CSI Research Summary: 2017 - 2019

Dabi Ahn[10] **Hyun-Jin Bae**[6] **Jinhyun Bang**[3] **Younghwa Byeon**[6]
Moon-jung Chae[3] **Sungkyun Chang**[7] **Jaemin Cho**[9,*] **Sukhyun Cho**[3]
Yongwon Cho[6] **Kyoyun Choi**[3] **Da-in Eun**[6] **Sungwon Ham**[6] **Bohyung Han**[1]
Wan Huh[3] **Kyoungseok Jang**[8] **Sungwook Jeon**[3] **Jinhoon Jeong**[6] **Yeonwoo Jeong**[5]
Wanmo Kang[8,†] **Byeongchang Kim**[9] **Chang-Wook Kim**[10] **Chris Dongjoo Kim**[9]
Dokyun Kim[3] **Gunhee Kim**[9,*] **Hyunwoo Kim**[9] **Namju Kim**[10] **Namkug Kim**[6]
Yoonsung Kim[5] **Young-Gon Kim**[6] **A-Reum Lee**[6] **Cheolhyoung Lee**[8]
Donmoon Lee[7] **Hyunna Lee**[6] **Juheon Lee**[7] **Kyogu Lee**[7] **Seungjin Lee**[7]
Hyungui Lim[7] **Jonghwan Mun**[1,2] **Seil Na**[9] **Hyeonwoo Noh**[1,2] **Beomhee Park**[6]
Cesc Chunseong Park[9] **Heewoong Park**[3] **Jonggwon Park**[3] **Jonghun Park**[3]
Jonghyuk Park[3] **Kyubyong Park**[10] **Yookoon Park**[9] **Seonghee Ryu**[6]
Hyun Oh Song[5] **Youngji Song**[6] **Soobin Suh**[3,4] **Ilsang Woo**[6] **Jihye Yun**[6]
[1]Computer Vision Lab., Seoul National University [2]Computer Vision Lab., POSTECH
[3]Information Management Lab., Seoul National University [4]Naver Corporation
[5]Machine Learning Lab., Seoul National University
[6]Medical Imaging and Intelligent Reality Lab.,
University of Ulsan College of Medicine, Asan Medical Center
[7]Music and Audio Research Group, Seoul National University
[8]Probability and Optimization with Applications Lab.,
Korea Advanced Institute of Science and Technology
[9]Vision and Learning Lab., Seoul National University
[10]Kakao Brain Corporation
[1]{bhhan}@snu.ac.kr [1,2]{jonghwan.mun,shgusdnrhrh}@postech.ac.kr
[3]{jinhyun95, jjamjung, chosh90, andyhome1907, jrw8217, wookee3, kdk01,
hee188, jayg996, jonghun, chico2121}@snu.ac.kr [3,4]{soobin323025}@gmail.com
[5]{yeonwoo, yskim227, hyunoh}@mllab.snu.ac.kr
[6]{hjbae.astro, harmoniousbyh, dragon1won, eundain94, swlove920,
nineclas, namkugkim, younggon2.kim, lareum613, hyunnalee,
qkr18x, seonghee.ryu, youngji0428, bidfore, dool0120}@gmail.com
[7]{rayno1, lunideal, juheon2, kglee, joshua77, goongding7}@snu.ac.kr
[8]{jajajang, bloodwass}@kaist.ac.kr [8,†]{wanmo.kang}@kaist.edu
[9]{byeongchang.kim, cdjkim, hyunwoo.kim, seil.na, cs.park,
yookoon.park}@vision.snu.ac.kr [9,*]{jaemin895, gunhee}@snu.ac.kr
[10]{dabi.ahn, dade.ai, hexa.ai, kyubyong.park}@kakaobrain.com

Abstract

Center for SuperIntelligence[1] (CSI) was founded in April 2017 with generous support by Kakao and Kakao Brain corporations. CSI aims to be a hub for state-of-the-art deep learning research and strives to focus on making significant progress in the field of artificial intelligence (AI) for social benefit while serving as a bootcamp for future AI leaders. In this paper, we highlight some of our efforts from April 2017 to March 2018. We hope that our work could contribute to progress towards next-generation AI, and look forward to what will come in the following years.

[1]http://csi.snu.ac.kr/

1 Introduction

Supported by Kakao and Kakao Brain corporations, Center for SuperIntelligence (CSI) was founded in April 2017 with an aim to make significant contributions in the field of artificial intelligence (AI). Eight research laboratories from Asan Medical Center at University of Ulsan College of Medicine, Korea Advanced Institute of Science and Technology (KAIST), Seoul National University (SNU), and University of Southern California have been involved in CSI, carrying out cutting-edge research in deep learning theory and its application to various domains including vision, audio, speech, language, music, and medicine.

We believe that active collaboration between industry and academia is of paramount importance to the future of AI. Through developing state-of-the-art deep learning technology and sharing findings and breakthroughs in the form of open research between industry and academia, CSI strives to make significant progress in the field of AI for social benefit while serving as a bootcamp for future AI leaders.

The last two years were exciting and challenging for CSI research teams, with our work advancing the deep learning technology in many ways, including publications in journals and conference proceedings, open source software contributions, participating in competitions, and active collaborations with Kakao and Kakao Brain engineers. Selected research outcomes of CSI during the period from April 2017 to March 2019 are summarized in this paper, with each section reviewing research results from the individual laboratories that are currently affiliated with CSI. The paper is organized as follows.

In Section 2, computer vision laboratory of SNU reports its recent achievements on visual question answering. Information management laboratory at SNU, focusing on speech recognition/synthesis, reading comprehension, as well as music generation summarizes its work in Section 3. Section 4 is provided by machine learning laboratory which is a research group at SNU with a focus on advancing machine learning algorithms and their underlying theories. Research results from medical imaging and intelligent reality laboratory (MI2RL) at Asan Medical Center, whose interests span the areas of image-based clinical applications such as AI, 3D printing, medical image processing, computer-aided surgery, and robotic interventions are shown in Section 5. Music and audio research group (MARG) at SNU is a laboratory that concentrates on understanding and analyzing the characteristics of various sounds and music, and Section 6 reviews various deep learning based music analysis and recommendation methods developed. In Section 7, probability and optimization with applications laboratory from KAIST provides its research findings in theoretical analysis of deep learning methods. Recent advancements in natural language understanding research in Section 8 are contributed by vision and learning laboratory of SNU. Finally, we conclude the paper with some remarks in Section 9.

2 Computer Vision Lab., Seoul National University

2.1 Learning to Specialize with Knowledge Distillation for Visual Question Answering

Visual Question Answering (VQA) is a notoriously challenging problem because it involves various heterogeneous tasks defined by questions within a unified framework. Learning specialized models for individual types of tasks is intuitively attractive but surprisingly difficult; it is not straightforward to outperform naïve independent ensemble approaches. In [1], we tackled on how to associate models with individual types of tasks and how to learn the specialized models effectively.

For the purpose, we propose a novel framework referred to as Multiple Choice Learning with Knowledge Distillation (MCL-KD). Given a training dataset $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$ and M independently trained base models with fixed model parameters ϕ_m, we train M models parametrized by θ_m in the proposed MCL-KD framework using the following loss:

$$\begin{aligned} &\mathcal{L}_{\text{MCL-KD}}(\mathcal{D}) \qquad (1)\\ &= \sum_{i=1}^{N}\sum_{m=1}^{M} v_{i,m}\ell_{\text{task}}(y_i, P(y|x_i;\theta_m)) + \beta(1-v_{i,m})\ell_{\text{KD}}\left(P(y|x_i;\phi_m,T), P(y|x_i;\theta_m,T)\right), \\ &\qquad\qquad \text{subject to} \quad \sum_{m=1}^{M} v_{i,m} = k, \quad v_{i,m} \in \{0,1\}, \end{aligned}$$

where $\ell_{\text{task}}(\cdot,\cdot)$ denotes a task specific loss function (*i.e.* cross-entropy loss in classification), $v_{i,m}$ is an indicator variable for the assignment of x_i to the m-th model, $k\ (=1,\ldots,M)$ is the number of specialized models per example, $\beta > 0$ is a hyper-parameter to balance the two loss terms, and $T > 0$ is a temperature scaling parameter of the softmax function. Here, we employ a knowledge distillation loss between the m-th base model and the corresponding specialized model $\ell_{\text{KD}}(\cdot)$, which is formally given by

$$\ell_{\text{KD}}\left(P(y|x_i;\phi_m,T), P(y|x_i;\theta_m,T)\right) = D_{KL}\left(P(y|x_i;\phi_m,T)||P(y|x_i;\theta_m,T)\right), \tag{2}$$

where $P(y|x;\theta,T) = \frac{\exp(f_y(x;\theta)/T)}{\sum_{y'}\exp(f_{y'}(x;\theta)/T)}$ is a calibrated softmax distribution, and $f(\cdot)$ denotes a logit of deep models. Our proposed loss drives specialized models to predict the ground-truth answers while non-specialized ones preserve the representations of the corresponding base models.

