IMPROVING ULTRASOUND IMAGE CLASSIFICATION WITH LOCAL TEXTURE QUANTISATION

Xiao Li, Huizhi Liang

University of Reading, UK

Sidhartha Nagala, Jane Chen

Royal Berkshire Hospital, UK

ABSTRACT

Ultrasound image classification is important for disease diagnosis. It is more challenging than usual image classification tasks since ultrasound images are difficult to collect and usually contain lots of noise. This paper proposes a novel image classification framework for small-scaled and noisy ultrasound image datasets. The framework first transforms images into discrete *index grids*. The index grids use discrete indices encoding the local texture patterns of the images. Then, it will conduct classification based on index grids. The proposed framework can significantly reduce the impact of noise as well as the amount of training data that needed. Comparing with existing models, the proposed framework is a lite model and has better explainability. We evaluated the proposed approach on two public ultrasound image datasets for thyroid nodule classification and breast nodule classification. The experiment results show that the proposed approach achieves the new state-of-the-art. [1]

***Index Terms*—** Classification, lite model, quantisation, ultrasound image, thyroid nodule, breast nodule

1. INTRODUCTION

Ultrasound image analysis and classification play an important role in Computer-Aided Medical Systems. They can help radiologists quickly filter a large number of images, prevent radiologists from missing important information, reduce the need for radiologists to interpret images, and support disease diagnosis. They also can mitigate the problem of lacking high-quality medical equipment or well-trained radiologists in less-developed areas.

Ultrasound image classification is more challenging comparing to the usual image classification tasks. It has two main obstacles: 1) lack of large-scaled labelled datasets; 2) the images contain lots of noise. Collecting ultrasound images require expensive professional equipment and well-trained radiologists. Furthermore, annotating the labels of images is labour intensive. These issues limit the size of available labelled ultrasound image datasets. However, training deep learning models requires large labelled datasets. The lack of large labelled data is a bottleneck for the use of deep learning in ultrasound image analysis. Moreover, ultrasound images usually have low quality and contain lots of noise (e.g. speckle noise) [1]. They are difficult for human users to read and neural networks to process.

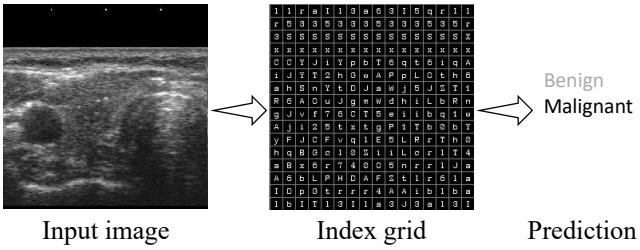

Fig. 1. The proposed model transform images to index grids, then makes the prediction according to the index grids.

Work [2, 3] collected large-scaled datasets and trained deep learning models on them; however, these datasets are not publicly available. Because of these limitations and challenges, many recent works [4, 5] used shallow models like SVM, KNN, and logistic regression rather than deep learning based models. Some other work [6] manually pre-processed the images to reduce noises before a neural network model. How to design novel deep learning based models for small and noisy ultrasound image datasets still need to be explored.

To bridge the gap, this paper proposes a novel image classification framework. The proposed framework adopts a two-step classification architecture. Inspired by computational linguistics whose input signals are represented as symbols [7, 8], first, our model transforms images into discrete *index grids*. Then, the classification is based on the *index grids* (Fig. 1). Index grids are grids of discrete indices encoding the local texture patterns of the images. When a classifier does the classification based on the index grids, it can concentrate on the global information. Since the vector-formed local texture information is replaced with the indices (the vector names) the local texture details mixed with noises are hidden. On the other hand, when a model predicts the local indices, it can concentrate on the local areas of the images. Since each image can be divided into many small local areas, the size of

Thanks Royal Berkshire NHS Foundation Trust for funding this project.

