



Semi-supervised Learning by Disentangling and Self-ensembling over Stochastic Latent Space

Prashnna Kumar Gyawali, Zhiyuan Li^(✉), Sandesh Ghimire, and Linwei Wang

Rochester Institute of Technology, Rochester, NY, USA
{pkg2182,z17904,sg9872,lxwast}@rit.edu

Abstract. The success of deep learning in medical imaging is mostly achieved at the cost of a large labeled data set. Semi-supervised learning (SSL) provides a promising solution by leveraging the structure of unlabeled data to improve learning from a small set of labeled data. Self-ensembling is a simple approach used in SSL to encourage consensus among ensemble predictions of unknown labels, improving generalization of the model by making it more insensitive to the latent space. Currently, such an ensemble is obtained by randomization such as dropout regularization and random data augmentation. In this work, we hypothesize – from the generalization perspective – that self-ensembling can be improved by exploiting the stochasticity of a disentangled latent space. To this end, we present a stacked SSL model that utilizes unsupervised disentangled representation learning as the stochastic embedding for self-ensembling. We evaluate the presented model for multi-label classification using chest X-ray images, demonstrating its improved performance over related SSL models as well as the interpretability of its disentangled representations.

Keywords: Semi-supervised learning · Self-ensembling · Disentangled representation learning

1 Introduction

While deep learning has seen tremendous successes in a variety of medical image analysis problems [9], many of these successes are achieved at the cost of a large labeled data set. Labeling a large collection of data however requires substantial human and financial resources, creating a primary hurdle to the wide-spread adoption of deep learning in clinical practice [9]. Semi-supervised learning (SSL) provides a promising solution by leveraging the structure of unlabeled data to improve the learning from a small set of labeled data [6, 8].

Among existing SSL methods, self-ensembling is a simple approach that encourages consensus among ensemble predictions of unknown labels [8]. Such

P. K. Gyawali and Z. Li—Both authors contributed equally to this work.

© Springer Nature Switzerland AG 2019

D. Shen et al. (Eds.): MICCAI 2019, LNCS 11769, pp. 766–774, 2019.

https://doi.org/10.1007/978-3-030-32226-7_85

ensemble predictions can be formed by randomization such as network regularization (*e.g.*, dropout) and random input augmentation [8]. As later rationalized in [5], based on the analytical learning theory, these randomization techniques are critical as they improve the generalization of the model by making it more *insensitive* to the latent space. From this theoretical perspective, it is natural to hypothesize that the design of this randomization will benefit from the knowledge of the latent space, especially its stochasticity. This however is not considered in existing works. The input augmentation functions, for instance, are typically hand-crafted considering random translations or rotations of images [8], with little consideration to the distribution of latent variables.

In parallel, advances in representation learning – especially that of the variational auto-encoder (VAE) – has allowed us to infer posterior distributions of the latent variables in an unsupervised manner [7]. In the classic semi-supervised deep generative model presented in [6], it has also been shown that such an unsupervised embedding can largely facilitate the subsequent SSL training by providing a disentangled and thereby more separable latent space [6].

In this paper, drawing on the analytical learning theory [5], we rationalize that (1) disentangling and (2) self-ensembling over the stochastic latent space will improve the generalization ability of the model. Based on this rationale, we investigate using unsupervised disentangled representation learning as the stochastic input embedding in self-ensembling. The presented SSL model consists of a VAE-based unsupervised embedding of the data, followed by a semi-supervised self-ensembling network utilizing the stochastic embedding as the inherent random augmentation of the inputs. We evaluate the presented SSL model on the recently open-sourced Chexpert data set for *multi-label* classification of thoracic disease using chest X-rays [4]. To demonstrate the benefits gained by exploiting the stochastic latent space in self-ensembling, we compare the performance of the presented method with the standard self-ensembling method considering different image-level input augmentation methods [8], VAE-based embedding with and without a subsequent deep generative SSL [6], along with a generative adversarial network (GAN)-based SSL [11]. We further qualitatively demonstrate the disentangled representation obtained via unsupervised embedding, and discuss its use for data analysis and model interpretability.