Table 1: Classification accuracy (%) on CLEVR validation set with varying the number of specialized models (k). The bold-faced numbers mean the best algorithm for each k in top-1 accuracy.

Answering Network	k	Single	IE		MCL		CMCL		MCL-KD	
		Top-1	Top1	Oracle	Top-1	Oracle	Top-1	Oracle	Top-1	Oracle
MLP	1	58.40	60.10	80.73	41.31	98.92	59.12	63.76	**60.22**	80.75
	2				48.94	97.57	60.27	76.00	**60.38**	81.20
	3				58.63	95.67	60.49	82.67	**60.89**	81.86
SAN	1	82.30	85.23	94.93	42.19	98.67	83.99	91.55	**85.98**	95.38
	2				58.39	98.62	84.83	96.64	**87.02**	95.78
	3				83.73	98.62	86.18	96.26	**88.16**	96.12

For evaluation, we adopt two models as our answering networks: a simple MLP-based model with 2 hidden layers of 1,024 units after an image and question fusion layer, and a well-known stacked attention network (SAN)[2]. We measure both top-1 and oracle accuracy; the top-1 accuracy is computed by the ratio of correctly predicted examples while the oracle accuracy measures whether at least one of the models predicts the correct answers. Generally speaking, higher oracle accuracy implies that trained models are more specialized to subsets of data.

Table 1 summarizes the results from models learned by independent ensemble (IE), other two MCL-based algorithms (MCL and CMCL) and our proposed method (MCL-KD) on the validation set of CLEVR dataset[3]. MCL is the best in terms of oracle accuracy, but its top-1 accuracy is not satisfactory due to overconfidence issue. CMCL is substantially better than MCL, but still not sufficient to achieve clear accuracy improvement with respect to IE. On the contrary, MCL-KD consistently outperforms IE, MCL, and CMCL regardless of k. This implies that applying knowledge distillation loss for non-specialized models is important to balance specialization and generalization of ensemble models in visual question answering.

2.2 Transfer Learning via Unsupervised Task Discovery for Visual Question Answering

Existing large-scale visual data with annotations such as image class labels, bounding boxes and region descriptions are good sources for learning rich and diverse visual concepts. In [2], we study exploiting these sources of information to cope with out-of-vocabulary answers in visual question answering (VQA) problem. We tackle this problem in two steps: 1) learning a task conditional visual classifier based on unsupervised task discovery and 2) transferring and adapting the task conditional visual classifier to visual question answering models.

Our problem setting considers VQA with out-of-vocabulary answers. External visual data provides a set of labels $\mathcal{A}$ and only a subset of these labels $\mathcal{B} \subset \mathcal{A}$ appears in VQA training set as answers. The goal of this setting is to handle out-of-vocabulary answers $\mathbf{a} \in \mathcal{A} - \mathcal{B}$ successfully by exploiting visual concepts learned from external visual data.

The main idea of the proposed method is to learn visual concepts using off-the-shelf visual data and transfer the concepts to VQA models. To adapt the learned visual concepts to diverse questions about

[2]Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *CVPR*, 2016.

[3]Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017.

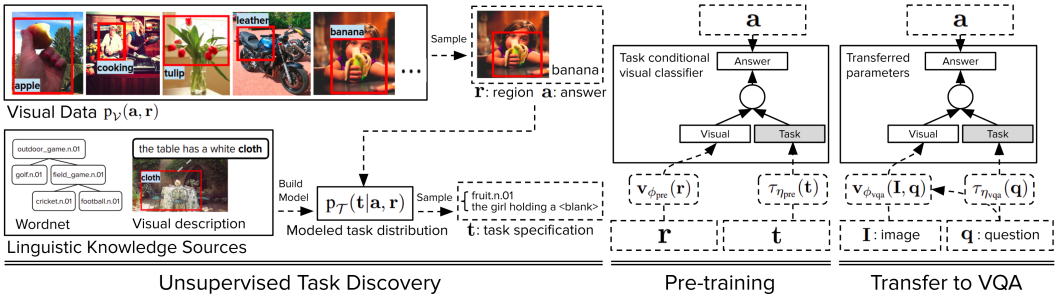


Figure 1: **Overview of the proposed algorithm.** (Left) Unsupervised task discovery learns a task conditional visual classifier by leveraging off-the-shelf visual data without task specification $\mathbf{t}$. It also defines a task distribution $p_{\mathcal{T}}(\mathbf{t}|\mathbf{a}, \mathbf{r})$ using linguistic knowledge sources, where stochastic sampling associates a task specification $\mathbf{t}$ with a visual annotation $(\mathbf{a}, \mathbf{r})$. (Center) A visual annotation with a task specification, denoted by $(\mathbf{a}, \mathbf{r}, \mathbf{t})$, is employed to pre-train a task conditional visual classifier. (Right) The pre-trained task conditional visual classifier is transferred to VQA with the learned parameters and adapts input representations $\mathbf{v}_{\phi_{\text{vqa}}}(\mathbf{I}, \mathbf{q})$ and $\tau_{\eta_{\text{vqa}}}(\mathbf{q})$ without fine-tuning.

a visual scene, we should learn not only a name of a visual concept but also a task related to the concepts. We propose unsupervised task discovery to identify visual recognition tasks from visual data without question or task annotations. Figure 1 illustrates an overview of the proposed algorithm.

Figure 2 illustrates model comparison results. The standard VQA model fails to predict any out-of-vocabulary answers (i.e., 0 VQA score) because there is no clue for inferring out-of-vocabulary answers. Using off-the-shelf visual data and task specifications from linguistic knowledge sources dramatically improves performance both on the separable classifier baseline and the proposed model. Especially, the proposed model based on transfer learning task conditional visual classifier pre-trained with unsupervised task discovery outperforms all other baselines significantly in VQA with out-of-vocabulary answers.

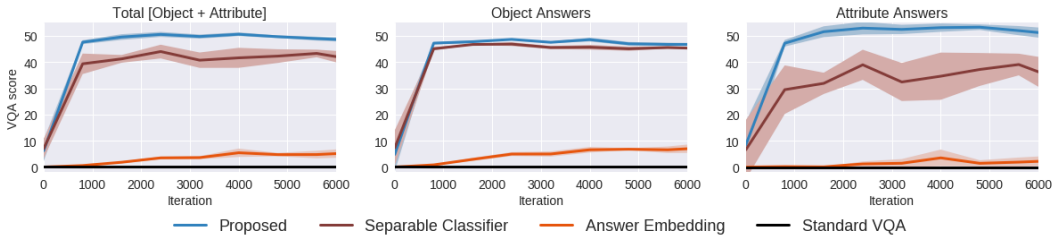


Figure 2: **Model comparisons.** Exploiting external data with unsupervised task discovery boosts performance of the proposed model and separable classifier significantly. However, the separable classifier shows limited performance gain on attribute answers, which have significant variations depending on tasks.

3 Information Management Lab., Seoul National University

3.1 Speech Recognition & Synthesis

We proposed a non-sequential greedy decoding (NSGD) method that generalizes the greedy decoding schemes presented in the past [3]. The greedy decoding method used in the conventional sequence-to-sequence models is prone to producing a model with a compounding of errors, mainly because it makes inferences in a fixed order, regardless of whether or not the model's previous guesses are correct. The proposed method determines not only which token to consider, but also which position in the output sequence to infer at each inference step. Specifically, it allows the model to consider easy parts first, helping the model infer hard parts more easily later by providing more information.

We also studied a grapheme-to-phoneme conversion task with a fully convolutional encoder-decoder model that embeds the proposed decoding method. Figure 3 shows an example of the inference

procedure of our model, generating the phoneme sequence from the grapheme sequence 'KNIGHT'. We showed the effectiveness of the proposed model and decoding method by achieving the state-of-the-art performances on the G2P task with various datasets and demonstrating the decoding results. Our implementation code is available at [4].