[1]The code is available at: https://github.com/lissomx/LTQ

the training data can be largely enriched and therefore it is effective to train deep learning models on small-scaled datasets. Therefore, index grids that enabling local texture quantisation of images can dramatically reduce the impact of noise as well as the amount of data needed for model training. Index grids also can facilitate image classifiers to adopt lite models.

2. PROPOSED METHOD

The proposed framework consists of three models, a *discrete-encoder*, a *classifier*, and a *recogniser* (Fig. 2). The discrete-encoder improves VQ-VAE [9, 10] to generate the index grids. The classifier is to predict the image labels based on the index grids, and the recogniser predicts the index grids.

The discrete-encoder adopts autoencoder-based framework (see Fig. 2), integrating the ResNet [11] and VQ-VAE [9, 10]. It encodes images (denoted by $\mathbf{x}$) with 256x256 resolution to 16x16 grids of latent vectors (i.e. $16 \times 16 = 256$ latent vectors, denoted by $\mathbf{h}_1, ..., \mathbf{h}_{256}$). The encoder ($\mathbf{E}(\cdot)$) is formed of *bottleneck blocks* borrowed from ResNet. The encoder structure is shown in Fig. 3.

Then, we employ a quantiser ($\mathbf{Q_e}(\cdot)$) from the VQ-VAE, which maintains a group of *code vectors* ($\mathbf{e}_1, ..., \mathbf{e}_k$), and always replaces latent vectors with the nearest code vectors. Given a latent vector $\mathbf{h}$, the quantiser is defined by Eq. 1, and, each replaced latent vector ($\mathbf{h}'$) can be represented by the index ($id(\cdot)$) of the code vector q (Eq. 2).

$$\mathbf{h}' = \mathbf{Q_e}(\mathbf{h}) = arg\min_{\mathbf{e}_q} ||\mathbf{e}_q - \mathbf{h}||_2^2,\ \mathbf{e}_q \in \{\mathbf{e}_0, ..., \mathbf{e}_k\} \quad (1)$$

$$q = id(\mathbf{e}_q),\ q \in [0, k] \quad (2)$$

We use a decoder ($\mathbf{D}(\cdot)$) to reconstruct the images based on the replaced latent vectors. Like the encoder, the decoder is formed of *bottleneck blocks*, with their *convolutional layers* replaced with the *transposed convolutional layers* for the upsampling propose. As soon as the decoder is able to reconstruct images (the reconstructed images $\mathbf{x}'$ is close to $\mathbf{x}$), all the necessary information (in $\mathbf{x}$) is preserved in the replaced latent vectors as well as the index grids. The discrete-encoder is defined by Eq. 3.

$$\begin{aligned} \mathbf{h}_1, ..., \mathbf{h}_{256} &= \mathbf{E}(\mathbf{x}) \\ \mathbf{h}'_1, ..., \mathbf{h}'_{256} &= \mathbf{Q_e}(\mathbf{h}_1), ..., \mathbf{Q_e}(\mathbf{h}_{256}) \\ \mathbf{x}' &= \mathbf{D}(\mathbf{h}'_1, ..., \mathbf{h}'_{256}) \end{aligned} \quad (3)$$

The index grid denoted as $\mathbf{z}$ is defined by Eq. 4

$$\mathbf{z} = id(\mathbf{h}'_1), ..., id(\mathbf{h}'_{256}) \quad (4)$$

Comparing to the VQ-VAE (2nd version [10]), we removed the skip connections to force the decoder to reconstruct the images only according to the discrete encoding. Comparing to VQ-VAE (1st version [9]) we improve the reconstruction quality by using the ResNet-based autoencoder

