Accurate BAPL Score Classification of Brain PET Images Based on Convolutional Neural Networks with a Joint Discriminative Loss Function †

Ryosuke Sato 1,‡, Yutaro Iwamoto 1,‡, Kook Cho 2,‡, Do-Young Kang 3,4,‡ and Yen-Wei Chen 1,*

1 College of Information Science and Engineering, Ritsumeikan University, Shiga 603-8577, Japan; is0259vf@ed.ritsumei.ac.jp (R.S.); yiwamoto@fc.ritsumei.ac.jp (Y.I.)
2 College of General Education, Dong-A University, Busan 49315, Korea; kcho@dau.ac.kr
3 Institute of Convergence Bio-Health, Dong-A University College of Medicine, Busan 49201, Korea; dykang@dau.ac.kr
4 Department of Nuclear Medicine, Dong-A University College of Medicine, Busan 49201, Korea
* Correspondence: chen@is.ritsumei.ac.jp
† This paper is an extended version of paper published in the 7th International Conference on Innovation in Medicine and Healthcare held in St. Julians, Malta, 17–19 June 2019.
‡ These authors contribute equally to this work.

Received: 17 October 2019; Accepted: 13 December 2019; Published: 2 February 2020

Abstract: Alzheimer’s disease (AD) is an irreversible progressive cerebral disease with most of its symptoms appearing after 60 years of age. Alzheimer’s disease has been largely attributed to accumulation of amyloid beta (Aβ), but a complete cure has remained elusive. 18F-Florbetaben amyloid positron emission tomography (PET) has been shown as a more powerful tool for understanding AD-related brain changes than magnetic resonance imaging and computed tomography. In this paper, we propose an accurate classification method for scoring brain amyloid plaque load (BAPL) based on deep convolutional neural networks. A joint discriminative loss function was formulated by adding a discriminative intra-loss function to the conventional (cross-entropy) loss function. The performance of the proposed joint loss function was compared with that of the conventional loss function in three state-of-the-art deep neural network architectures. The intra-loss function significantly improved the BAPL classification performance. In addition, we showed that the mix-up data augmentation method, originally proposed for natural image classification, was also useful for medical image classification.

Keywords: Alzheimer’s disease; deep learning; convolutional neural network; PET image; brain amyloid plaque load (BAPL) score; coronal plane; intra-loss function; joint loss function; mix-up; data augmentation

1. Introduction

Alzheimer’s disease (AD) is an irreversible progressive cerebral disease with most of its symptoms appearing after 60 years of age. Alzheimer’s disease currently affects approximately 46.8 million people worldwide and may reach 132 million people in 2050 [1]. To date, a complete cure for AD has remained elusive. Accurate and early diagnosis of AD is essential for treating patients and for developing future treatments. Accumulation of amyloid beta (Aβ) is considered as a diagnostic feature of Alzheimer’s disease. Accordingly, 18F-florbetaben (FBB) amyloid positron emission tomography (PET) is a more powerful tool for understanding AD-related brain changes than magnetic resonance imaging (MRI) and computed tomography (CT) [2,3].

According to the International Working Group, both pathophysiological markers and clinical phenotypes are required for an AD diagnosis. Pathophysiological markers include CSF (cerebrospinal fluid) analysis for amyloid beta and phosphorylated tau proteins, and magnetic resonance imaging (MRI) for atrophy. Clinical phenotypes include cognitive decline, functional impairment, and behavioral changes. Early diagnosis is crucial for providing appropriate care and intervention, as well as for enabling timely enrollment in clinical trials for new treatments.
fluid) Aβ/tau or amyloid PET [4], but clinical trials of Aβ-targeting drugs have been unsuccessful in clinical trials. This failure might be attributable to the late application of the treatment, highlighting the need for early treatment after early diagnosis [5]. Amyloid markers are the fastest appearing biomarkers early in the disease [6–8].

