthodontic diagnosis task, improving performance even in low data regimes. As can be seen from these results, research that incorporates representation learning should be considered rather than simply evaluating performance by learning by model.

Reference

[1] Deng, Jiankang, J. Guo and S. Zafeiriou. “ArcFace: Additive Angular Margin Loss for Deep Face Recognition.” *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019): 4685-4694.

[2] Wu, Yuxin and Kaiming He. “Group Normalization.” *ECCV* (2018).

[3] Ioffe, S. and Christian Szegedy. “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift.” *ArXiv* abs/1502.03167 (2015).

[4] Buda, Mateusz, A. Maki and M. Mazurowski. “A systematic study of the class imbalance problem in convolutional neural networks.” *Neural networks : the official journal of the International Neural Network Society* 106 (2018): 249-259 .

[5] Chen, Xinlei and Kaiming He. “Exploring Simple Siamese Representation Learning.” *ArXiv* abs/2011.10566 (2020).

[6] Steiner CC. “Cephalometrics for you and me.” *Am J Orthod*. 1953;39: 729-755.

[7] Jarabak JA, Fizzel JR. Technique and treatment with light wire appliances. Am J Orthod Dentofacial Orthop. 1972;64: 317-318.

[8] Tweed CH. The Frankfort-mandibular plane angle in orthodontic diagnosis, classification, treatment planning, and prognosis. Am J Orthod Oral Surg. 1946;32: 175-230.

[9] Korean Association of Orthodontics Malocclusion White Paper Publication Committee. Cephalometric analysis of normal occlusion in Korean adults. Korean Association of Orthodontists. 1997.

[10] Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition," in Proceedings of the IEEE, vol. 86, no. 11, pp. 2278-2324, Nov. 1998, doi: 10.1109/5.726791.

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet classification with deep convolutional neural networks. *In Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1* (*NIPS'12*). Curran Associates Inc., Red Hook, NY, USA, 1097–1105.

[12] Simonyan, K. and Andrew Zisserman. “Very Deep Convolutional Networks for Large-Scale Image Recognition.” *CoRR* abs/1409.1556 (2015).

[13] Szegedy, Christian, W. Liu, Y. Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, D. Erhan, V. Vanhoucke and Andrew Rabinovich. “Going deeper with convolutions.” *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015): 1-9.

[14] He, Kaiming, X. Zhang, Shaoqing Ren and Jian Sun. “Deep Residual Learning for Image Recognition.” *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016): 770-778.

[15] Huang, Gao, Zhuang Liu and Kilian Q. Weinberger. “Densely Connected Convolutional Networks.” *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017): 2261-2269.

[16] Hu, Jie, L. Shen and G. Sun. “Squeeze-and-Excitation Networks.” *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018): 7132-7141.

[17] Zoph, Barret and Quoc V. Le. “Neural Architecture Search with Reinforcement Learning.” *ArXiv* abs/1611.01578 (2017).

[18] Zoph, Barret, Vijay Vasudevan, Jonathon Shlens and Quoc V. Le. “Learning Transferable Architectures for Scalable Image Recognition.” *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018): 8697-8710.

[19] Tan, Mingxing and Quoc V. Le. “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks.” *ArXiv* abs/1905.11946 (2019).

[20] Radosavovic, Ilija, Raj Prateek Kosaraju, Ross B. Girshick, Kaiming He and Piotr Dollár.

"Designing Network Design Spaces." *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020): 10425-10433.

[21] Goyal, Priya, Piotr Dollár, Ross B. Girshick, P. Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Y. Jia and Kaiming He. "Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour." *ArXiv* abs/1706.02677 (2017).

[22] Jia, X., S. Song, W. He, Yangzihao Wang, Haidong Rong, F. Zhou, L. Xie, Zhenyu Guo, Yuanzhou Yang, L. Yu, Tiegang Chen, G. Hu, S. Shi and Xiaowen Chu. "Highly Scalable Deep Learning Training System with Mixed-Precision: Training ImageNet in Four Minutes." *ArXiv* abs/1807.11205 (2018).

[23] Cubuk, E. D., Barret Zoph, Dandelion Mané, Vijay Vasudevan and Quoc V. Le. "AutoAugment: Learning Augmentation Policies from Data." *ArXiv* abs/1805.09501 (2018).

[24] Cubuk, E. D., Barret Zoph, Jonathon Shlens and Quoc V. Le. "Randaugment: Practical automated data augmentation with a reduced search space." *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2020): 3008-3017.