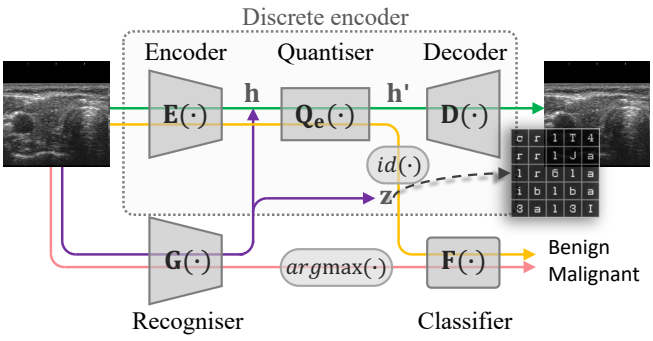

Fig. 2. The components and training process of the proposed model, where '⟶' denotes the discrete encoder training, '⟶' denoting the classifier training, '⟶' denoting the recogniser training, and '⟶' denoting the testing process.

and large-sized latent vectors ($\mathbf{h}_i \in \mathbb{R}^{2048}, i \in [1, 256]$). This strategy successfully generates high quality 256x256 images in our experiments, which will be further discussed in §4 (see Fig. 4).

The input of the classifier ($\mathbf{F}(\cdot)$) is the index grids ($\mathbf{z}$) of the images. Since the indices are discrete symbols, we treat them as texts and employ a method of text modelling to learn the embedding of the indices. Here, we uses word2vec [12] to initialise the embedding vectors ($\mathbf{m}_0, ..., \mathbf{m}_k$) for the indices ($0, ..., k$). We use a relatively small embedding vector size ($\mathbf{m}_0, ..., \mathbf{m}_k \in \mathbb{R}^{16}$) compared to the latent vectors (i.e. $\mathbb{R}^{2048}$). Specifically, the word2vec model is trained under CBOW architecture, regarding the code vector indices as words and every 5x5 area in the index grids as the word-bags (i.e. the contexts).

The classifier $\mathbf{F}(\cdot)$ is a lite CNN-based model (see Fig. 3). It contains original convolutional layers, instance norm layer, and ReLu layers. Instance norm and ReLu layers are used in between every two convolutional layers. The classifier $\mathbf{F}(\cdot)$ predicts the image labels in one-hot formed vector ($\mathbf{y}'$).

$$\mathbf{y}' = \mathbf{F}(\mathbf{z}) \quad (5)$$

Like usual autoencoders [13, 14, 15], the discrete encoder uses a large number of parameters for reconstruction. Since our final target is classification, we distil the knowledge from the discrete encoder into the recogniser to reduce the model size, that is, the discrete encoder is not included in the final model.

The recogniser $\mathbf{G}(\cdot)$ is also a lite CNN-based model. It directly predicts the index grids. $\mathbf{G}(\cdot)$ estimates the index distributions for each position (denoted by $\mathbf{t}_1, ..., \mathbf{t}_{256} \in \mathbb{R}^k$) in an index grid. So we use argmax function to find the most likely index grids (i.e. $\mathbf{z}$ in Eq. 4 and Eq. 6).

$$\begin{aligned} \mathbf{t}_1, ..., \mathbf{t}_{256} &= \mathbf{G}(\mathbf{x}) \\ \mathbf{z} &\approx arg\max_i \mathbf{t}_1^{(i)}, ..., arg\max_i \mathbf{t}_{256}^{(i)} \end{aligned} \quad (6)$$