In recent years, AD in MRI or PET brain images has been identified by various machine learning methods including deep learning [3,9–13]. Zhang [9] classified AD and normal control images by a combined kernel technique with a support vector machine. Sarraf [10] developed the program DeepAD for AD diagnosis which analyzes sMRI and fMRI brain scans separately on the slice and subject levels by two convolutional neural networks (CNNs) (LeNet and GoogleNet). Farooq [11] classified AD in MRI scans by a deep CNN-based multi-class classification algorithm based on GoogleNet and ResNet. In a related study of PET images, Kang [12] proposed a classification method for scoring brain amyloid plaque load (BAPL) in FBB PET images which is based on a deep CNN developed by the Visual Geometry Group. Liu [13] combined a CNN with recurrent neural networks (RNN) for classifying amyloids in fluorodeoxyglucose (FDG) PET images. Although Kang’s method satisfactorily classified images as positive or negative for amyloids, it could not accurately identify BAPL2 in a ternary classification, because BAPL2 is a weaker amyloid load than BAPL1 [12]. In some cases, the interpreter cannot easily distinguish between BAPL1 and BAPL2. Liu [13] studied FDG PET images, which are more appropriate for identifying progression markers than diagnostic markers, and clinically classified them as normal, mild cognitive impairment (MCI), or AD. Choi [7] combined florbetapir (not FBB) amyloid PET and FDG PET images for clinical classification of normal, MCI, and AD diagnoses.

In this paper, we added a discriminative intra-loss function to the conventional loss function (i.e., the cross-entropy loss function). The resulting joint discriminative loss function was intended to improve the accuracy of deep-learning-based AD classification. The performances of the proposed joint loss function and the conventional loss function were compared in three state-of-the-art deep neural network architectures. The intra-loss function significantly improved the classification accuracy, especially that of BAPL2 (a marker of MCI). In addition, we show that the mix-up data augmentation method, originally proposed for natural image classification [14], is also useful in medical image classification.

The remainder of this paper is organized as follows. Sections 2 and 3 describe the dataset and methods of the study, respectively. The experimental results are presented and discussed in Section 4. Conclusions are given in Section 5.

2. Materials

The analyzed data were PET images provided by Dong-A University of Korea [12]. All participants underwent 18F FBB PET/CT. All PET examinations were performed using a Biograph 40 mCT flow PET/CT scanner (Siemens Healthcare, Knoxville, TN, USA). The PET images were reconstructed by UltraHD-PET (TrueX-TOF). Participants were intravenously injected with a 300 MBq dose of 18F FBB (NeuraCeq, Piramal Pharma, Mumbai, India) followed by a saline flush. Thereafter, the participants rested in a bed within a quiet room with dim lighting. The PET/CT scans were acquired 90 min after the radiotracer injection. The participants’ heads were secured with Velcro straps to minimize motion artifacts. Helical CT was performed with a rotation time of 0.5 s at 100 kVp and 228 mAs with no intravenous contrast agent. Immediately afterwards, the PET was acquired for 20 min in three-dimensional mode. All images were acquired from the skull vertex to the skull base.

The PET images were categorized into three classes based on their BAPL scores: 1 (BAPL1), 2 (BAPL2) and 3 (BAPL3) [8]. The database comprised 380 PET images: 188 BAPL1 subjects, 48 BAPL2 subjects, and 144 BAPL3 subjects. Table 1 shows the demographic distribution of the dataset. Each PET image was sized (400 × 400) pixels$^2$ with a pixel size of (1.02 × 1.02) mm$^2$ and a slice thickness of 1.5 mm. We normalized each 3D PET image to a standard brain image by statistical parametric mapping [15]. Typical central normalized slice images of BAPL1, BAPL2, and BAPL3 are shown in panels (a), (b), and (c) of Figure 1, respectively. The left, center, and right images in each row were acquired in the axial
regional cortical tracer binding (RCTB) was then resized to 224 × 224 pixels as the input image for classification in this study. Comparison of the axial plane and the coronal plane with the proposed method is shown in Section 4.4. To obtain a sufficiently large dataset, we extracted 11 slices at the center of the normalized hippocampus area (central slice ± 5 slices corresponding to a total thickness of 16.5 mm) from the coronal images as our dataset. Each slice image was then resized to 224 × 224 pixels.

Table 1. Demographic distribution of our dataset.

| Diagnosis | Number | Age      | Gender (F/M) |
|-----------|--------|----------|--------------|
| BAPL1     | 188    | 67.1 ± 8.7 | 71/117       |
| BAPL2     | 48     | 72.8 ± 6.1 | 21/27        |
| BAPL3     | 144    | 69.3 ± 7.9 | 62/82        |

BAPL: brain amyloid plaque load score.