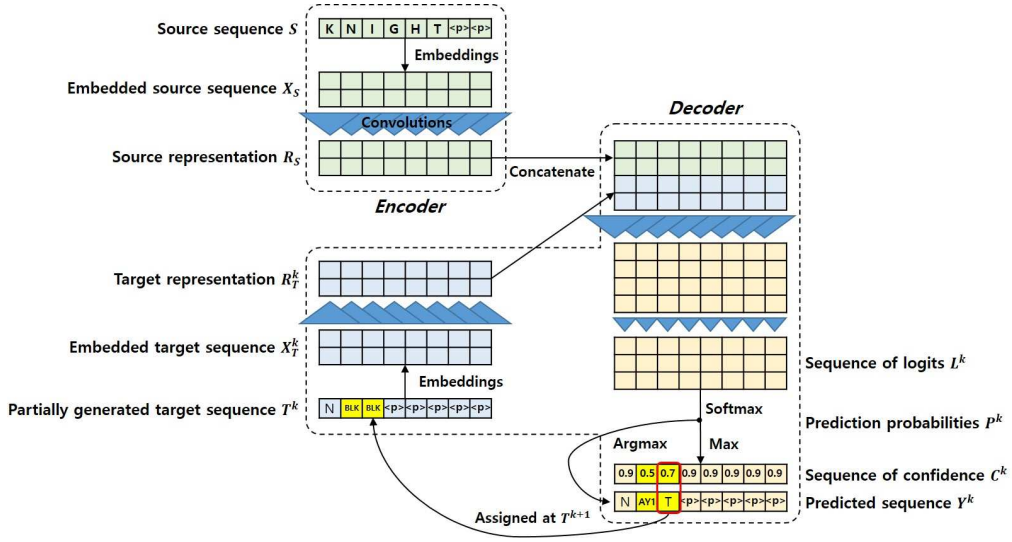


Figure 3: The proposed convolutional encoder-decoder model with NSGD during test.

A method that learns speaker embeddings for identification and verification while requiring relatively small training data and computing resources was presented in [5]. For speaker recognition systems, extracting speaker-specific information from speech signals and then representing it with fixed length vectors is a crucial component. We proposed a deep neural network based embedding model where the similarity of embedded utterance vectors directly serves as a score for a back-end classifier. Technically, the network was optimized with not only a typical cross-entropy loss but also an additional loss based on the embedding similarities. In addition, we suggested a new feature sampling method to augment training data. Our experimental results on two databases demonstrated that the method was more effective at speaker identification and verification when compared to decent baseline methods.

In [6], we studied how to reflect a target speaker's various speaking patterns in voice conversion. Previous voice conversion methods have a limitation in that they generate deterministic output, which leads to only a fixed speaking style for each source speech segment. In this study, we employed a conditional variational autoencoder (VAE) for voice conversion in order to learn a set of speaking styles that an individual person has in an unsupervised manner. To obtain more realistic speech, inverse autoregressive flow, which enables more flexible posterior, was applied between the encoder and the decoder. As a result, our method was capable of generating converted utterances with diverse intonation patterns without severe quality degradation in terms of naturalness.

Another research outcome pertaining to speech recognition and synthesis can be found in [7] where an improved training algorithm for semi-supervised learning methods that jointly train TTS and ASR models by back-translation is presented. This work was the first study to improve the efficiency of training TTS and ASR models using unpaired datasets through a novel approach that updates data synthesizers in back-translation by considering the inherent characteristics of TTS and ASR tasks. In contrast to the previous studies that used a large amounts of paired data, which is not applicable in the low-resource languages, we verified that TTS and ASR models could be trained in the low-resource situation by using the proposed training algorithm. Along with this research, we made a PyTorch implementation of Tacotron[4] which is available at [8].

[4]Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. Tacotron: A fully end-to-end text-to-speech synthesis model. *arXiv preprint arXiv:1703.10135*, 2017.

3.2 Machine Reading Comprehension

In addition to application of deep learning models to the speech domain mentioned above, in [9], we focused on machine learning approach for sentence completion, which aims to pick the correct word or phrase to complete a sentence out of given candidates. Since the sentence completion task requires diverse skills and knowledge including linguistic ability, logical reasoning, and common sense, the task has been used to assess the comprehension level of not only humans but also machines. We demonstrated that a word-level Recurrent Neural Network (RNN), a popular neural language model, can be a competitive baseline for sentence completion by improving the performance with a proper network structure and hyper-parameters. We also proposed a bidirectional version of word RNN, which was confirmed to give further improvement. Despite its simple architecture, the accuracy results of the word RNNs exceeded the previous best results on Microsoft Research Sentence Completion Challenge and the Scholastic Aptitude Test (SAT) sentence completion questions. Our implementation code for the experimentation is available at [10], and an implementation of ByteNet[5] that tried to improve the model with a tensor masking technique can be found in [11].

3.3 Music Generation

In [12], we applied a bi-directional transformer model to music chord recognition task. The goal of chord recognition task is to output a sequence of time-synchronized chord labels given a raw audio recording of music as input. Most previous deep learning based automatic chord recognition systems rely on recurrent neural networks (RNN). While RNNs are known to fit well with sequential data, they have a drawback in that they fail to capture long-term dependency as the sequence length gets longer. To overcome this, we used a bi-directional attention mechanism. The proposed model well captures long-term dependency and its training procedure is relatively simple. For a chord recognition task, it is important to identify the point where the chord changes and segment the chord interval. The proposed model is able to segment the interval using attention and use the related information for chord recognition, whereas RNNs rely on the sequence of the previous information without screening. Therefore, the proposed model appears to be more suitable for the task. We evaluated the proposed model with weighted chord symbol recall (WCSR) score and achieved performance that was comparable to the other state-of-the-art models. In addition, the proposed approach enables visualization of how the model works with attention map. Through attention map, we verified that the model was properly trained for the chord recognition task.

Furthermore, with the recent advancements of deep generative models such as VAEs and generative adversarial networks (GANs), it has become possible to create more realistic music. Researches of automatic music composition have extensively used pianoroll or lead sheets which usually contain melody line and a sequence of accompanying chords. While the latent space based generative models such as VAEs and GANs are able to produce realistic music, the existing methods are only using chord information additionally rather than reflecting the relationship between chord and melody on the latent space. To address this problem, we separately projected chord and melody on latent space through posterior network and prior network respectively [13]. As similar chord sequences are distributed closely in latent space, it became possible to compose a melody that is more aligned with a given chord sequence.

4 Machine Learning Lab., Seoul National University

Learning the feature representations with similarity information among data has generated considerable research interest from machine learning and computer vision community recently. Deep metric learning methods have shown great performance on training representations where similar pairs of data are close to each other and vice versa for dissimilar pairs.

However, there remains a need for improving the inference efficiency and handing scalability of the representation in end-to-end learning. To handle this, practitioners of deep metric learning methods implement the post-processing technique such as thresholding and vector quantization while compromising the performance of the models.

[5]Nal Kalchbrenner, Lasse Espeholt, Karen Simonyan, Aaron van den Oord, Alex Graves, and Koray Kavukcuoglu. Neural machine translation in linear time. *arXiv preprint arXiv:1610.10099*, 2016.

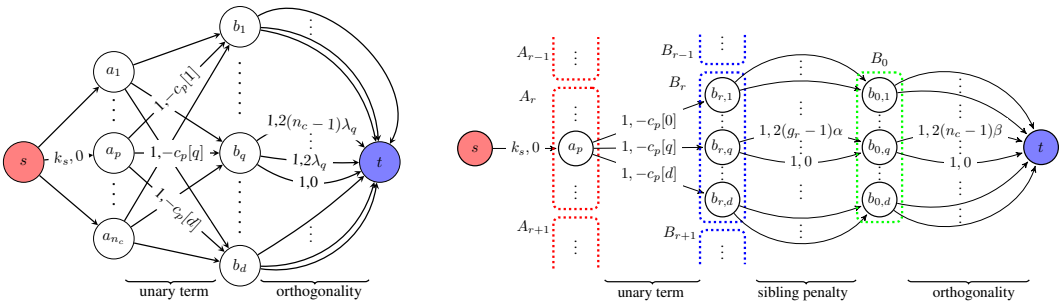
In this regard, [14] and [15] proposed algorithms for quantizable representation by training the deep convolutional neural network and optimizing the sparse constrained binary code at the same time. Quantizable representation via hierarchical structure improves the scalability of the representation [15].

Equation (3) and Equation (4) are the proposed hash functions of [14] and [15], respectively. Both equations generate binary codes with k_s activated dimensions from the representation. Binary hash code generates hash table by hashing the data $\mathbf{x}$ into the corresponding activated hash buckets.

$$r(\mathbf{x}) = \underset{\mathbf{h}\in\{0,1\}^d}{\operatorname{argmin}} -f(\mathbf{x};\boldsymbol{\theta})^\intercal \mathbf{h} \quad (3) \qquad \text{subject to } \|\mathbf{h}\|_1 = k_s$$

$$r(\mathbf{x}) = \bigotimes_{v=1}^{k} \underset{\mathbf{h}^v}{\operatorname{argmin}} -\left(f(\mathbf{x};\boldsymbol{\theta})^v\right)^\intercal \mathbf{h}^v \quad (4) \qquad \text{subject to } \|\mathbf{h}^v\|_1 = \begin{cases} 1 & \forall v \neq k \\ k_s & v = k \end{cases} \text{ and } \mathbf{h}^v \in \{0,1\}^d$$

The representation $f(\mathbf{x};\boldsymbol{\theta}) \in \mathbb{R}^d$ generates d-dimensional sparsity constrained binary hash code in Equation (3), while the representation $f(\mathbf{x};\boldsymbol{\theta}) = [f(\mathbf{x};\boldsymbol{\theta})^1, \ldots, f(\mathbf{x};\boldsymbol{\theta})^k] \in \mathbb{R}^{d\times k}$ generates d^k-dimensional sparsity constrained binary hash code in Equation (4) with k hash codes $h^1, \ldots, h^k$. The representation in [14] suffers from scalability of the dimension in hash codes, while that in [15] improves scalability through k-level hierachical structure proposed in Equation (4). Embedding representations $f(\mathbf{x};\boldsymbol{\theta})$ and binary hash codes $r(\mathbf{x})$ are optimized via alternating minimization scheme through solving discrete optimization problem for sparse-constrained binary hash codes and updating parameters of the deep neural network, $\boldsymbol{\theta}$.