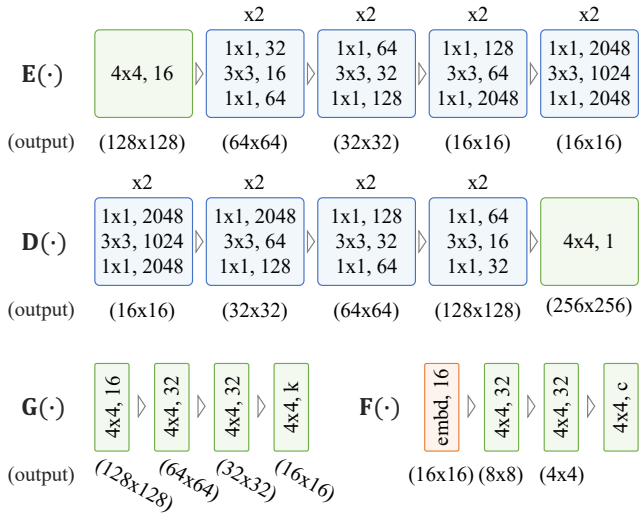

Fig. 3. The frameworks of our model components. The illustration follows the ResNet [11] – '1x1', '3x3' and '4x4' are the convolution kernel size followed by the output channel dimension. The blue blocks denotes the layers with residual connections (green and red blocks denote the normal layers). k is the number of indices, and c is the number of label classes. NB: $\mathbf{Q_e}(\cdot)$ has no layer so it is not listed in the figure.

3. TRAINING

The discrete-encoder, classifier, and recogniser are trained separately (Fig. 2). First, we train the discrete-encoder (i.e. $\mathbf{E}(\cdot)$, $\mathbf{Q}(\cdot)$, and $\mathbf{D}(\cdot)$), and since it does not require labels, it can be trained on both the labelled and non-labelled images. We use L_2 loss for the image reconstruction; its code vector size (i.e. k) is set to 64.

$$\mathcal{L}_{rec} = ||\mathbf{x} - \mathbf{x}'||_2^2 \tag{7}$$

Because there is no gradient can pass the $argmin(\cdot)$ in Eq 1, we use Straight-Through Estimator [16] to estimate the gradient of $\mathbf{h}_1, ..., \mathbf{h}_{256}$, which to directly defines the gradient of $\mathbf{h}$ by the gradient of the corresponding $\mathbf{e}_q$ (i.e. $\bigtriangledown_{\mathbf{h}} \stackrel{def}{=} \bigtriangledown_{\mathbf{e}_q}$). Meanwhile, the code vectors in our quantiser also need to be trained. We also use the L_2 loss to let the code vectors be close to the latent vectors:

$$\mathcal{L}_{ker} = \frac{1}{n}\sum_{i=0}^{n} ||\mathbf{h}_i - \mathbf{e}_{q_i}||_2^2 \tag{8}$$

The entire loss for the discrete-encoder is the combination of the two losses, which is $\mathcal{L}_{dis} = \mathcal{L}_{rec} + \mathcal{L}_{ker}$.

After the training of discrete-encoder, we use the discrete-encoder to transform all the images into index grids. Then, we collect the 5x5 areas of the index grids for the word2vec training, and obtain the embedding vector for the code vector indices.

Second, we train the classifier $\mathbf{F}(\cdot)$ with the image labels ($\mathbf{y}$) and the corresponding index grids ($\mathbf{z}$). The index grids of unlabelled images are not used in this process. In the training process, the embedding vectors ($\mathbf{m}_0, ..., \mathbf{m}_k$) are allowed to be updated, and the gradients for each embedding vector are scaled by the inverse frequency of the symbols in the mini-batch. The training process uses cross entropy loss.

$$\mathcal{L}_{cls} = CrossEntropy(\mathbf{y}', \mathbf{y}) \tag{9}$$

Finally, we distil the knowledge from the discrete-encoder into the recogniser. We use $\mathbf{E}(\cdot)$ and $\mathbf{Q_e}(\cdot)$ as the teacher model. The student model (i.e. $\mathbf{G}(\cdot)$) is trained to predict the index grids ($\mathbf{z}$). Unlike normal distil processes (e.g. [17]), we create the soft targets by considering both the outcomes of $\mathbf{E}(\cdot)$ and $\mathbf{Q_e}(\cdot)$ i.e. the latent vectors and the target indices (q). Given a latent vector $\mathbf{h}$, the soft target ($t(\mathbf{h})$) is created by Eq. 10.

$$t(\mathbf{h}) = \delta\left(id(\mathbf{Q_e}(\mathbf{h}))\right) - \left[||\mathbf{h} - \mathbf{e}_1||_2^2, ..., ||\mathbf{h} - \mathbf{e}_k||_2^2\right] \tag{10}$$