Although the entire brain image (3D image) provides a more comprehensive structural knowledge than plane images, the 3D implementation requires the estimation of many network parameters which increases the size of the training dataset. Thus, we classified the BAPL score by a 2D deep learning algorithm applied to 2D slices (central hippocampus slices) of PET images as done in most deep learning-based methods. In our previous study [16], we showed that inputting coronal plane images achieved better results than inputting the axial and sagittal planes. Therefore, we used the coronal plane image as the input image for classification in this study. Comparison of the axial plane and the coronal plane with the proposed method is shown in Section 4.4. To obtain a sufficiently large dataset, we extracted 11 slices at the center of the normalized hippocampus area (central slice ± 5 slices corresponding to a total thickness of 16.5 mm) from the coronal images as our dataset. Each slice image was then resized to 224 × 224 pixels.

3. Methods

3.1. FBB PET Interpretations

The FBB PET images were visually assessed by two nuclear medicine physicians. The subjects’ amyloid depositions were rated in four brain regions (frontal cortex, posterior cingulate, lateral temporal cortex, and parietal cortex) using a predefined regional cortical tracer binding (RCTB) scoring system (1 = no binding, 2 = minor binding, 3 = pronounced binding). The RCTB scores of each PET scan

Figure 1. Typical central slice images of (a) BAPL1; (b) BAPL2; and (c) BAPL3.
were condensed into a single score on a predefined three-grade scoring system. The final score, called the brain amyloid-β plaque load (BAPL) score, was divided into three grades: 1 (BAPL1), 2 (BAPL2), and 3 (BAPL3), indicating no Aβ load, minor Aβ load, and significant Aβ load, respectively [8].

3.2. CNN Models

As mentioned in Section 2, we classified the BAPL score of PET images by a deep-learning-based 2D implementation in standard deep-learning-based methods. The useful high-level features were extracted from each coronal slice image by 2D CNNs [16]. The image was then classified as BAPL1, BAPL2, or BAPL3. The network structure is shown in Figure 2. As the backbone networks for feature extraction, we employed three state-of-the-art CNNs: VGG 19 [17], ResNet 50 [18], and DenseNet 121 [19]. The performances of the three networks is discussed in Section 4.

![Figure 2. Convolutional neural network (CNN) architecture.](image)

### 3.2.1. VGG19

The Visual Geometry Group network (VGGNet), proposed in 2014 [17], combines multiple $3 \times 3$ convolutional filters into a large receptive field (the output size depends on the effective area of the input image). Multiple stacked smaller-sized filters achieve a deeper network than a single large filter, so more complex features can be learned at lower cost. The VGG19 is a VGGNet with 19 layers (16 convolutional layers follow three fully connected layers) as shown in Figure 3.

![Figure 3. VGG19 architecture.](image)

### 3.2.2. ResNet50

In the traditional CNN, the output of the upper convolutional layer is input to the next convolutional layer. In 2016, He et al. [18] proposed a deep residual network (ResNet) which learns the residual knowledge ($F(x) = H(x) - x$, where $x$ is the input feature and $H(x)$ is the output) through a shortcut connection (see Figure 4a). The shortcut effectively prevents the network degradation and improves the network performance. ResNet won the ILSVRC-2015 challenge and is widely used in medical image classification [11,20]. ResNet50 is a 50 layer ResNet constructed as shown in Figure 4b.
The densely connected network (DenseNet) architecture is an extension of ResNet which groups the convolution layers into a block of layers called a dense block [19]. DenseNet differs from other networks in its connectivity pattern. DenseNet is a stack of dense blocks, each consisting of multiple convolutional layers. It requires fewer parameters than ResNet. It also solves the information (and gradient) flow problem, because each layer can directly access the gradients from the loss function and the original input image [19]. Figure 5a shows a representative DenseNet with two dense blocks. The detailed architecture of DenseNet121 is shown in Figure 5b.

The cross-entropy is a common loss function in image classification tasks. In this paper, we first defined the cross-entropy as the following inter-loss function:

$$\mathcal{L}_{\text{inter}} = - \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{3} y_i \ln p_{i,k}$$  \hspace{1cm} (1)
where \( N \) is the number of samples. \( p_{i,k} \) is the softmax output, representing the probability that the \( i \)-th sample belongs to the \( k \)-th class, and \( y_i \) represents the label. Here, \( y = 1, 2, \) and \( 3 \) correspond to BAPL1, BAPL2, and BAPL3, respectively.