(a) G_1 (b) G_2

Figure 4: Equivalent flow network diagrams G_1, G_2 corresponding to discrete optimizations. Edge labels show the capacity and the cost, respectively. The amount of total flow to be sent is $n_c k_s$.

Class mean vector is defined as in $\mathbf{c}_p^v = \frac{1}{m}\sum_{i:y_i=p} f(\mathbf{x}_i;\boldsymbol{\theta})^v$ [6] where y_i is a temporal class label for similarity information in a mini-batch setting in [14] and [15], respectively. Binary hash codes can be exactly optimized through solving Equation (5) and Equation (6) via minimum cost flow problem in Figure 4(a) and Figure 4(b), respectively. [15] solves discrete optimization problem k times sequentially with sibling penalty updating $\mathcal{S}_z^v = \left\{(p,q) \mid \mathbf{z}_p^w = \mathbf{z}_q^w,\ \forall w = 1, \ldots, v-1\right\}$ from the previous hash codes. Concretely, $\mathbf{h}_i = \mathbf{z}_p$ where $y_i = p$.

$$\underset{\mathbf{z}_1,\ldots,\mathbf{z}_{n_c}}{\text{minimize}} \underbrace{\sum_{p=1}^{n_c} -\mathbf{c}_p^\intercal \mathbf{z}_p}_{\text{unary term}} + \underbrace{\sum_{p\neq q} \mathbf{z}_p^\intercal P \mathbf{z}_q}_{\text{orthogonality}} \quad (5) \qquad \text{subject to } \mathbf{z}_p \in \{0,1\}^d, ||\mathbf{z}_p||_1 = k_s,\ \forall p$$

$$\underset{\mathbf{z}_1,\ldots,\mathbf{z}_{n_c}}{\text{minimize}} \underbrace{\sum_{p=1}^{n_c} -(\mathbf{c}_p^v)^\intercal \mathbf{z}_p}_{\text{unary term}} + \underbrace{\sum_{\substack{(p,q)\in\mathcal{S}_z^v \\ p\neq q}} \mathbf{z}_p{}^\intercal Q \mathbf{z}_q}_{\text{sibling penalty}} + \underbrace{\sum_{p\neq q} \mathbf{z}_p{}^\intercal P \mathbf{z}_q}_{\text{orthogonality}} \quad (6)$$

$$\text{subject to } \|\mathbf{z}_p\| = \begin{cases} 1 & \forall v \neq k \\ k_s & v = k \end{cases},\ \mathbf{z}_p \in \{0,1\}^d,\ \forall p$$

[6] We omit v in [14].

Given hash codes, the parameters for deep neural network, $\boldsymbol{\theta}$, are updated via the state-of-the-art metric learning losses such as triplet loss with semi-hard negative mining[7] and Npairs loss[8] where the distance between each data is defined with continuous representation and hash codes as in $d_{ij}^{v} = \| \left(\mathbf{h}_i^v \vee \mathbf{h}_j^v\right) \odot \left(f(\mathbf{x}_i;\boldsymbol{\theta})^v - f(\mathbf{x}_j;\boldsymbol{\theta})^v\right) \|_1$ [6].

Table 2: Results with Triplet network with hard negative mining and Npairs network. Querying test data against a hash table built on *test* set and a hash table built on *train* set on Cifar-100.

		Triplet								Npairs							
		test				*train*				*test*				*train*			
	Method	SUF	Pr@1	Pr@4	Pr@16	SUF	Pr@1	Pr@4	Pr@16	SUF	Pr@1	Pr@4	Pr@16	SUF	Pr@1	Pr@4	Pr@16
k_S	Metric	1.00	56.78	55.99	53.95	1.00	62.64	61.91	61.22	1.00	57.05	55.70	53.91	1.00	61.78	60.63	59.73
1	Th	41.21	54.82	52.88	48.03	43.19	61.56	60.24	58.23	12.72	54.95	52.60	47.16	13.65	60.80	59.49	57.27
	VQ	22.78	56.74	55.94	53.77	40.35	62.54	61.78	60.98	34.86	56.76	55.35	53.75	31.35	61.22	60.24	59.34
	[14]	97.67	57.63	57.16	55.76	97.77	63.85	63.40	63.39	54.85	58.19	57.22	55.87	54.90	**63.11**	62.29	61.94
	[15]	97.67	**58.42**	**57.88**	**56.58**	97.28	**64.73**	**64.63**	**64.69**	**101.1**	**58.28**	**57.79**	**56.92**	**97.47**	63.06	**62.62**	**62.44**
2	Th	14.82	56.55	55.62	52.90	15.34	62.41	61.68	60.89	5.09	56.52	55.28	53.04	5.36	61.65	60.50	59.50
	VQ	5.63	56.78	56.00	53.99	6.94	62.66	61.92	61.26	6.08	57.13	55.74	53.90	5.44	61.82	60.56	59.70
	[14]	76.12	57.30	56.70	55.19	78.28	63.60	63.19	63.09	16.20	57.27	55.98	54.42	16.51	61.98	60.93	60.15
	[15]	**98.38**	**58.39**	**57.51**	**56.09**	**97.20**	**64.35**	**63.91**	**63.81**	**69.48**	**57.60**	**56.98**	**55.82**	**69.91**	**62.19**	**61.71**	**61.27**
3	Th	7.84	56.78	55.91	53.64	8.04	62.66	61.88	61.16	3.10	56.97	55.56	53.76	3.21	61.75	60.66	59.73
	VQ	2.83	56.78	55.99	53.95	2.96	62.62	61.92	61.22	2.66	57.01	55.69	53.90	2.36	61.78	60.62	59.73
	[14]	42.12	56.97	56.25	54.40	44.36	62.87	62.22	61.84	7.25	57.15	55.81	54.10	7.32	61.90	60.80	59.96
	[15]	**94.55**	**58.19**	**57.42**	**56.02**	**93.69**	**63.60**	**63.35**	**63.32**	**57.09**	**57.56**	**56.70**	**55.41**	**58.62**	**62.30**	**61.44**	**60.91**
4	Th	4.90	56.84	56.01	53.86	5.00	62.66	61.94	61.24	2.25	57.02	55.64	53.88	2.30	61.78	60.66	59.75
	VQ	1.91	56.77	55.99	53.94	1.97	62.62	61.91	61.22	1.66	57.03	55.70	53.91	1.55	61.78	60.62	59.73
	[14]	16.19	57.11	56.21	54.20	16.52	62.81	62.14	61.58	4.51	57.15	55.77	54.01	4.52	61.81	60.69	59.77
	[15]	**92.18**	**58.52**	**57.79**	**56.22**	**91.27**	**64.20**	**63.95**	**63.63**	**49.43**	**57.75**	**56.79**	**55.50**	**50.80**	**62.43**	**61.65**	**61.01**

The proposed models show the experiment results with speed up factor(SUF), precision@k(pr@k), and normalized mutual information(NMI) on Cifar-100[9] and Imagenet[10]. Table 2 shows the results from the triplet network and the Npairs network on Cifar-100. The results show that [14] and [15] not only outperform search accuracies of the state-of-the-art deep metric learning based models but also provide the superior speedup over other baselines.
The source code for [14] is published on [16].

5 Medical Imaging and Intelligent Reality Lab., Asan Medical Center, University of Ulsan College of Medicine

5.1 Pathologic Diagnosis System for Renal Allograft Rejection

Complement degradation product (C4d) is a sensitive and specific marker for antibody-mediated allograft rejection. Evaluation of C4d expression on peritubular capillary of renal allografts is labor-intensive and subject to observer variability. Using CNN-based classification and detection algorithm, we proposed not only a fully-automated system to predict allograft rejection but also novel ways to improve detection performance by using enlarging mask and AI-assisted labeling technique [17] (Figure 5(a)).

5.2 Frozen Section Analysis for Breast Cancer

Assessing the status of metastasis in sentinel lymph nodes (SLNs) by pathologists is an essential task for the accurate staging of breast cancer. However, histopathological evaluation of SLNs by a pathologist is not easy, tedious and time-consuming task especially using frozen tissue sections. We made a fully automated system to predict not only whether digital slide has metastasis or not, but also localize the active tumor regions using an ensemble of CNN-based classification models in real-time.