[25] Berthelot, David, Nicholas Carlini, E. D. Cubuk, Alexey Kurakin, Kihyuk Sohn, Han Zhang and Colin Raffel. "ReMixMatch: Semi-Supervised Learning with Distribution Alignment and Augmentation Anchoring." *ArXiv* abs/1911.09785 (2019).

[26] C. Huang, Y. Li, C. C. Loy and X. Tang, "Learning Deep Representation for Imbalanced Classification," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5375-5384, doi: 10.1109/CVPR.2016.580.

[27] Huang, C., Y. Li, Chen Change Loy and X. Tang. "Deep Imbalanced Learning for Face Recognition and Attribute Prediction." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42 (2020): 2781-2794.

[28] Lin, Tsung-Yi, Priya Goyal, Ross B. Girshick, Kaiming He and Piotr Dollár. "Focal Loss for Dense Object Detection." *2017 IEEE International Conference on Computer Vision (ICCV)* (2017): 2999-3007.

[29] Liu, Weiyang, Y. Wen, Zhiding Yu, Ming Li, B. Raj and Le Song. "SphereFace: Deep Hypersphere Embedding for Face Recognition." *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017): 6738-6746.

[30] Wang, Feng, Jian Cheng, Weiyang Liu and H. Liu. "Additive Margin Softmax for Face Verification." *IEEE Signal Processing Letters* 25 (2018): 926-930.

[31] Wardlaw DW, Smith RJ, Hertweck DW, Hildebolt CF. "Cephalometrics of anterior open bite: a receiver operating characteristic (ROC) analysis." *Am J Orthod Dentofacial Orthop*. 1992;101: 234-243.

[32] L.J.P. van der Maaten and G.E. Hinton. "Visualizing High-Dimensional Data Using t-SNE.**"** *Journal of Machine Learning Research* 9(Nov):2579-2605, 2008.

[33] Selvaraju, Ramprasaath R., Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh and Dhruv Batra. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization." *International Journal of Computer Vision* 128 (2019): 336-359.

[34] Dosovitskiy, A., Jost Tobias Springenberg, Martin A. Riedmiller and T. Brox. "Discriminative Unsupervised Feature Learning with Convolutional Neural Networks." *NIPS* (2014).

[35] Doersch, Carl, A. Gupta and Alexei A. Efros. "Unsupervised Visual Representation Learning by Context Prediction." *2015 IEEE International Conference on Computer Vision (ICCV)* (2015): 1422-1430.

[36] Noroozi, M. and P. Favaro. "Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles." *ECCV* (2016).

[37] Larsson, Gustav, M. Maire and Gregory Shakhnarovich. "Colorization as a Proxy Task for Visual Understanding." *2017 IEEE Conference on Computer Vision and Pattern Recognition*

(CVPR) (2017): 840-849.

[38] Gidaris, Spyros, Praveer Singh and Nikos Komodakis. “Unsupervised Representation Learning by Predicting Image Rotations.” *ArXiv* abs/1803.07728 (2018).

[39] Oord, Aäron van den, Y. Li and Oriol Vinyals. “Representation Learning with Contrastive Predictive Coding.” *ArXiv* abs/1807.03748 (2018).

[40] Hénaff, Olivier J., A. Srinivas, J. Fauw, Ali Razavi, Carl Doersch, S. Eslami and Aäron van den Oord. “Data-Efficient Image Recognition with Contrastive Predictive Coding.” *ArXiv* abs/1905.09272 (2020).

[41] He, Kaiming, Haoqi Fan, Yuxin Wu, Saining Xie and Ross B. Girshick. “Momentum Contrast for Unsupervised Visual Representation Learning.” *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020): 9726-9735.

[42] Chen, Xinlei, Haoqi Fan, Ross B. Girshick and Kaiming He. “Improved Baselines with Momentum Contrastive Learning.” *ArXiv* abs/2003.04297 (2020).

[43] Chen, Xinlei, Saining Xie and Kaiming He. “An Empirical Study of Training Self-Supervised Vision Transformers.” *ArXiv* abs/2104.02057 (2021).

[44] Chen, Ting, Simon Kornblith, Mohammad Norouzi and Geoffrey E. Hinton. “A Simple Framework for Contrastive Learning of Visual Representations.” *ArXiv* abs/2002.05709 (2020).

[45] Chen, Ting, Simon Kornblith, Kevin Swersky, Mohammad Norouzi and Geoffrey E. Hinton. “Big Self-Supervised Models are Strong Semi-Supervised Learners.” *ArXiv* abs/2006.10029 (2020).