We then defined an intra-loss (i.e., center loss [20, 21]) function that handles the intra-variation problem:

\[
L_{\text{intra}} = \frac{1}{2} \sum_{i=1}^{N} ||f_i - c_{y_i}||^2
\]  

(2)

where \( f_i \) is the feature vector of the \( i \)-th sample and \( c_{y_i} \) denotes the center of the feature class to which \( y_i \) belongs which is updated during the training by Equation (3). The update is operated on each batch (not on the entire training set).

\[
c_j = c_j - \alpha \cdot \Delta c_j = c_j - \alpha \cdot \frac{\partial L_{\text{intra}}}{\partial c_j}
\]

\[
= c_j - \alpha \cdot \frac{\sum_{i=1}^{N} \delta(y_i = j)(c_j - f_i)}{1 + \sum_{i=1}^{N} \delta(y_i = j)}
\]  

(3)

where \( \delta \) is the delta function. In this paper, the learning rate \( \alpha \) was set to 0.5.

Finally, the inter-loss (Equation (1)) and intra-loss (Equation (2)) are combined into a joint loss for training the network:

\[
\mathcal{L} = \mathcal{L}_{\text{inter}} + \lambda \mathcal{L}_{\text{intra}}
\]  

(4)

where the parameter \( \lambda \) adjusts the balance between the inter-class and intra-class losses. Note that when \( \lambda = 0 \), the joint loss reduces to the conventional cross-entropy loss function used in the existing methods [3, 9–13].

3.4. Data Augmentation with Mix-Up

The dataset was augmented with mix-up [14], an augmentation method that mixes two training samples to create a new sample pair composed of an input image (\( X \)) and its class label (\( y \)). Suppose that \( (X_1, y_1) \) and \( (X_2, y_2) \) are two samples in the existing dataset. The augmented sample \( (X, y) \) is then generated as follows:

\[
X = bX_1 + (1 - b)X_2
\]  

(5)

\[
y = by_1 + (1 - b)y_2
\]  

(6)

where \( b \in [0,1] \) is sampled from a beta distribution [14]. Table 2 shows the distribution of the augmentation dataset for the eight-fold cross-validation.

4. Results and Discussion

4.1. Experimental Setup

The networks (VGG19, ResNet50, and DenseNet121) are based on an implementation using the Keras [22] and TensorFlow [23] libraries. All computations were performed on a NVIDIA GeForce GTX TITAN X GPU with 12 GB of memory. We first loaded the weights of the pre-trained CNNs from ImageNet [24] provided by Keras. All layers of the networks were then fine-tuned on our training dataset (PET images).

The dataset was divided into eight groups. Each group contained 23 subjects in BAPL1, 6 subjects in BAPL2, and 18 subjects in BAPL3. The classification accuracy, defined by Equation (7), was obtained by eight-fold cross-validation. The dataset was augmented with mix-up [14] as described in Section 3.4. The sizes of the training and test datasets after the data augmentation are listed in Table 2.

\[
\text{Accuracy} = \frac{TP(\text{BAPL1}) + TP(\text{BAPL2}) + TP(\text{BAPL3})}{TP(\text{BAPL1}) + TP(\text{BAPL2}) + TP(\text{BAPL3}) + FP(\text{BAPL1}) + FP(\text{BAPL2}) + FP(\text{BAPL3})}
\]  

(7)
where TP and FP denote true positive and false positive, respectively. Note that the accuracy shown in following subsections was based on classification of the slice images. We also evaluated the accuracy per case (per patient) in Section 4.4. The case label (patient label) was based on the major vote of its 11 slice labels.

Table 2. Datasets after data augmentation for eight-fold cross-validation.

|       | Train                      | Test          |
|-------|----------------------------|---------------|
| BAPL1 | 165 subject (1,815→5,445 slices) | 23 subjects (253 slices) |
| BAPL2 | 42 subjects (462→4620 slices)   | 6 subjects (66 slices)  |
| BAPL3 | 126 subjects (1386→4,620 slices) | 18 subjects (198 slices) |
| Total | 15,609 slices               | 517 slices     |

4.2. Effectiveness of Data Augmentation Based on Mix-up

We first confirmed the effectiveness of the data augmentation based on mix-up. The validation was performed on ResNet50 with the conventional loss function (λ = 0.00). We performed three experiments: without data augmentation, with data augmentation based on conventional translation and inversion, and with data augmentation based on mix-up. The comparison results are shown in Table 3. The effectiveness of the data augmentation was statistically tested in a t-test. Both methods with data augmentation outperformed the method without data augmentation (translation and inversion, p = 0.003; mix-up, p = 0.0001), and the results were statistically significant (p < 0.05). Moreover, the data augmentation based on mix-up outperformed the conventional augmentation based on translation and inversion (p = 0.0052 < 0.05). To our knowledge, we present the first evidence that mix-up data augmentation improves medical image classification. Mix-up-based data augmentation was therefore applied in our further experiments.