[7] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015.

[8] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *NeurIps*, 2016.

[9] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-100 (Canadian Institute for Advanced Research). 2009.

[10] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

5.3 Detection and Segmentation of Tools and Anatomical Landmarks on the Surgical Videos

Although it is important to teach and guide unskilled surgeons, there are not enough number of experts. A deep learning based real-time automated surgery guidance system can be a solution to this prevalent problem. In this study, we performed the detection of various kinds of surgical tools and segmentation of anatomical landmarks on the surgery videos as the first step for developing AI surgery guide. Using 70 videos of simple mastoidectomy surgery acquired from the Asan Medical Center, we succeeded in the detection of five surgical tools and segmentation of five anatomic landmarks.

5.4 Pancreas Segmentation using Domain Adaptation

Segmentation using deep learning has been evolving and showing high performance in many studies. Despite strong performances on segmentation using deep neural networks, it gave different results depending on where the data was obtained due to differing supplies and parameters used in the process. So, we proposed and evaluated a domain adaptation strategy based on generative adversarial learning for pancreas segmentation in the multi-center abdominal CT scans [18] (Figure 5(b)).

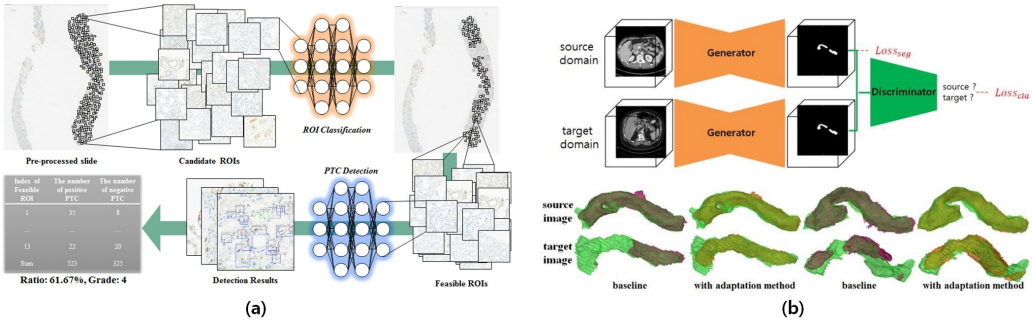


Figure 5: (a) Pathologic diagnosis system for renal allograft rejection (b) Pancreas segmentation using domain adaptation

5.5 Brain Infarct Segmentation

In this study, we proposed a semantic segmentation model with Squeeze-and-Excitation (SE) block for infarct segmentation in diffusion-weighted imaging [19, 20] (Figure 6(a)). Our method has an advantage to enhance channel-wise information with feature maps in each semantic segmentation model. We compared various 2D networks with and without SE blocks. This method showed the best performance than those of other semantic segmentation networks in case of 2D. Applying SENet to segmentation networks could be applied to various kinds of medical segmentation.

5.6 Traumatic Brain Injury Detection

We proposed a fully automated detection system using deep neural networks for brain injury patients in emergency rooms [21] (Figure 6(b)). Weakly labeled data could lead to training failure due to high complexity and dimensionality problems. To address these problems, we collected and used additional 169 patient data with information on areas of fracture and hemorrhage. Using this hard labeled data, 3D patch images were extracted and the model was trained under weak supervision. After that, the network was fine-tuned using weak supervision with a relatively large amount of data. Our proposed method helps radiologists and physicians in emergency rooms to reducing the diagnosis time and human errors.

5.7 Pattern Classification of Diffuse Interstitial Lung Disease

We developed a general data augmentation strategy using Perlin noise, applying it to pixel-by-pixel image classification and quantification of six typical disease patterns of diffuse interstitial lung disease [22] (Figure 6(c)). For supervised learning, one of the most challenging issues is a paucity of human-labeled data and its expenses, especially in medical research. To overcome this problem, we developed a technique to make a theoretically infinite number of labeled datasets, achieving around

10% performance improvement than the conventional augmentation. This data augmentation strategy using Perlin noise could be widely applied to deep learning studies for image classification and segmentation, especially in cases with relatively small datasets.

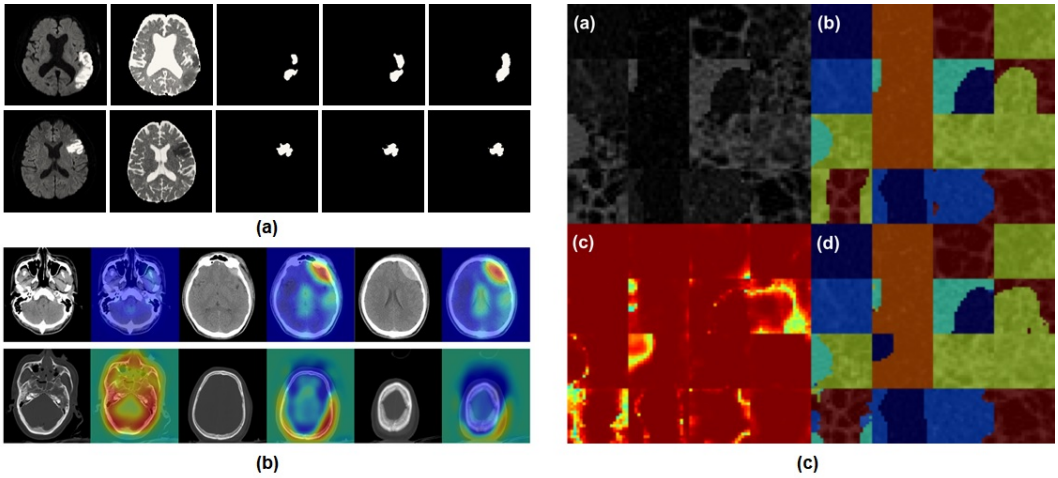


Figure 6: (a) Brain infarct segmentation (b) Traumatic brain injury detection (c) Pattern classification of diffuse interstitial lung disease

5.8 Reproducibility Analysis on Computer Aided Abnormality Detection

The chest PA X-ray images could be used for computer-aided diagnosis (CAD) to improve the radiologist's reading quality. However, reproducibility of CAD has not been studied intensively yet in chest PA X-ray images of the same patient acquired within a short-term period. We evaluated the reproducibility of CAD for classifying and detecting 5-class abnormalities in chest X-ray including nodule, consolidation, interstitial opacity, pleural effusion, and pneumothorax [23].

5.9 CT Kernel Conversions

In a clinical setting, different reconstruction of a CT image is required for a different task. We developed a CT kernel conversion technique based on CNNs that can produce images with the desired kernel from an existing reconstructed image with a specific kernel in the absence of the raw data (sinogram) [24]. We applied a simplified SE in the layers of CNN to efficiently learn the noise distribution. Progressive learning with auxiliary loss functions corresponding to intermediate kernel losses enables the conversion from an image reconstructed with a very sharp kernel to one reconstructed with a very smooth kernel and vice versa. Our technique outperformed previously developed denoising techniques in term of SSIM and MSE.

6 Music & Audio Research Group, Seoul National University

6.1 Chord Generation

We first studied the chord generation task, which deals with symbolic data to understand the structural nature of music. The chord generation is a task for generating a chord sequence by inferring matching harmonics for a given monophonic melody. Many previous studies have used HMM-based methods, but they have the disadvantage that they are heavily influenced by the frequency of chords in the training data. Therefore, the chord generation result of the HMM-based method has a limitation that only a dominant tonic chord is generated in one song. We have designed a system that can predict more harmonic structure by using BLSTM network in [25]. The proposed network in both quantitative and qualitative evaluation showed better performance than HMM and DNN-HMM networks. Listeners also tend to prefer chord sequences generated by our proposed network.

6.2 Cover Song Identification

Recently, various cover song contents have been produced and distributed due to the development of personal media. In order to prevent copyright problems, an effective cover song identification algorithm is needed. We have studied the cover song identification algorithm as a basic research to define and compare the similarity between music. In [26], we proposed a cover song identification algorithm that uses a method of classifying cross similarity matrices between two songs using convolutional neural network (CNN). We observed a diagonal shape in the cross similarity matrix due to melody similarity when the two songs were in the cover relationship, and used the CNN based classification method to recognize it effectively. In addition, In [27], we proposed a ranking algorithm that enables higher performance cover song identification using a representation vector based on the cover-probability between two songs from CNN based cover classifier. The proposed model showed 12.8%p improvement over the existing state-of-the-art algorithm, and the first identification performance for the mixed dataset was recorded in the MIREX cover song identification task conducted in 2018 [28]. As a follow-up study, we confirmed that the performance of identifying the cover song was deteriorated by the transpositioned tone in the middle of the song. To prevent this, we propose a preprocessing algorithm that predicts and corrects the transposition with patch-wise. We found that when adding arbitrary transpositions to an existing cover song dataset, the performance of identifying existing cover songs was significantly degraded. In addition, we have shown that the performance degradation does not occur when the proposed transpose correction algorithm is applied. These studies have significance in that they proposed the highest level of cover song identification performance and it is meaningful that it enabled a deep learning based approach for the similarity definition problem between two sequences of music.