2 Related Work

This work is mostly related to two lines of research: (1) SSL based on regularization with random transformations, and (2) disentangled representation learning and its use in SSL. In the former, consistency-based regularization is applied on ensemble predictions obtained by randomization techniques such as random data augmentation, dropout, and random max-pooling [8]. This randomization was empirically shown to improve the generalization and stability of the SSL model, while its theoretical basis was recently shown to be related to the reduction of model sensitivity to the latent space [3, 5]. Motivated by this theory, in this work, we attempt to utilize the knowledge about the stochastic latent space – obtained in unsupervised learning – in this randomization process.

In the latter representation learning, deep neural networks have been combined with variational inference to jointly realize generative modeling of unlabeled data and posterior inference of latent variables [7]. Furthermore, the learned latent representations are encouraged to be semantically interpretable and mutually invariant, which is empirically shown to be useful for the downstream tasks [6]. For instance, an unsupervised VAE was used to provide a disentangled embedding (M1) for a subsequent VAE-based semi-supervised model (M2), commonly known as the M1+M2 model [6]. In this work, we make the first attempt to use the analytical learning theory to support the effect a disentangled embedding can have on the generalization ability of a model. Furthermore, we improve the M1+M2 model by replacing M2 with a self-ensembling SSL network, taking VAE's ability to model stochastic latent space to support self-ensembling.

Besides the approaches discussed above, there is also an active line of research in GAN-based SSL methods [10, 11]. The general idea is to add a classification objective to the original mini-max game and increases the capacity of the discriminator to associate the inputs to the corresponding labels. The presented work differs from this line of research by the emphasis on obtaining, regularizing, and interpreting the latent representations in SSL.

An increased interest in SSL has also been seen in medical image analysis. The use of an unsupervised representation learning for better generalization has been investigated for the task of myocardial segmentation [2]. In [10], SSL was used in a similar X-ray data set, although the scope was limited to binary classifications between normal and abnormal categories. To our knowledge, we present the first multi-label SSL that investigates disentangled learning and self-ensembling of stochastic latent space in medical image classification.

3 Model

We consider training data $\mathcal{D} = \mathcal{D}_l \cup \mathcal{D}_u$, where $\mathcal{D}_l = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{N_l}$ is the labeled set and $\mathcal{D}_u = \{\mathbf{x}_j\}_{j=1}^{N_u}$ the unlabeled set. As outlined in Fig. 1, a stochastic latent space will be learned in an unsupervised and disentangled manner (Sect. 3.1), which will then be regularized via self-ensembling for SSL (Sect. 3.2).

3.1 Unsupervised and Disentangled Learning of Stochastic Latent Space

To disentangle and obtain the posterior distribution in the latent space, we first utilize a VAE for unsupervised learning of the generative factors in the data. We assume data \mathbf{x} to be generated by a likelihood $p_\theta(\mathbf{x}|\mathbf{z})$, involving a latent variable \mathbf{z} with a certain prior distribution $p(\mathbf{z})$. Given the intractability of exact posterior inference, a distribution $q_\phi(\mathbf{z}|\mathbf{x})$ is introduced to approximate the true posterior $p(\mathbf{z}|\mathbf{x})$ using variational inference [7]. Both the approximated posterior $q_\phi(\mathbf{z}|\mathbf{x})$ and the likelihood $p_\theta(\mathbf{x}|\mathbf{z})$ are parameterized by deep neural networks.