Table 3. Effectiveness of mix-up-based data augmentation (accuracy of ResNet50 with λ = 0.00).

| Without Data Augmentation | With Data Augmentation |
|---------------------------|------------------------|
|                           | Translation and Inversion | Mixup       |
| 83.87%                    | 86.13%                  | 88.90%      |

4.3. Impact of λ

This subsection checks the effectiveness of the proposed joint loss for different values of the hyperparameter λ which weights the combined conventional cross-entropy loss (Equation (1)) and the intra-loss (Equation (2)). These experiments were performed on ResNet50 with mix-up-based data augmentation. In the proposed method, λ is always greater than 0.00. Recall that when λ = 0.00, the method reduces to the existing method with conventional cross-entropy loss. The classification accuracies, defined as the ratios of the correctly classified samples over all classified samples, are presented for different λ in Table 4. The proposed joint loss (λ > 0.00) consistently outperformed the conventional loss (λ = 0.00) and was optimized at λ = 0.0025. Therefore, we set λ = 0.0025 in further detailed comparisons.

Table 4. Classification accuracies of the proposed joint loss with different λ (CNN: ResNet50) (λ = 0.00 corresponds to the conventional cross-entropy loss).

| λ     | Accurary         |
|-------|-----------------|
| 0     | 88.90%          |
| 0.001 | 89.53%          |
| 0.0025| 91.03%          |
| 0.005 | 90.12%          |
| 0.01  | 89.32%          |
| 0.1   | 89.53%          |
4.4. Coronal Plane versus Axial Plane

We compared the classification accuracy with the axial plane and the coronal plane. The results are shown in Table 5. We also showed the accuracy per case (per patient) in Table 5 which is defined in Section 4.1. As shown in Table 5, both the classification accuracies per slice and per case with the coronal plane are higher than those with the axial plane as well as our previous results using the conventional loss function [16]. The results suggest that the coronal plane contains more useful information for the classification of BAPL scores which are extracted by deep learning. In the future, we will develop a method to visualize the useful features of the trained deep convolutional neural network for BAPL score classification.

Table 5. Comparison of classification accuracy with the axial plane and the coronal plane (CNN: ResNet50).

| Plane for Analysis | Accuracy | Accuracy per Case |
|--------------------|----------|-------------------|
| Axial Plane        | 85.18%   | 87.50%            |
| Coronal Plane      | 91.03%   | 93.75%            |

4.5. Detailed Comparison between the Proposed and Conventional Loss Functions on Three Network Architectures

For a detailed performance comparison, we compared the precisions, recalls, F1 scores, and accuracies of three state-of-the-art networks (i.e., VGG19, ResNet50, and DenseNet121) with the proposed and conventional loss functions. The precision and recall define the ratios of correctly classified positive samples to all classified positive samples and all true positive samples, respectively. The F1 score is the harmonic mean (average) of the precision and recall. The results are summarized in Table 6. As shown in Table 6, the proposed joint loss function improved the classification performance of all three networks, confirming its usefulness in AD classification by any state-of-the-art network. Also, notably, the joint loss function greatly improved the classification performance of BAPL2 (MCI). In the networks with the conventional loss function ($\lambda = 0.00$), the classification performances were high for BAPL1 and BAPL3 but very low for BAPL2. The poor result for BAPL2 can be explained as follows. As the preliminary stage of AD (BAPL3), BAPL2 (MCI) shares some similar visual features with AD (BAPL3) and possibly with NC (BAPL1). Therefore, BAPL2 images were easily misclassified as AD (BAPL3) or NC (BAPL1). After adding the intra-loss, the performance indices (precision, recall, F1) of BAPL2 (MCI) were improved from 40.25%, 17.50%, and 24.39% to 100%, 51.50%, and 67.99% in VGG19, from 73.7%, 40.20%, and 52.02% to 67.80%, 62.10%, and 64.82% in ResNet50, and from 67.79%, 53.00%, 59.49% to 85.32%, 72.70%, and 78.51% in DenseNet121. Moreover DenseNet121 achieved the highest performance among the three network architectures. The classification accuracy of DenseNet121 was approximately 94.86%. In a statistical analysis (i.e., t-test) of the accuracy comparisons of the proposed joint loss ($\lambda = 0.0025$) and conventional loss ($\lambda = 0.00$), the p-values for VGG19, ResNet50, and DenseNet121 were 0.0033, 0.0042, and 0.00001, respectively. As all p-values were less than 0.05, we statistically confirmed the superior performance of the proposed joint loss over the conventional loss in all three networks.
### Table 6. Comparison of classification performances between the proposed joint loss function and the conventional loss function in the three network architectures.