6.3 Music Information Retrieval

Recent advances in deep learning have shown to be very successful in many application areas, and music information retrieval (MIR) is not an exception. However, there is still limited availability in the MIR since the performance of the deep learning-based algorithms is highly dependent on the amount of annotated training data. The process of labeling and verifying music data requires more time and effort than other types of data, such as images. Furthermore, the fact that one music piece can have multiple attributes or labels make the process even more difficult. To address this problem, we proposed knowledge transfer methods from large-scale auto-labeled data [29]. We used the user's history in the music streaming service as the training data. The user's listening history accumulates continuously without human effort. These logs define the characteristics of each song, which is one of the key features used in the music recommendation system and also can be inferred from the audio data[11]. It has been confirmed that cross-domain knowledge transfer is possible in experiments using various knowledge transfer methods. In particular, the proposed distillation losses have advantages in terms of expandability.

6.4 Music Recommendation

We studied music recommendation task based on music information retrieval. Due to technological advances, the market trends in music content consumption have become digitized and mobilized, and users are able to enjoy music without time and space constraints. Increased accessibility has created difficulties in selection of music content, and there has been a growing demand for recommender systems. In terms of real-world recommendation applications, it is important to provide an explainable recommendation. Providing explanations usually builds user's trust in recommender system and its results. Lyrics contain a wealth of information about the content, mood, genre, and style of music, as they contain a variety of words that reflect the emotion and mood of the music. In [30], we proposed an explainable content feature extraction algorithm that uses lyrics and music signals. In order to extract explainable features of lyrics and music signals, we use self-attentive genre classification and knowledge distillation of pre-trained models. As a result of visualizing attention weights of lyrics, there is a tendency to focus on words that have an important role in genre classification. Visualization of the attention weights of music signals shows a tendency to focus on the beginning of the vocal part. Retrieved similar songs from lyrics include cover songs and have similar lyrical mood and theme. Retrieved similar songs from music signals include similar genre, similar artists and similar vocal

[11] Aaron Van den Oord, Sander Dieleman, and Benjamin Schrauwen. Deep content-based music recommendation. In *NeurIps*, 2013.

timbre. In further research, we extracted two additional features lyrics and music signals in terms of cold-start recommendation, and found that the methods using the additional features outperformed the existing method[11] in terms of mAP, NDCG, recall@N. Our work has significance in that we found the efficiency of lyrics in cold-start recommendation and indicated acoustic signals are as good features as lyrics in terms of interpretation, and both of them reflect the details that they do not reflect each other.

6.5 Sequential Skip Prediction

Today, a huge number of active users are interacting within a database of audio tracks in online music streaming services such as *Melon*[12] and *Spotify*[13]. While conventional music recommender systems[11,14] have been successful in general, the problem of personalization[15] with few-shot of user behavioral data has not been discovered well. The *WSDM Cup 2019 Spotify Challenge*[16] tackled this issue by defining a new task: predicting whether users would skip tracks or not, given their immediately preceding interactions in their listening session. In [31], we proposed two different meta-learning approaches to solve the sequential skip prediction task, based on 1) metric learning, and 2) sequence learning.

- In metric learning-based approach, one key feature was that they do not assume the presence of orders in a sequence. This allowed us to formulate the skip prediction problem in a similar way with the previous work[17] on few-shot learning that *learns to compare*.
- In sequence learning-based approach, temporal convolution was employed to learn or memorize information by assuming the presence of orders in a sequence. In this fashion, we formulated the skip prediction problem as a meta-learning[18] that *learns to refer to past experience*.

In the work [31], the above-mentioned approaches were evaluated using a real-world dataset[19]. The total number of user-interaction logs was larger than one billion. The main results showed that the sequence learning-based approaches consistently outperformed the metric learning by at least 5.9%p, in terms of *mean average accuracy* defined by the cup challenge. Our final model for sequence learning-based approach was more efficient than other models with attention units[18,20]. The consistently poorer results of metric learning-based approach implied the importance of sequential information in this task. In additional experiments, we verified that giving a complete information to the query set could improve the prediction accuracy by about 20%p.

These findings implicate us a promising future direction [31] for "learning to teach": a meta-learner that learns to distill knowledge from a model trained with completely labeled inputs is expected to leverage the performance of existing model trained with incomplete inputs, in various few-shot learning tasks.

7 Probability and Optimization with Applications Lab., KAIST

Although stochastic gradient descent (SGD) is a major method to train a deep neural network, our understanding of SGD dynamics is limited due to its high dimensionality. In recent years, some

[12] `https://www.melon.com/`

[13] `https://www.spotify.com/`

[14] Oscar Celma. Music recommendation. *In Music Recommendation and Discovery*, pages 43–85. Springer, 2010.

[15] Yoon Ho Cho, Jae Kyeong Kim, and Soung Hie Kim. A personalized recommender system based on web usage mining and decision tree induction. *Expert Systems with Applications*, 23(3):329–342, 2002.

[16] `https://www.crowdai.org/challenges/spotify-sequential-skip-prediction-challenge`

[17] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *CVPR*, 2018.

[18] Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. A simple neural attentive meta-learner. In *ICLR*, 2018.

[19] Brian Brost, Rishabh Mehrotra, and Tristan Jehan. The music streaming sessions dataset. In *The Web Conference*, 2019.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIps*, 2017.

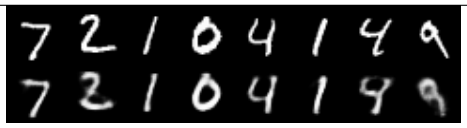
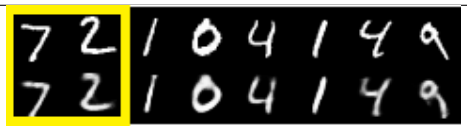
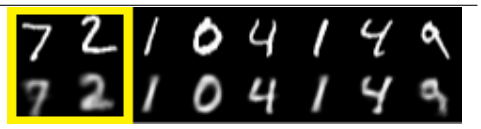
researchers have measured the stochasticity of minibatch gradients to better characterize the learning dynamics of SGD [32]. They have focused on the directional stochasticity of minibatch gradients and measured this by the concentration parameter of the von Mises-Fisher distribution (vMF), κ. The vMF approximation of this paper on the directions of minibatch gradients is original, and this work theoretically and empirically shows that the directional uniformity of minibatch gradients increases over the course of SGD. By doing so, the paper provides some insights warning the importance about the directions of stochastic gradients.

In recent studies, some researchers have been interested in applications of hyperbolic spaces because of their tree-likeliness and the exponential growth of capacity. In order to verify the usefulness of the hyperbolic space in the context of deep learning, a variational autoencoder (VAE) with the hyperbolic latent space for the MNIST dataset has been implemented [33]. This is trained under the following loss function:

$$\mathcal{L} = -E_{z\sim q_\theta(z|x)}[\log p(x|z)] + D_{KL}(q_\theta(z|x)\|p(z)) - \sum_{(x,\, y)\in\mathcal{D}} \log \frac{e^{-d(x,\, y)}}{\sum_{n\in\mathcal{N}_x} e^{-d(x,\, n)}}.$$

Here, the first two terms are usual VAE losses , and the third term is the modified ranking loss that is introduced in the previous research[21]. In their implementation, they consider the positive pair set ($\mathcal{D}$) as the set of same number pairs, and $\mathcal{N}_x = \{s : (x, s) \notin \mathcal{D}\}$ as the set of negative examples for x. $d(\cdot,\, \cdot)$ is the hyperbolic distance function which gives the main difference of geometry in latent space. It is shown that the VAE with the hyperbolic latent space (Hyperbolic VAE) reconstructs better images than the usual VAE when the latent space is low dimensional (Table 3).

Table 3: Sample comparisons between VAE and Hyperbolic VAE with the ranking loss across various dimensions. The first and the second lines of each cell are input data and corresponding reconstructed outputs, respectively. The hyperbolic VAE produces better images in low dimensional latent spaces (red boxes), but it struggles to reconstruct clear images in the high dimensional latent space (yellow boxes).

Dim.	VAE	Hyperbolic VAE
2		
10		
20		

8 Vision and Learning Lab., Seoul National University (SNUVL)

We mainly focused on natural language processing, including (i) generating sentences from various types of contexts (*e.g.* image, text and audio) [34, 35, 36], (ii) improving variational autoencoder (VAE) for conversation modeling [37], and (iii) analyzing neural networks to understand how it works on NLP tasks [38].

We explain our papers with brief summary.