[46] Caron, Mathilde, Ishan Misra, J. Mairal, Priya Goyal, P. Bojanowski and Armand Joulin. “Unsupervised Learning of Visual Features by Contrasting Cluster Assignments.” *ArXiv* abs/2006.09882 (2020).

[47] Wang, Xinlong, Rufeng Zhang, Chunhua Shen, Tao Kong and Lei Li. “Dense Contrastive Learning for Self-Supervised Visual Pre-Training.” *ArXiv* abs/2011.09157 (2020).

[48] Xie, Zhenda, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin and Han Hu. “Propagate Yourself: Exploring Pixel-Level Consistency for Unsupervised Visual Representation Learning.” *ArXiv* abs/2011.10043 (2020).

[49] Roh, Byung-Seok, W. Shin, Ildoo Kim and S. Kim. “Spatially Consistent Representation Learning.” *ArXiv* abs/2103.06122 (2021).

[50] Pinheiro, Pedro H. O., Amjad Almahairi, Ryan Y. Benmaleck, Florian Golemo and Aaron C. Courville. “Unsupervised Learning of Dense Visual Representations.” *ArXiv* abs/2011.05499 (2020).

[51] H'enaff, Olivier J., Skanda Koppula, Jean-Baptiste Alayrac, Aäron van den Oord, Oriol Vinyals and João Carreira. “Efficient Visual Pretraining with Contrastive Detection.” *ArXiv* abs/2103.10957 (2021).

[52] Grill, Jean-Bastien, Florian Strub, Florent Altch'e, C. Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, B. A. Pires, Zhaohan Daniel Guo, M. G. Azar, Bilal Piot, K. Kavukcuoglu, R. Munos and Michal Valko. “Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning.” *ArXiv* abs/2006.07733 (2020).

[53] Chen, Xinlei and Kaiming He. “Exploring Simple Siamese Representation Learning.” *ArXiv* abs/2011.10566 (2020).

[54] Zbontar, J., L. Jing, Ishan Misra, Y. LeCun and Stéphane Deny. “Barlow Twins: Self-Supervised Learning via Redundancy Reduction.” *ArXiv* abs/2103.03230 (2021).

Abstract (with Korean)

인공지능 기술 중 딥러닝은 컴퓨터비전 뿐만 아니라 자연어 처리, 강화 학습, 추천 시스템, 데이터 베이스 등 다양한 분야에 적용되어 눈에 띄는 성능 향상을 보이고 있다. 현재 분류 (classification), 검출 (detection), 분할 (segmentation), 자동 요약 (automatic summarization), 기계 번역 (machine translation), 질의 응답 (question answering) 등 큰 범위의 테스크들은 어느 정도 성능이 수렴된 것처럼 보인다. 실제로 2-3 년 전처럼 큰 폭의 성능 향상을 이루는 연구들보다 기존의 알고리즘의 특정 부분을 개선하여 1-2%의 성능 향상을 이뤄낸 연구들이 주를 이루고 있다.

하지만 이런 테스크들에 기반하여 파생된 세부 테스크 또는 다른 도메인에 동일한 연구가 적용되었을 때 성능 향상은 기대만큼 크지 않거나 오히려 하락하는 경우가 많다. 대표적인 원인으로 큰 테스크를 목표로 구현된 모델 구조가 세부 테스크나 다른 도메인에는 적합하지 않은 경우가 될 수 있고, 주어진 데이터 분포가 달라지는 경우도 될 수 있다. 이를 해결하기위해 최근에는 단순히 모델 구조와 알고리즘으로 해결하는 연구가 아닌, 모델로부터 얻은 특징을 잘 처리해내는 연구인 표현 학습 (representation learning)이 활발하게 연구되고 있다.

표현 학습에는 얼굴 인식에서 많이 사용되는 거리 학습 (metric learning)과 데이터의 본질적인 특징을 얻기 위한 자기 지도 학습 (self-supervised learning) 등 다양한 연구 분야가 존재한다. 이 연구들의 공통점은 주어진 데이터 간의 특정한 규칙을 이용해 군집화를 하는 과정으로 학습하는 것이다. 이 과정을 거치고 나면 성능 개선은 물론, 연구자가 데이터의 분포를 시각적으로 확인할 수 있도록 도와준다는 장점이 있다.