| Method | λ = 0  | λ = 0.0025 |
|--------|--------|------------|
|        | BAPL1  | Precision  | Recall  | F1 Score| Accuracy |
| VGG19  |        | 88.16%     | 91.60%  | 89.85%  | 81.50%   |
|        | BAPL2  | 40.25%     | 17.50%  | 24.39%  |          |
|        | BAPL3  | 78.52%     | 89.40%  | 83.61%  |          |
|        |        | 87.96%     | 93.20%  | 90.50%  |          |
|        | BAPL2  | 100%       | 51.50%  | 67.99%  | 88.60%   |
|        | BAPL3  | 87.51%     | 94.90%  | 91.06%  |          |
| ResNet50 |        |            |         |         |          |
|        | λ = 0  | BAPL1  | 92%     | 92.60%  | 92.40%   |
|        | BAPL2  | 73.70%    | 40.20%  | 52.02%  |          |
|        | BAPL3  | 81.90%    | 94.10%  | 87.58%  |
|        |        | 93.63%    | 92.00%  | 92.81%  |
|        | BAPL2  | 67.80%    | 62.10%  | 64.82%  |
|        | BAPL3  | 89.80%    | 92.00%  | 90.89%  |
|        |        | 97.38%    | 88.30%  | 92.62%  |
|        | BAPL2  | 67.79%    | 53.00%  | 59.49%  |
|        | BAPL3  | 89.76%    | 99.40%  | 94.33%  |
| DenseNet121 |        |            |         |         |          |
|        | λ = 0  | BAPL1  | 90.91%  |         |          |
|        | BAPL2  | 92.42%    |         |          |
|        | BAPL3  | 88.90%    |         |          |
|        | λ = 0.0025 | BAPL1 | 96.97% | 95.55% |
|        | BAPL2  | 92.42%    |         |          |

### 4.6. Two-Class Classification

We also calculated the accuracy, receiver operating characteristic curve (ROC), and area under curve (AUC) in two-class classifications: BAPL3 + BAPL2 versus BAPL1 and BAPL3 versus BAPL1 + BAPL2. In the first case, both BAPL3 and BAPL2 were treated as positive, whereas in the second case, only BAPL3 was treated as positive. The ROC curves and AUCs of BAPL3 + BAPL2 versus BAPL1 and BAPL3 versus BAPL1 + BAPL2 are shown in Figure 6, and the two-class classification accuracies are given in Table 7. Both the AUCs and accuracies of cases 1 and 2 were significantly higher for the proposed joint loss function (λ = 0.0025) than for the conventional loss function (λ = 0.00).

### Table 7. Comparison of two-class classification accuracies.

| Method | BAPL1 versus (BAPL2 + BAPL3) | (BAPL1 + BAPL2) versus BAPL3 |
|--------|-------------------------------|-------------------------------|
| λ = 0  | 90.91%                        | 92.42%                        |
| λ = 0.0025 | 96.97%                       | 95.55%                        |
Figure 6. ROC curves and AUCs of the conventional loss (top) and the proposed joint loss (bottom).

4.7. Data Visualization and Discussion

To confirm the effectiveness of the proposed joint loss function, we reduced the dimension of the feature vector from 1779 to 2 using a fully connected layer [25]. The distributions of training samples represented by the reduced two-dimensional features are visualized in Figures 7–9. Figures 7–9 are results obtained by VGG19, ResNet50, and DenseNet121, respectively. Different classes of samples are represented by different colors.