Towards Personalized Image Captioning via Multimodal Memory Networks, IEEE TPAMI 2018 [34, 39] (Figure 7(a)). We address personalized image captioning, which generates a descriptive sentence for a user's image, accounting for prior knowledge such as her active vocabulary or writing

[21] Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In *NeurIps*, 2017.

style in her previous documents. As applications of personalized image captioning, we solve two post automation tasks in social networks: *hashtag prediction* and *post generation*. The hashtag prediction predicts a list of hashtags for an image, while the post generation creates a natural text consisting of normal words, emojis, and even hashtags. We propose a novel personalized captioning model named *Context Sequence Memory Network* (*CSMN*). Its unique updates over existing memory networks include (i) exploiting memory as a repository for multiple types of context information, (ii) appending previously generated words into memory to capture long-term information, and (iii) adopting CNN memory structure to jointly represent nearby ordered memory slots for better context understanding. For evaluation, we collect a new dataset InstaPIC-1.1M, comprising 1.1M Instagram posts from 6.3K users. We further use the benchmark YFCC100M dataset[22] to validate the generality of our approach. With quantitative evaluation and user studies via Amazon Mechanical Turk, we show that the three novel features of the CSMN help enhance the performance of personalized image captioning over state-of-the-art captioning models.

A Hierarchical Latent Structure for Variational Conversation Modeling, NAACL-HLT 2018 [37, 40] (Figure 8(a)). Variational autoencoders (VAE)[23] combined with hierarchical RNNs have emerged as a powerful framework for conversation modeling. However, they suffer from the notorious degeneration problem, where the decoders learn to ignore latent variables and reduce to vanilla RNNs. We empirically show that this degeneracy occurs mostly due to two reasons. First, the expressive power of hierarchical RNN decoders is often high enough to model the data using only its decoding distributions without relying on the latent variables. Second, the conditional VAE structure whose generation process is conditioned on a context, makes the range of training targets very sparse; that is, the RNN decoders can easily overfit to the training data ignoring the latent variables. To solve the degeneration problem, we propose a novel model named Variational Hierarchical Conversation RNNs (VHCR), involving two key ideas of (1) using a hierarchical structure of latent variables, and (2) exploiting an utterance drop regularization. With evaluations on two datasets of Cornell Movie Dialog[24] and Ubuntu Dialog Corpus[25], we show that our VHCR successfully utilizes latent variables and outperforms state-of-the-art models for conversation generation. Moreover, it can perform several new utterance control tasks, thanks to its hierarchical latent structure.

[22]Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. YFCC100M: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

[23]Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

[24]Cristian Danescu-Niculescu-Mizil and Lillian Lee. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *CMCL Workshop*, 2011.

[25]Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *SIGDIAL*, 2015.

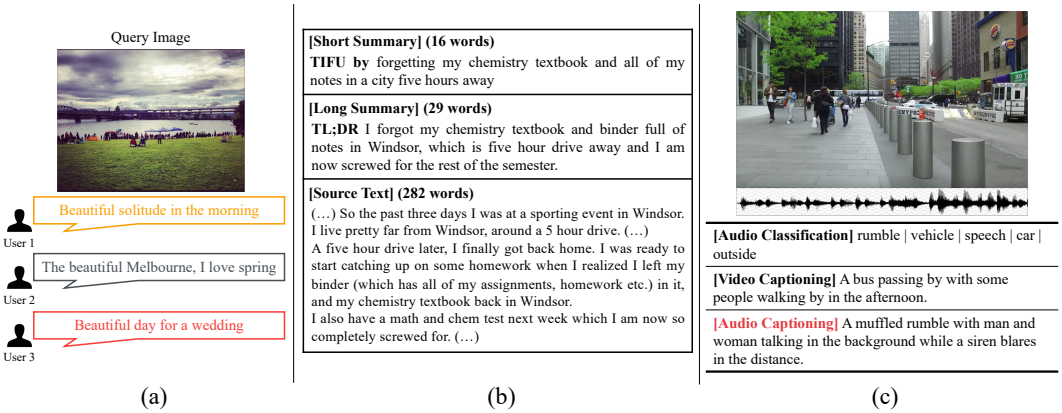
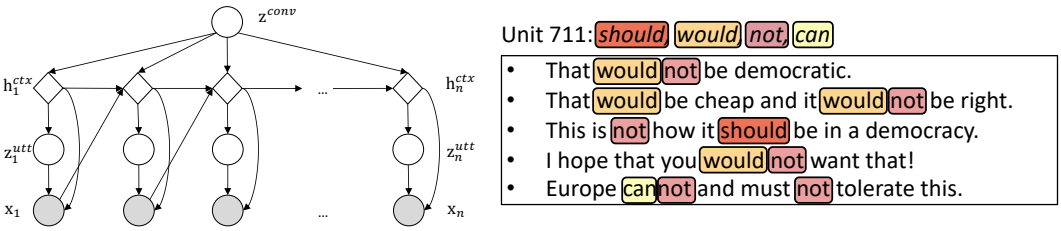


Figure 7: (a) **Personalized Image Captioning [34]**: Personalized image captioning is motivated by that different users are likely to generate different sentences for the same image, according to their own experiences, thoughts, or writing styles. (b) **Reddit Posts Summarization [35]**: An example (source text)-(short/long summary) pair of the *Reddit TIFU* dataset (c) **Audio Captioning [36]**: Comparison of audio captioning with audio classification and video captioning tasks.



(a) (b)

Figure 8: (a) **Hierarchical Latent Structure for Conversation [37]**: Graphical representation of the Variational Hierarchical Conversation RNN (VHCR). The global latent variable $\mathbf{z}^{conv}$ provides a global context in which the conversation takes place. (b) **Language Concepts in Individual Units [38]**: Most activated sentences and aligned concepts to the units in hidden representations of deep convolutional networks. Aligned concepts appear frequently in most activated sentences, implying that those units respond selectively to specific natural language concepts.

Abstractive Summarization of Reddit Posts with Multi-level Memory Networks, NAACL-HLT 2019 [35, 41] (Figure 7(b)). We address the problem of abstractive summarization in two directions: proposing a novel dataset and a new model. First, we collect *Reddit TIFU* dataset, consisting of 120K posts from the online discussion forum Reddit. We use such informal crowd-generated posts as text source, in contrast with existing datasets that mostly use formal documents as source such as news articles. Thus, our dataset could less suffer from some biases that key sentences usually locate at the beginning of the text and favorable summary candidates are already inside the text in similar forms. Second, we propose a novel abstractive summarization model named *multi-level memory networks* (MMN), equipped with multi-level memory to store the information of text from different levels of abstraction. With quantitative evaluation and user studies via Amazon Mechanical Turk, we show the *Reddit TIFU* dataset is highly abstractive and the MMN outperforms the state-of-the-art summarization models.

AuidoCaps: Generating Captions for Audio in The Wild, NAACL-HLT 2019 [36] (Figure 7(c)). We explore the problem of *audio captioning*: generating natural language description for any kind of audio in the wild, which has been surprisingly unexplored in previous research. We contribute a large-scale dataset of 46K audio clips with human-written text pairs collected via crowdsourcing on the AudioSet dataset[26]. Our thorough empirical studies not only show that our collected captions are indeed faithful to audio inputs but also discover what forms of audio representation and captioning models are effective for the audio captioning. From extensive experiments, we also propose two novel components that help improve audio captioning performance: the top-down multi-scale encoder and aligned semantic attention.

Discovery of Natural Language Concepts in Individual Units of CNNs, ICLR 2019 [38] (Figure 8(b)). Although deep convolutional networks have achieved improved performance in many natural language tasks, they have been treated as black boxes because they are difficult to interpret. Especially, little is known about how they represent language in their intermediate layers. In an attempt to understand the representations of deep convolutional networks trained on language tasks, we show that individual units are selectively responsive to specific morphemes, words, and phrases, rather than responding to arbitrary and uninterpretable patterns. In order to quantitatively analyze such an intriguing phenomenon, we propose a concept alignment method based on how units respond to the replicated text. We conduct analyses with different architectures on multiple datasets for classification and translation tasks and provide new insights into how deep models understand natural language.

9 Conclusion

This paper summarized selected research results from seven CSI affiliated laboratories during the period from April 2017 to March 2019. Our achievements include the publications in journals and

[26] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *ICASSP*, 2017.

conference proceedings, open source software contributions, several competition participations, as well as active research collaborations with Kakao and Kakao Brain engineers. Furthermore, our contributions were not just limited to application of deep learning models to real-world problems, but also involved advancement of deep learning theories.

As we look back on the past, we are excited about the depth and breadth of what CSI has accomplished. In the future, we look forward to making even greater impact on the deep learning research community. Thanks to the generous support from Kakao and Kakao Brain corporations, we were able to set solid foundations for carrying out state-of-the-art deep learning research and also for educating deep learning experts from several different disciplines. More importantly, the experiences gained from the CSI activities helped us to build a firm basis to make substantial progress for the years to come.

Acknowledgments

This work was supported by Kakao and Kakao Brain corporations.