The first term (i.e. $\delta\left(id(\mathbf{Q_e}(\mathbf{h}))\right)$) denotes one-hot formed indices, the hard target, where $\delta(\cdot)$ returns one-hot vector with its argument referring to the non-zero item. The remaining term provides extra weights to those indices whose code vectors are close to the latent vector. This aims to encourage the recogniser to learn the generalisation ability of the discrete-encoder [17]. The distil training also adopts the L_2 loss:

$$\mathcal{L}_{recog} = ||\mathbf{G}(\mathbf{x}) - t(\mathbf{E}(\mathbf{x}))||_2^2 \tag{11}$$

4. EXPERIMENTS

We test our model on two open ultrasound image datasets. DDIT [18] is *thyroid nodules* classification dataset, containing 602 ultrasound images – 102 benign, 368 malignant and others are unlabelled. BUSI [19] is for *breast nodules* classification. It contains 780 ultrasound images with 437 benign and 210 malignant images. Our models do not use the nodule boundary information in the experiments. We compared our model with the existing models which are trained and tested on the open datasets, and all the selected models are recently published,including the model based on SVM[20], FFT/CNN[6], Res-Gan[21], DRS/ResNet[22], and DenseNet[23]. Because both the two datasets do not pre-split the images into training and testing sets, like these existing works, we randomly split the datasets with 80% images for training and 20% for testing. 5-fold cross-validations were used in our experiments.

We subtract the margins of the images and resized them to 256x256 resolution (like in Fig. 1). We uses small batch (size = 16), and AdamW optimiser [24] to train the models, with learning rate 3e-4, weight decay 1e-2. We only use the random horizontal flip to extend the datasets. The training processes are stopped when the losses stop decreasing in the next

Model	P	Sp	R/Se	F1	Acc	Φ
DDTI dataset						
FFT/CNN [6]	-	65.7	96.1	-	92.1	22M
Res-GAN [21]	-	86.5	95.0	-	92.2	10M
SVM [20]	-	94.6	88.8	-	93.8	n/a
LTQ (ours)	99.8	**99.5**	94.5	97.1	95.5	0.1M
LTQ+ (ours)	99.8	**99.5**	**95.7**	97.7	**96.5**	0.1M
BUSI dataset						
DRS/ResNet [22][b]	-	89.7	97.6	-	92.3	47M
DenseNet [23][b]	**90.0**	**95.6**	92.3	91.4	**94.6**	29M
LTQ (ours)	85.8	92.4	98.6	91.7	94.4	0.1M
LTQ+ (ours)	86.2	92.7	**98.7**	**92.0**	**94.6**	0.1M

Table 1. Experiment results on DDTI and BUSI datasets. [b] denotes the model requires the nodule boundary information. '-' means the papers do not report the corresponding results.

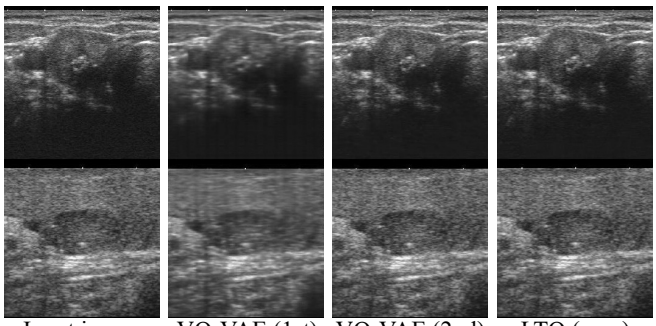

Fig. 4. The reconstruction comparison. The images are generated by our discrete-encoder (LTQ), VQ-VAE (1st [9]), and VQ-VAE (2nd [10]). Please zoom in to see the difference.