The VAE is trained by maximizing the variational evidence lower bound of the marginal likelihood of the training data with respect to parameters θ and ϕ :

$$\log p(\mathbf{x}) \geq \mathcal{L} = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - KL(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) \quad (1)$$

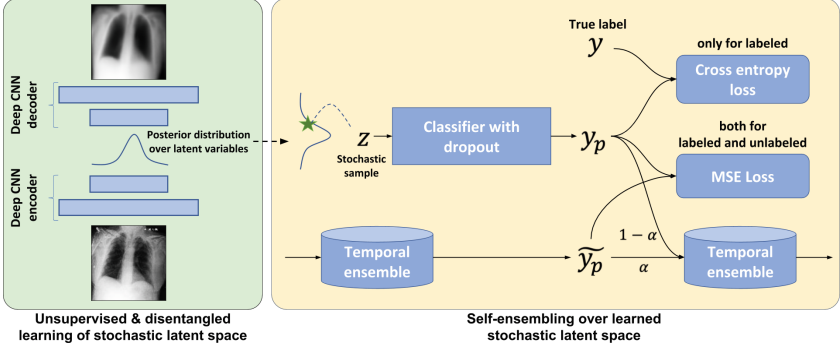


Fig. 1. Schematic diagram of the presented model.

where the first term can be interpreted as minimizing the reconstruction error, and the second term regularizes the learned posterior density $q_\phi(\mathbf{z}|\mathbf{x})$ by a prior $p(\mathbf{z})$ via the Kullback-Leibler (KL) divergence measure. We set the prior $p(\mathbf{z})$ to be an isotropic Gaussian which, through mutually independent latent dimensions, encourages disentangled latent representations in $q_\phi(\mathbf{z}|\mathbf{x})$ [7].

3.2 Regularizing the Stochastic Latent Space via Self-ensemble SSL

The unsupervised VAE embedding is then used as the input to a self-ensembling SSL model which, on unsupervised targets, applies coherence-based regularization to form consensus among ensemble predictions. For each training sample $\mathbf{x}^{(i)}$, its ensemble predictions are obtained from three sources:

1. Sampling from VAE-learned posterior density $q(\mathbf{z}^{(i)}|\mathbf{x}^{(i)})$. This utilizes a distribution learned from unlabeled data to replace the commonly-used hand-crafted augmentation functions to perturb $\mathbf{x}^{(i)}$ [8].
2. Network dropout that randomly neglects some units and utilizes a sub-network during training.
3. Temporal ensemble [8] achieved by accumulating the predicted label \mathbf{y}_p after every training epoch into an ensemble output $\tilde{\mathbf{y}}_p$ by $\tilde{\mathbf{y}}_p \leftarrow \alpha \tilde{\mathbf{y}}_p + (1 - \alpha)\mathbf{y}_p$, where α controls how far the ensemble reaches into training history.

Given each pair of ensemble predictions \mathbf{y}_p and $\tilde{\mathbf{y}}_p$, the network is trained, in each batch B , with the objective of minimizing the ensemble loss \mathcal{L}^e :

$$\mathcal{L}^e = \underbrace{\frac{1}{|B|} \sum_{n \sim (B \cap \mathcal{D}_l)} \sum_{l=1}^L [-y_{n,l} \log f(y_{n,p}|q(\mathbf{z}|\mathbf{x}))]}_{\text{only for labeled}} + \underbrace{\zeta \cdot \frac{1}{|B|} \sum_{n \sim B} \|\mathbf{y}_p - \tilde{\mathbf{y}}_p\|^2}_{\text{both for labeled and unlabeled}} \quad (2)$$

where the first term corresponds to the standard cross entropy loss and is evaluated for labeled data. The second term is evaluated for all data, encouraging consensus among ensemble predictions via mean squared loss. A ramp-up weighted function, starting from zero, is used for ζ as described in [8].

3.3 Stacked Self-ensembling SSL from the Generalization Perspective

In this section, we examine the generalization ability of the presented method – via the recently introduced analytical learning theory [5] – from two viewpoints: self-ensembling and disentangling. Theorem 1 in [5] provides an upper bound on the generalization gap (difference between expected and empirical error) Δ_g :

$$\Delta_g \leq V[f] \cdot \mathbb{D}^* \quad (3)$$

where $V[f]$ is the *variation* that computes how a function f varies in total w.r.t each small perturbation of every cross-combination of its variables, and \mathbb{D}^* is the *discrepancy* between the latent projections of an available data set \mathcal{D} and true data distribution. In our case, f is the composition of the loss function ℓ of coherence-based regularization and the mapping function f_y between stochastic latent sample \mathbf{z} and the prediction \mathbf{y}_p , i.e., $f = \ell \circ f_y$. Here, we rationalize how the presented method decreases the generalization gap by decreasing $V[f_y]$.