Figure 7. Distributions of training samples, represented by the reduced two-dimensional VGG19 features: (a) conventional loss function; (b) proposed joint loss function. Left: training data; right: test data.
As shown in Figure 7a, Figure 8a, and Figure 9a, when we only used the conventional loss function, parts of the BAPL2 data (yellow) were mixed with BAPL3 (red) and BAPL1 (green) in the test samples. On the other hand, the BAPL2 samples (yellow) were separated from those of BAPL1 (red) and BAPL3 (green) by the use of the proposed joint loss function as shown in Figure 7b, Figure 8b, and Figure 9b. From these results, we can easily understand why the classification performance, especially for BAPL2, was significantly improved by the proposed joint loss function.
5. Conclusions

To improve the classification accuracy of Alzheimer’s disease by deep learning, we proposed a joint discriminative loss function that adds a discriminative intra-loss function to the conventional loss function (the cross-entropy loss function). Detailed comparisons between the proposed joint loss and conventional loss functions were performed on three state-of-the-art deep neural network architectures. In the experiments, the intra-loss function significantly improved the classification performance. Especially, the classification accuracy of intermediate brain amyloid plaque load (BAPL2) was greatly improved. In addition, we showed that the mix-up data augmentation method, originally proposed for natural image classification, was also useful in medical image classification. The proposed joint loss can be combined with any deep learning network, including the 3D CNN [26] and the combined CNN and RNN network [13], which includes the 3D information for accurate classification. The weakness of this study was the small sample size, especially for BAPL 2. In future work, we will increase the number of samples and apply the proposed joint loss function to a 3D CNN.

Author Contributions: Software development and experiments, R.S.; methodology and analysis, Y.I. and K.C.; data collection and validation, D.-Y.K.; and conceptualization and validation, Y.-W.C.; funding acquisition, K.C. and D.-Y.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in part by the project at Institute of Convergence Bio-Health, Dong-A University; funded by Busan Institute of S&T Evaluation and Planning and in part by the Grant-in Aid for Scientific Research from the Japanese Ministry for Education, Science, Culture and Sports (MEXT) under the Grant No. 18H03267 and No. 18K18078.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. World Alzheimer Report. Available online: https://www.alz.co.uk/research/WorldAlzheimerReport2015.pdf (accessed on 13 November 2019).
2. Cheng, D.; Liu, M. Combining Convolutional and Recurrent Neural Networks for Alzheimer’s Disease Diagnosis Using PET Images. In Proceedings of the 2017 IEEE International Conference on Imaging Systems and Techniques, Beijing, China, 18–20 October 2017.
3. Plant, C.; Teipel, S.J.; Oswald, A.; Böhm, C.; Meindl, T.; Mourao-Miranda, J.; Bokde, A.W.; Hampel, H.; Ewers, M. Automated detection of brain atrophy patterns based on MRI for the prediction of Alzheimer’s disease. Neuroimage 2010, 50, 162–174. [CrossRef] [PubMed]
4. Dubois, B.; Feldman, H.H.; Jacova, C.; Hampel, H.; Molinuevo, J.L.; Blennow, K.; DeKosky, S.T.; Gauthier, S.; Selkoe, D.; Bateman, R.; et al. Advancing research diagnostic criteria for Alzheimer’s disease: The IWG-2 criteria. Lancet Neurol. 2014, 13, 614–629. [CrossRef]
5. Pillai, J.A.; Cummings, J.L. Clinical trials in predementia stages of Alzheimer disease. Med. Clin. N. Am. 2013, 97, 439–457. [CrossRef] [PubMed]
6. Jack, C.R.; Knopman, D.S.; Jagust, W.J.; Petersen, R.C.; Weiner, M.W.; Aisen, P.S.; Shaw, L.M.; Vemuri, P.; Wiste, H.J.; Weigand, S.D.; et al. Tracking pathophysiological processes in Alzheimer’s disease: An updated hypothetical model of dynamic biomarkers. Lancet Neurol. 2013, 12, 207–216. [CrossRef]
7. Choi, H.; Jin, K.H. Predicting cognitive decline with deep learning of brain metabolism and amyloid imaging. Behav. Brain Res. 2018, 344, 103–109. [CrossRef] [PubMed]
8. Barthel, H.; Gertz, H.J.; Dresel, S.; Peters, O.; Bartenstein, P.; Buerger, K.; Hiemeyer, F.; Wittemer-Rump, S.M.; Seibyl, J.; Reininger, C.; et al. Cerebral amyloid-β PET with florbetaben (18F) in patients with Alzheimer’s disease and healthy controls: A multicentre phase 2 diagnostic study. Lancet Neurol. 2011, 10, 424–435. [CrossRef]
9. Zhang, D.; Wang, Y.; Zhou, L.; Yuan, H.; Shen, D. Multimodel classification of Alzheimer’s disease and mild cognitive impairment. Neuroimage 2011, 55, 856–867. [CrossRef] [PubMed]
10. Sarraf, S.; Tofighi, G. DeepAD: Alzheimer’s disease classification via deep convolutional neural networks using MRI and fMRI. bioRxiv 2016. [CrossRef]
11. Farooq, A.; Anwar, S.; Muhammad Awais, M.; Saad Rehman, S. A Deep CNN based Multi-class Classification of Alzheimer’s Disease Using MRI. In Proceedings of the IEEE Instrumentation and Measurement, Beijing, China, 18–20 October 2017.