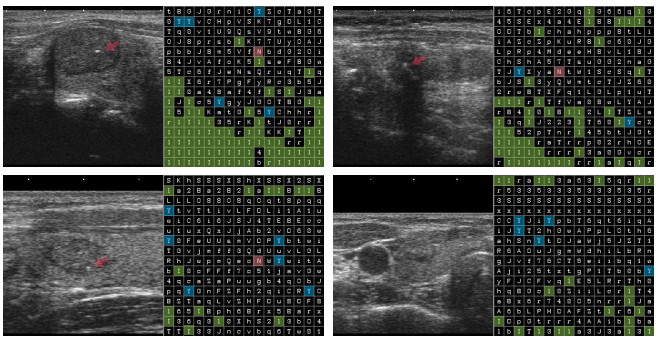

Fig. 5. Images with their index grids, highlighting the possible mappings between indices and texture patterns. We can see 'N' refers to the nodule calcification, 'I' and 'l' for shadow areas, and 'Y' for strong reflection areas.

20 epochs. The experimental results (in Table 1) are comparing to the existing approaches. In the tables, P, R, F1, Acc denote the Precision, Recall, F1-score, and Accuracy respectively, while, Se and Sp are Sensitivity and Specificity (NB: Recall equals Sensitivity), and Φ denote the number of model parameters. LTQ denotes our model that is trained on labelled training data, and LTQ+ denotes the version of our model with a discrete-encoder that is trained with both labelled training data and unlabelled images.

We can see that overall our model performed better than the compared baseline models on both DDTI and BUSI datasets. The proposed deep learning based models are effective on small datasets. Meanwhile, our model is a lite model. It uses 0.1 million trainable parameters, which is significantly less than other deep learning models. Our model is easy to be adopted on those medical devices or other IoT devices that have limited computational capacity.

We also subjectively evaluate the reconstruction quality of the proposed model. We can see that the reconstructed image of our discrete-encoder is significantly clear than the first version of VQ-VAE [9], and similar to the second version of VQ-VAE [10]. This shows that all the necessary information of the original input images have been preserved in the index grids. Although the second version of VQ-VAE can reconstruct high-quality images, it cannot ensure all the necessary information is preserved in the latent vectors due to using skip connections.

Finally, we discuss the model explainability via analysing what indices refer to what texture patterns. Although how indices map to the pattern is arbitrary, and multiple indices may refer to the same patterns, the indices should refer to the patterns near to them, because of the residual connections. Fig. 5 shows the visualisation of the index grids. There are 64 code vectors ($k = 64$) for index grids. Numbers ('0'-'9'), lower case and upper case letters ('a'-'z', 'A'-'Z'), and character '+' and '-' are used to name the code vectors. We manually coloured the cells to show some obvious mappings. We can see that the index 'N' appears in the areas that the nodule has the calcification, 'I' and 'l' are related to shadow areas, and 'Y' may be related to the strong reflection areas. This shows that the proposed model has certain degrees of explainability.

5. CONCLUSION

This paper proposed a novel ultrasound image classification framework through image local texture quantisation. We used the discrete encoding to separate the classification task into two steps: first transforming images into index grids, then conducting the classification based on the index grids. The framework allows the models to be effectively trained on small datasets and reduce the impact of noise. We also showed the proposed discrete-encoder performs better than VQ-VAE in terms of extracting indices, and discussed the model explainability through index grids analysis. The experiments conducted on two open ultrasound image datasets demonstrate the effectiveness of the proposed framework.

6. REFERENCES

[1] Juan L Mateo and Antonio Fernández-Caballero, "Finding out general tendencies in speckle noise reduction in ultrasound images," *Expert systems with applications*, vol. 36, no. 4, pp. 7786–7797, 2009.

[2] Jinlian Ma, Fa Wu, Tian'an Jiang, Qiyu Zhao, and Dexing Kong, "Ultrasound image-based thyroid nodule automatic segmentation using convolutional neural networks," *International journal of computer assisted radiology and surgery*, vol. 12, no. 11, 2017.