Self-ensembling: The second regularization term in (2) can be re-written over the input samples drawn from the posterior density $q(\mathbf{z}|\mathbf{x})$ from VAE embedding:

$$\ell = \int_{\mathbf{z}_1, \mathbf{z}_2} \|f_y(\mathbf{z}_1) - f_y(\mathbf{z}_2)\|^2 dP(\mathbf{z}_1, \mathbf{z}_2|\mathbf{x})$$

where $P(\mathbf{z}_1, \mathbf{z}_2|\mathbf{x}) = q(\mathbf{z}_1|\mathbf{x})q(\mathbf{z}_2|\mathbf{x})$. Minimizing this loss explicitly requires f_y to be more insensitive over the space of \mathbf{z} which implicitly minimizes $V[f_y]$ and thus the bound on the generalization gap.

Disentangling: Given representation $(\mathbf{z}_y, \mathbf{z}_o)$ where \mathbf{z}_y and \mathbf{z}_o are respectively the latent variables related and unrelated to \mathbf{y} . Based on [5], the *variation* $V[f_y]$ is minimal when the mapping f_y from latent space $(\mathbf{z}_y, \mathbf{z}_o)$ is invariant over \mathbf{z}_o . A disentangled latent space by design thereby reduces the generalization gap.

4 Experiments

Data and Model Architecture: We evaluated the presented model on the recently open-sourced Chexpert data set that has strong reference standards compared to other similar large-scale chest X-ray data set [4]. It consists of 224,316 chest radiographs from 65,240 patients, with labels for 14 pathology categories extracted from radiology reports. Given uncertainty labels provided for all images, we first removed all uncertain samples from the data set. We also removed all lateral-view samples. Small labeled training sets were then created by balancing among each disease category, ranging from 100 to 500 samples per category. Another 5000 and 50000 samples were randomly selected as the validation and test sets, while the rest were used as unlabeled training data. All images were re-sized to 128×128 in dimension.

The encoder of the VAE had five convolutional layers followed by three fully-connected (FC) layers. The output of the last FC layer was divided into the mean and log variance of the posterior distribution of the latent variables. The decoder is symmetrical to the encoder, with two FC layers followed by five transposed convolutional layers. ReLU activation was used for all the layers, except the last encoder layer that has no activation and the last decoder layer that used sigmoid activation. The self-ensembling network consisted of three FC layers with dropout layers ($p = 0.5$) in between. ReLU activation was used for the first two FC layers, and sigmoid activation for the last one. We use the Adam optimizer throughout all our experiments.

Table 1. Mean AUROC for multi-label classification for 14 pathology categories. trained with a fixed number of unlabeled data and a varying number of labeled data.

Model	Size of labeled data set (k)				
	100	200	300	400	500
VAE embedding	0.5853	0.6119	0.6331	0.6416	0.6556
VAE embedding + SSL (M1 + M2)	0.6080	0.6272	0.6340	0.6390	0.6491
Image-space self-ensemble (with noise)	0.6012	0.6277	0.6444	0.6550	0.6626
Image-space self-ensemble (with augmentation)	0.6089	0.6301	0.6423	0.6530	0.6617
ACGAN	0.5865	0.6036	0.6064	0.6284	0.6372
Latent-space self-ensemble (presented)	0.6200	0.6386	0.6484	0.6637	0.6697

4.1 Multi-label Semi-supervised Classification

We quantitatively evaluated the SSL performance of the presented model in comparison to existing models as summarized in Table 1. We first trained a classifier (with the same architecture as the self-ensembling network) using the unsupervised VAE embedding on labeled training data (*VAE embedding*). We then considered two most-related models to the presented work: (1) the M1+M2 SSL model where the unsupervised VAE embedding was used to support a subsequent VAE-based SSL model (*VAE embedding + SSL*) [6], and (2) the self-ensembling SSL (with the same architecture as the VAE encoder followed by self-ensembling in the stacked model) with two different types of image-level augmentation: adding Gaussian noises (with $std=0.15$) or random translation and rotation (maximum 12 pixels and 10°) to the images (*image-space self-ensembling*) [8]. To demonstrate how the presented work relates to GAN-based SSL methods, we also added a comparison to ACGAN [11] that was shown to generate globally coherent and discriminative samples assisting in SSL.