12. Kang, H.; Kim, W.G.; Yang, G.S.; Kim, H.W.; Jeong, J.E.; Yoon, H.J.; Cho, K.; Jeong, Y.J.; Kang, D.Y. VGG-based BAPL Score Classification of 18F-Florbetaben Amyloid Brain PET. *Biomed. Sci. Lett.* 2018, 24, 418–425. [CrossRef]

13. Liu, M.; Cheng, D.; Yan, W. Classification of Alzheimer’s Disease by Combination of Convolutional and Recurrent Neural Networks Using FDG-PET Images. *Front. Neuroinform.* 2018, 12, 35. [CrossRef] [PubMed]

14. Zhang, H.; Cissé, M.; Lopez-Paz, D. Mix-Up: Beyond Empirical Risk Minimization. In Proceedings of the International Conference on Learning Representations (ICLR), Vancouver, BC, Canada, 30 April–3 May 2018.

15. Friston, K.J.; Ashburner, J.; Frith, C.D.; Poline, J.B.; Heather, J.D.; Frackowiak, R.S.J. Spatial registration and normalization of images. *Hum. Brain Map.* 1995, 3, 165–189. [CrossRef]

16. Sato, R.; Iwamoto, Y.; Cho, K.; Kang, D.Y.; Chen, Y.W. Comparison of CNN Models with Different Plane Images and Their Combinations for Classification of Alzheimer’s Disease Using PET Images. In *Innovation in Medicine and Healthcare Systems, and Multimedia*; Chen, Y.W., Zimmermann, A., Howlett, R., Jain, L., Eds.; Smart Innovation, Systems and Technologies (Proc. of InMed2019); Springer: St. Julians, Malta, 2019; Volume 145, pp. 169–177.

17. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the International Conference on Learning Representations (ICLR), Sandiego, CA, USA, 7–9 May 2015.

18. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. *arXiv 2015*, arXiv:1512.03385.

19. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1097–1105.

20. Liang, D.; Lin, L.; Hu, H.; Zhang, Q. Combining Convolutional and Recurrent Neural Networks for Classification of Focal Liver Lesions in Multi-Phase CT Images. In *MICCAI 2018, LNCS 11071*; Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G., Eds.; Springer: Cham, Switzerland, 2018; pp. 666–675.

21. Wen, Y.; Zhang, K.; Li, Z.; Qiao, Y. A discriminative feature learning approach for deep face recognition. In *ECCV2016, LNCS*; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer: Cham, Switzerland, 2016; Volume 9911, pp. 499–515.

22. Chollet, F.; Rahman, F.; Lee, T.; Marmiesse, G.; Zabluda, O.; Pumperla, M.; Santana, E.; McColgan, T.; Snellgrove, X.; Branchaud-Charron, F.; et al. Keras: The Python Deep Learning Library. 2015. Available online: https://keras.io (accessed on 13 November 2019).

23. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv 2016*, arXiv:1603.04467.

24. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Li, F.F. Imagenet: A large-scale hierarchical image database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009.

25. Center Loss Implementation in Keras. Available online: https://github.com/handongfeng/MNIST-center-loss (accessed on 13 November 2019).

26. Payan, A.; Montana, G. Predicting Alzheimer’s Disease: A Neuroimaging Study with 3D Convolutional Neural Networks. *arXiv 2015*, arXiv:1502.02506.