[3] Xiangchun Li, Sheng Zhang, Qiang Zhang, Xi Wei, et al., "Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study," *The Lancet Oncology*, vol. 20, no. 2, pp. 193–201, 2019.

[4] Gesheng Song, Fuzhong Xue, and Chengqi Zhang, "A model using texture features to differentiate the nature of thyroid nodules on sonography," *Journal of Ultrasound in Medicine*, vol. 34, no. 10, 2015.

[5] Noura Aboudi, Ramzi Guetari, and Nawres Khlifa, "Multi-objectives optimisation of features selection for the classification of thyroid nodules in ultrasound images," *IET Image Processing*, vol. 14, no. 9, 2020.

[6] Dat Tien Nguyen, Jin Kyu Kang, Tuyen Danh Pham, Ganbayar Batchuluun, and Kang Ryoung Park, "Ultrasound image-based diagnosis of malignant thyroid nodule using artificial intelligence," *Sensors*, vol. 20, 2020.

[7] Rui Mao, Chenghua Lin, and Frank Guerin, "End-to-end sequential metaphor identification inspired by linguistic theories," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 3888–3898.

[8] Rui Mao and Xiao Li, "Bridging towers of multi-task learning with a gating mechanism for aspect-based sentiment analysis and sequential metaphor identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, pp. 13534–13542.

[9] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu, "Neural discrete representation learning," *arXiv preprint arXiv:1711.00937*, 2017.

[10] Ali Razavi, Aaron van den Oord, and Oriol Vinyals, "Generating diverse high-fidelity images with vq-vae-2," in *Advances in neural information processing systems*, 2019, pp. 14866–14876.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[12] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[13] Xiao Li, Chenghua Lin, Ruizhe Li, Chaozheng Wang, and Frank Guerin, "Latent space factorisation and manipulation via matrix subspace projection," in *International Conference on Machine Learning*. PMLR, 2020.

[14] Xiao Li, Guanyi Chen, Chenghua Lin, and Ruizhe Li, "Dgst: a dual-generator network for text style transfer," *arXiv preprint arXiv:2010.14557*, 2020.

[15] Ruizhe Li, Xiao Li, Chenghua Lin, Matthew Collinson, and Rui Mao, "A stable variational autoencoder for text modelling," *arXiv preprint arXiv:1911.05343*, 2019.

[16] Yoshua Bengio, Nicholas Léonard, and Aaron Courville, "Estimating or propagating gradients through stochastic neurons for conditional computation," *arXiv preprint arXiv:1308.3432*, 2013.

[17] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[18] Lina Pedraza, Carlos Vargas, Fabián Narváez, Oscar Durán, et al., "An open access thyroid ultrasound image database," in *10th International Symposium on Medical Information Processing and Analysis*, 2015, vol. 9287.

[19] Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy, "Dataset of breast ultrasound images," *Data in brief*, vol. 28, pp. 104863, 2020.

[20] U Raghavendra, Anjan Gudigar, Mangalore Maithri, Arkadiusz Gertych, et al., "Optimized multi-level elongated quinary patterns for the assessment of thyroid nodules in ultrasound images," *Computers in biology and medicine*, vol. 95, pp. 55–62, 2018.

[21] Yuan Hang, "Thyroid nodule classification in ultrasound images by fusion of conventional features and resgan deep features," *Journal of Healthcare Engineering*, 2021.

[22] Michal Byra, "Breast mass classification with transfer learning based on scaling of deep representations," *Biomedical Signal Processing and Control*, 2021.

[23] Woo Kyung Moon, Yan-Wei Lee, Hao-Hsiang Ke, Su Hyun Lee, et al., "Computer-aided diagnosis of breast ultrasound images using ensemble learning from convolutional neural networks," *Computer methods and programs in biomedicine*, vol. 190, 2020.

[24] Ilya Loshchilov and Frank Hutter, "Fixing weight decay regularization in adam," 2018.