We tested all models on a varying number of labeled training data whilst keeping the number of unlabeled data fixed. Each model was tested on 10 different testing sets (5000 samples each). As shown, the three self-ensembling methods in general achieved better performance. Among the three self-ensembling methods, all the standard deviation is less than 0.008, and the improvement of the presented method over the two baseline self-ensembling methods is statistical significant ($p < 0.05$). This verified our hypothesis that, in comparison to

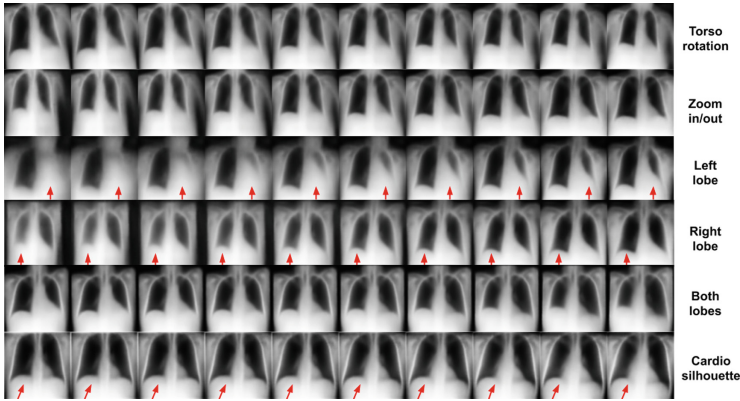


Fig. 2. Images generated by traversing over the $[-3, 3]$ range of a specific latent dimension, based on latent values inferred from a different seed image for each row.

ad-hoc image-level augmentations, utilizing the stochasticity of the latent space can improve the performance of self-ensembling.

It was surprising that ACGAN performed worse (except $k = 100$) than VAE embedding. We speculated that, unlike natural images where ACGAN has seen superior performance, disease-related factors in X-ray images may be more difficult to capture among other disease-irrelevant factors (see Sect. 5).

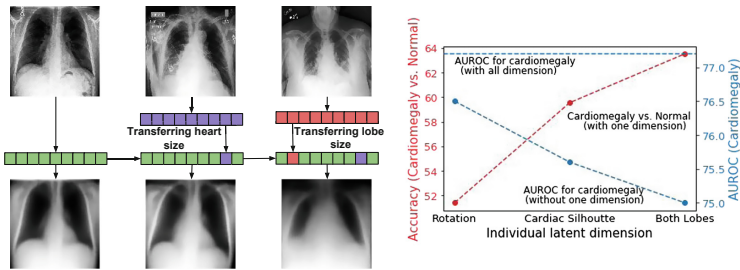


Fig. 3. [Best viewed in color] Left: demonstration of feature transfer by exchanging specific latent units. Top row: original images. Bottom row: reconstructed with original or exchanged latent codes. Right: Discriminative power of individual latent dimensions.

4.2 Interpreting the Disentangled Latent Representations

We then qualitatively evaluated the stochastic latent variables learned from the VAE. As shown in Fig. 2, as we traversed along a particular latent dimension and kept the others fixed, we were able to generate images reflecting changes in a particular semantic factor. With this knowledge, in Fig. 3 (left), we demonstrated

that it is possible to transfer specific features (such as heart size and lobe size) between X-ray images by swapping the corresponding latent units.