References

[1] Jonghwan Mun, Kimin Lee, Jinwoo Shin, and Bohyung Han. Learning to specialize with knowledge distillation for visual question answering. In *NeurIps*, 2018.

[2] Hyeonwoo Noh, Taehoon Kim, Jonghwan Mun, and Bohyung Han. Transfer learning via unsupervised task discovery for visual question answering. In *CVPR*, 2019.

[3] Moon-jung Chae, Kyubyong Park, Jinhyun Bang, Soobin Suh, Jonghyuk Park, Namju Kim, and Jonghun Park. Convolutional sequence to sequence model with non-sequential greedy decoding for grapheme to phoneme conversion. In *ICASSP*, 2018.

[4] Moon-jung Chae, Kyubyong Park, Jinhyun Bang, Soobin Suh, Jonghyuk Park, Namju Kim, and Jonghun Park. Convolutional seq2seq model with non-sequential greedy decoding. `https://github.com/ctr4si/NSGD_G2P`, 2018.

[5] Heewoong Park, Sukhyun Cho, Kyubyong Park, Namju Kim, and Jonghun Park. Training utterance-level embedding networks for speaker identification and verification. In *Interspeech*, 2018.

[6] Soobin Suh, Dabi Ahn, Heewoong Park, and Jonghun Park. Voice conversion with diverse intonation using conditional variational auto-encoder. In *MLSLP*, 2018.

[7] Moon-jung Chae, Jonghyuk Park, Jinhyun Bang, and Jonghun Park. Improving TTS and ASR models with semi-supervised training. To be submitted in April 2019.

[8] Soobin Suh, Moon-jung Chae, Jonghyuk Park, Jinhyun Bang, and Jonghun Park. Tacotron: A fully end-to-end text-to-speech synthesis model. `https://github.com/ctr4si/Tacotron`, 2017.

[9] Heewoong Park, Sukhyun Cho, and Jonghun Park. Word RNN as a baseline for sentence completion. In *IEEE MNLP*, 2018.

[10] Heewoong Park, Sukhyun Cho, and Jonghun Park. Word RNN for sentence completion. `https://github.com/ctr4si/sentence-completion`, 2018.

[11] Soobin Suh, Kyubyong Park, and Jonghun Park. Bytenet with masking. `https://github.com/ctr4si/Bytenet`, 2017.

[12] Jonggwon Park, Kyoyun Choi, Sungwook Jeon, Dokyun Kim, and Jonghun Park. A bi-directional self-attention model for musical chord recognition. To be submitted in April 2019.

[13] Sungwook Jeon, Wan Huh, Kyoyun Choi, and Jonghun Park. Chord based melody generation using conditional variational autoencoder. To be submitted in April 2019.

[14] Yeonwoo Jeong and Hyun Oh Song. Efficient end-to-end learning for quantizable representations. In *ICML*, 2018.

[15] Yeonwoo Jeong, Yoonsung Kim, and Hyun Oh Song. End-to-end efficient representation learning via cascading combinatorial optimization. In *CVPR*, 2019.

[16] Yeonwoo Jeong and Hyun Oh Song. Efficient end-to-end learning for quantizable representations. `https://github.com/ctr4si/Deep-Hash-Table-ICML18`, 2018.

[17] Young-Gon Kim, Gyuheon Choi, Heounjeong Go, Yongwon Cho, Hyunna Lee, A-Reum Lee, Beomhee Park, and Namkug Kim. A fully automated system using a convolutional neural network to predict renal allograft rejection: Extra-validation with giga-pixel immunostained slides. *Scientific Reports*, 2019.

[18] Beomhee Park, Young Ji Song, Hyoung Jung Kim, Joon Beom Seo, and Namkug Kim. Robustness and accuracy enhancement of pancreas segmentation using domain adaptation with 3D U-Net between multi-center abdominal CT scans. In *KCR*, 2018.

[19] Ilsang Woo, A-Reum Lee, Hyunna Lee, Keummi Choi, Dong-Wha Kang, Seung Chai Jung, and Namkug Kim. Semantic segmentation with squeeze-and-excitation block: Application to infarct segmentation in DWI. In *Medical Imaging meets NeurIps*, 2017.

[20] A-Reum Lee, Ilsang Woo, Hyunna Lee, Seung Chai Jung, and Namkug Kim. Enhancement of brain infarct semantic segmentation with squeeze-and-excitation block in diffusion weigthed MRI. In *SPIE Medical Imaging*, 2019.

[21] A-Reum Lee, Beomhee Park, Younghwa Byeon, Jeong Hyun Lee, Gilsun Hong, and Namkug Kim. Detecting hemorrhage and fracture regions using 3D convolutional neural networks with strong and weak labels in head and neck CT of trauma patient in emergency rooms. In *RSNA*, 2018.

[22] Hyun-Jin Bae, Chang-Wook Kim, Namju Kim, BeomHee Park, Namkug Kim, Joon Beom Seo, and Sang Min Lee. A perlin noise-based augmentation strategy for deep learning with small data samples of HRCT images. *Scientific Reports*, 8(1):17687, 2018.

[23] Yongwon Cho, Young-Gon Kim, Sejin Park, Geunhwi Ahn, Eunsol Lee, Younghoon Cho, Sangmin B Lee, Joonbeom Seo, and Namkug Kim. Reproducibility analysis on computer aided abnormality detection with convolutional neural net in chest PA X-Ray images within short-term period. In *RSNA*, 2018.

[24] Da-in Eun, Ilsang Woo, Seonghee Ryu, Beomhee Park, Namkug Kim, Sang Min Lee, and Joon Beom Seo. CT kernel conversions using convolutional neural net for super-resolution with SE block. In *SPIE Medical Imaging*, 2019.

[25] Hyungui Lim, Seungyeon Rhyu, and Kyogu Lee. Chord generation from symbolic melody using BLSTM networks. *arXiv preprint arXiv:1712.01011*, 2017.

[26] Sungkyun Chang, Juheon Lee, Sang Keun Choe, and Kyogu Lee. Audio cover song identification using convolutional neural network. *arXiv preprint arXiv:1712.00166*, 2017.

[27] Juheon Lee, Sungkyun Chang, Sang Keun Choe, and Kyogu Lee. Cover song identification using song-to-song cross-similarity matrix with convolutional neural network. In *ICASSP*, 2018.

[28] Juheon Lee, Sungkyun Chang, Donmoon Lee, and Kyogu Lee. Covernet: Cover song identification using cross-similarity matrix with convolutional neural network. In *MIREX*, 2018.

[29] Donmoon Lee, Jaejun Lee, Jeongsoo Park, and Kyogu Lee. Enhancing music features by knowledge transfer from user-item log data. *arXiv preprint arXiv:1903.02794*, 2019.

[30] Seungjin Lee, Juheon Lee, and Kyogu Lee. Content-based feature exploration for transparent music recommendation using self-attentive genre classification. *arXiv preprint arXiv:1808.10600*, 2018.

[31] Sungkyun Chang, Seungjin Lee, and Kyogu Lee. Sequential skip prediction with few-shot in streamed music contents. *arXiv preprint arXiv:1901.08203*, 2019.

[32] Cheolhyoung Lee, Kyunghyun Cho, and Wanmo Kang. Directional analysis of stochastic gradient descent via von mises-fisher distributions in deep learning. *CoRR*, abs/1810.00150, 2018.

[33] Kyoungseok Jang and Wanmo Kang. Hyperbolic variational autoencoder with ranking loss. `https://github.com/ctr4si/Hyperbolic-VAE-with-Ranking-Loss`, 2018.

[34] Cesc Chunseong Park, Byeongchang Kim, and Gunhee Kim. Towards personalized image captioning via multimodal memory networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(4):999–1012, 2019.

[35] Byeongchang Kim, Hyunwoo Kim, and Gunhee Kim. Abstractive summarization of reddit posts with multi-level memory networks. In *NAACL-HLT*, 2019.

[36] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audio in the wild. In *NAACL-HLT*, 2019.

[37] Yookoon Park, Jaemin Cho, and Gunhee Kim. A hierarchical latent structure for variational conversation modeling. In *NAACL-HLT*, 2018.

[38] Seil Na, Yo Joong Choe, Dong-Hyun Lee, and Gunhee Kim. Discovery of natural language concepts in individual units of CNNs. In *ICLR*, 2019.

[39] Cesc Chunseong Park, Byeongchang Kim, and Gunhee Kim. Attend2u. `https://github.com/ctr4si/attend2u`, 2017.

[40] Yookoon Park, Jaemin Cho, and Gunhee Kim. Variational hierarchical conversation rnn. `https://github.com/ctr4si/A-Hierarchical-Latent-Structure-for-Variational-Conversation-Modeling`, 2018.

[41] Byeongchang Kim, Hyunwoo Kim, and Gunhee Kim. Multi-level memory networks. `https://github.com/ctr4si/MMN`, 2019.