위에서 설명한 방법들은 의료 영상을 이용한 인공지능 연구에서 꼭 필요한 요소들이다. 일반 영상과 달리 영상의 정보를 결정하는 부분이 단 몇 픽셀부터 영상 대부분의 픽셀까지 크기의 다양성이 존재하고, 화소 강도도 8bit 에서 16bit 로 다양하게 분포되어 있으며, 연관된 사전 정보 (prior)나 사후 정보 (posterior)가 주어진 데이터와 복잡하게 얽힌 경우가 많다. 그리고 데이터를 수집하는 과정에서 수집된 기관의 특이성, 수집된 기간 등 다양한 조건에 의해 데이터의 분포가 천차만별로 달라질 수 있다. 이러한 환경속에서 의료 데이터의 분포와 데이터 특징의 분포를 다루는 연구는 필수라고 할 수 있다.

이 연구에서는 cephalogram 을 이용하여 거리 학습과 자기 지도 학습이 의료 영상에 어떤 영향을 미치는지 확인하기위해 ‘cephalogram 을 이용한 교정 진단’과 ‘자기 지도 학습이 교정 진단에 미치는 영향’, 이 두 가지 실험을 진행했다.

먼저 ‘cephalogram 을 이용한 교정 진단’ 연구에서는 전후 골격 불일치 (anteroposterior skeletal discrepancies, APSD: Class I, Class II, and Class III), 수직 골격 불일치 (vertical skeletal discrepancies, VSD: normo-divergent, hyper-divergent, and hypo-divergent), 수직 치아 불일치 (vertical dental discrepancies, VDD: normal overbite, open bite, and deep bite) 총 3 가지 진단 예측을 진행한다. 모든 진단은 두부 계측 랜드마크 (cephalometric landmark)를 기준으로 계산된 계측치를 기준으로 분류될 수 있다. 따라서 개별 진단 사이의 기준 영역 (gray zone)에서 잘 구분하는 것이 이 연구의 핵심이다. 또한, 데이터를 수집한 10 개의 기관 모두 수술 중심 기관이기 때문에 수술과 관련이 깊은 레이블의 데이터가 많았다. 이를 위해 거리 학습 모델 중 ArcFace 를 기존 모델에 추가하여 분별력있는 특징을 유도하도록 했다. 또한, 적은 수의 학습 데이터를 효과적으로 사용하기위해 작은 배치 사이즈로 학습을 진행했고, 작은 배치 사이즈에 좋은 성능을 보이는 Group Normalization 을 Batch Normalization 대신 사용했다. 그 결과, 전체 데이터로 전후 골격 불일치 진단을 위한 학습을 진행했을 때, 제안된 모델이 다른 모델보다 좋은 성능을 보였다. 또한, 2 개 기관의 데이터로 전체 진단에 대해 학습을 진행했을 때 학습에 사용된 2 개 기관 데이터로 이루어진 테스트셋으로 진행된 internal validation 과 나머지 8 개 기관 데이터로 이루어진 테스트셋으로 진행된 external validation 에서 거의 동등한 성능을 얻을 수 있었다.

두 번째 ‘자기 지도 학습이 교정 진단에 미치는 영향’ 연구에서는 자기 지도 학습 모델 중 SimSiam 을 cephalogram 으로 사전 학습을 진행하고, 전후 골격 불일치 진단에 대해 추가 학습을 진행했다. 비교를 위해 임의로 초기화된 가중치와 ImageNet 데이터셋으로 사전 학습된 모델의 가중치를 사용하여 전후 골격 불일치 진단에 대해 학습을 진행했다. 그 결과, 가중치를 모두 고정한 선형 평가 (linear evaluation)와 모델 전체를 재학습하는 미세 조정 (fine-tuning)을 진행했을 때 SimSiam 으로 만든 가중치를 사용한 경우가 ImageNet 데이터셋으로 만든 가중치와 임의로 초기화된 가중치를 사용한 경우보다 전체 또는 적은 데이터 환경에서 좋은 성능을 보였고 과적합을 유발하지 않았다.

두 연구를 통해 의료 영상에서 거리 학습과 자기 지도 학습을 이용하면 성능 개선은 물론, 분별력있는 특징을 추출하고 데이터 분포와 데이터 수에 강인한 모델을 학습할 수 있음을 확인할 수 있었다. 앞으로 진행될 의료 영상 인공지능에서도 단순히 모델별로 학습하여 성능을 평가하는 것이 아닌, 표현 학습이 수반된 연구가 진행되어야 할 것이다.