In an attempt to quantify how these unsupervised disentangled representations may affect downstream classification tasks, we considered the pathology of cardiomegaly as an example which is reflected as changes in heart-chest ratio in X-ray images [1]. We built a binary classifier between the category of cardiomegaly and no-finding, *considering only one of the learned latent units at a time*. We randomly sampled respectively 500, 1000 and 2000 images that had either cardiomegaly or no-finding labels for training, validation, and testing. As shown in Fig. 3 (right: red curve), the more a specific unit captured the heart-chest ratio, the more discriminative it was for detecting cardiomegaly. We also re-trained the presented model (for $k=500$) by removing one latent unit at a time, and evaluated the resulting AUROC for cardiomegaly. Similarly (Fig. 3 (right: blue curve)), the more a specific unit captured the heart-chest ratio, the larger its removal caused the drop in AUROC. These results suggest that improved disentangling in the latent representations may facilitate down-stream classification tasks as well as increase the interpretability of the results.

5 Conclusions and Discussion

We presented a stacked SSL method that uses unsupervised disentangling of the stochastic latent space as the input randomization in self-ensembling. From the analytical learning theory, we rationalized the effect of disentangling and self-ensembling over the latent space on the generalizability of the model. Empirically, we demonstrated both the quantitative improvement of the presented model in SSL and the interpretability of its disentangled representations.

We noted that, compared to many visual benchmark data sets, disease-specific factors in medical images may be buried by other more significant factors of variations in terms of contribution to pixel reconstruction or image distribution (*e.g.*, heart-chest ratio *vs.* torso shape). For instance, we attempted to remove the inactive dimensions (defined as $A_u < 10^{-2}$ where $A_u = Cov_x(\mathbb{E}_{u \sim q(u|x)}[u])$ for each dimension u in \mathbf{z}) from VAE embedding, a strategy shown to improve the performance of the M1+M2 model in visual benchmarks [6]. The mean AUROC of the presented model, however, decreased around 2% to 0.658 (for $k=500$). This, we believe, may explain the relatively limited progress of unsupervised representation learning in medical images despite its recent traction in other visual domains, a pressing challenge to be resolved in order to leverage unlabeled data in a field where image labeling is especially costly and difficult. For future work, we plan to improve two-stage training strategy and disentangling by hierarchical generative models.

Acknowledgement. This work is supported by NSF CAREER ACI-1350374 and NIH NHLBI R15HL140500.

References

1. Battler, A., et al.: The initial chest x-ray in acute myocardial infarction. Prediction of early and late mortality and survival. *Circulation* **61**(5), 1004–1009 (1980)
2. Chartsias, A., et al.: Factorised spatial representation learning: application in semi-supervised myocardial segmentation. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) *MICCAI 2018*. LNCS, vol. 11071, pp. 490–498. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00934-2_55
3. Ghimire, S., Gyawali, P.K., Dhamala, J., Sapp, J.L., Horacek, M., Wang, L.: Improving generalization of deep networks for inverse reconstruction of image sequences. In: Chung, A.C.S., Gee, J.C., Yushkevich, P.A., Bao, S. (eds.) *IPMI 2019*. LNCS, vol. 11492, pp. 153–166. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-20351-1_12
4. Irvin, J., et al.: CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison. In: *AAAI* (2019)
5. Kawaguchi, K., Bengio, Y., Verma, V., Kaelbling, L.P.: Generalization in machine learning via analytical learning theory. arXiv preprint [arXiv:1802.07426](https://arxiv.org/abs/1802.07426) (2018)
6. Kingma, D.P., Mohamed, S., Rezende, D.J., Welling, M.: Semi-supervised learning with deep generative models. In: *NeurIPS* (2014)
7. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: *ICLR* (2013)
8. Laine, S., Aila, T.: Temporal ensembling for semi-supervised learning. In: *ICLR* (2017)
9. Litjens, G., et al.: A survey on deep learning in medical image analysis. *Med. Image Anal.* **42**, 60–88 (2017)
10. Madani, A., Moradi, M., Karargyris, A., Syeda-Mahmood, T.: Semi-supervised learning with generative adversarial networks for chest x-ray classification with ability of data domain adaptation. In: *ISBI* (2018)
11. Odena, A., Olah, C., Shlens, J.: Conditional image synthesis with auxiliary classifier GANs. In: *ICML* (2017)