

# Learning Combinations of Activation Functions

Franco Manessi

Strategic Analytics

lastminute.com group

Chiasso, Switzerland

Email: franco.manessi@lastminute.com

Alessandro Rozza

Strategic Analytics

lastminute.com group

Chiasso, Switzerland

Email: alessandro.rozza@lastminute.com

**Abstract**—In the last decade, an active area of research has been devoted to design novel activation functions that are able to help deep neural networks to converge, obtaining better performance. The training procedure of these architectures usually involves optimization of the weights of their layers only, while non-linearities are generally pre-specified and their (possible) parameters are usually considered as hyper-parameters to be tuned manually. In this paper, we introduce two approaches to automatically learn different combinations of base activation functions (such as the identity function, ReLU, and tanh) during the training phase. We present a thorough comparison of our novel approaches with well-known architectures (such as LeNet-5, AlexNet, and ResNet-56) on three standard datasets (Fashion-MNIST, CIFAR-10, and ILSVRC-2012), showing substantial improvements in the overall performance, such as an increase in the top-1 accuracy for AlexNet on ILSVRC-2012 of 3.01 percentage points.

## I. INTRODUCTION

In the last decade, deep learning has achieved remarkable results in computer vision, speech recognition and natural language processing, obtaining in some tasks human-like [1] or even super-human [2] performance. The roots of recent successes of deep learning can be found in: (i) the increase of the data available for training neural networks, (ii) the rising of commodity computational power needed to crunch the data, (iii) the development of new techniques, architectures, and activation functions that improve convergence during training of deeper networks, overcoming the obstacle of vanishing/exploding gradient [3], [4].

For many years, neural networks have usually employed logistic sigmoid activation functions. Unfortunately, this activation is affected by saturation issues. This problem reduces its effectiveness and, nowadays, its usage in feedforward networks is discouraged [5]. To overcome such weakness and improve accuracy results, an active area of research has been devoted to design novel activation functions.

The training procedure of these architectures usually involve optimization of the weights of their layers only, while non-linearities are generally pre-specified and their (possible) parameters are usually considered as hyper-parameters to be tuned manually. In this paper, we introduce two approaches able to automatically learn combinations of base activation functions (such as: the identity function, ReLU, and tanh) during training; our aim is to identify a search space for

the activation functions by means of convex combination or affine combination of the base functions. To the best of our knowledge, this is one of the first attempts to automatically combine and optimize activation functions during the training phase.

We tested different well-known networks employing customary activation functions on three standard datasets and we compared their results with those obtained applying our novel approaches. The techniques proposed in this paper outperformed the baselines in all the experiments, even when using deep architectures: we found a 3.01 percentage points increase in the top-1 accuracy for AlexNet on ILSVRC-2012.

This paper is organized as follows: in Section II the related works are summarized; in Section III the proposed methods are described; in Section IV the experimental results are presented; finally, conclusions and future works are summarized in Section V.

## II. RELATED WORK

Designing activation functions that enable fast training of accurate deep neural networks is an active area of research. The rectified linear activation function, introduced in [6] and argued in [7] to be a better biological model than logistic sigmoid activation function, has eased the training of deep neural networks by alleviating the problems related to weight initialization and vanishing gradient. Slight variations of ReLU have been proposed over the years, such as *leaky ReLU* (LReLU), [8], which addresses *dead neuron* issues in ReLU networks, *thresholded ReLU* [9], which tackles the problem of large negative biases in autoencoders, and *parametric ReLU* (PReLU) [10], which treats the leakage parameter of LReLU as a per-filter learnable weight.

A smooth version of ReLU, called *softplus*, has been proposed in [11]. Despite some theoretical advantages over ReLU (it is differentiable everywhere and it has less saturation issues), this activation function usually does not achieve better results [7] when compared to its basic version.

More recently, *maxout* has been introduced in [12] as an activation function aimed at enhancing dropouts abilities as a model averaging technique. Among its extensions, it is worth mentioning the *probabilistic maxout* [13], and the  $L_p$  norm pooling activation [14] that is able to recover the maxout activation when  $p \rightarrow \infty$ .

Considering the last developments of activation functions in neural networks, it is important to mention the *exponential linear unit function* (ELU) [15] and the *scaled exponential linear unit function* (SELU) [16]. Like ReLU, LReLU, and PReLU, ELU reduces the vanishing gradient problem. Furthermore, ELU has negative values, allowing to push mean unit activations closer to zero like batch normalization, and speeding up the learning. SELU extends this property ensuring that activations close to zero mean and unit variance that are propagated through many network layers will converge towards zero mean and unit variance, even under the presence of noise and perturbations.

With the exception of PReLU, all previous activations are pre-specified (i.e. non-learnable). A first attempt to learn activations in a neural network can be found in [17], where the authors propose to randomly add or remove logistic or Gaussian activation functions using an evolutionary programming method. On the other hand, in [18]–[21] the authors proposed novel approaches to learn the best activation function per neuron among a pool of allowed activations by means of genetic and evolutionary algorithms.

A different method has been proposed in [22], which is able to learn the *hardness* parameter of a sigmoid function, similarly to the approach employed by PReLU to learn the leakage parameter.

However, all the previous learning techniques are limited by the fact that the family of functions over which the learning takes place is either finite or a simple parameterization of customary activation functions.

Recently, in [23] the authors tackle the problem from a different angle using piecewise linear activation functions that are learned independently for each neuron using gradient descent. However, (i) the number of linear pieces is treated as a hyper-parameter of the optimization problem; (ii) the number of learned parameters increases proportionally to the amount of hidden units; (iii) all the learned piecewise linear activation functions are  $\text{ReLU}(x)$  for  $x$  large enough (i.e. there exists  $u \in \mathbb{R}$  such that  $g(x) = \text{ReLU}(x)$  for  $x \geq u$ ), thus reducing the expressivity of the learned activation functions by design.

It is worth mentioning that, in [24] the authors propose the *network in network* approach where they replace activation functions in convolutional layers with small multi-layer perceptrons.

In this paper we try to overcome some of the limitations of the aforementioned approaches. Indeed, the two techniques explained in Section III: (i) increase the expressivity power of the learned activation functions by enlarging the hypothesis space explored during training with respect to [17]–[22]; (ii) restrict the hypothesis space with respect to [23], [24], in order to allow a faster training without the need of a careful initialization of network weights (see Proposition 1 and the following lines).

### III. METHODS

A neural network  $N_d$  made of  $d$  hidden layers can be seen as the functional composition of  $d$  functions  $L_i$  followed by a

final mapping  $\bar{L}$  that depends on the task at hand (e.g. classification, regression):  $N_d = \bar{L} \circ L_d \circ \dots \circ L_1$ . In particular, each hidden layer function  $L_i$  can be written as the composition of two functions,  $g_i$  followed by  $\sigma_i$ , the former being a suitable remapping of the layer input neurons, the latter being the activation function of the layer:  $L_i = \sigma_i \circ g_i$ . In the most general case, both  $\sigma_i$  and  $g_i$  are parameterized and belongs to some hypothesis spaces  $\mathcal{H}_{\sigma_i}$  and  $\mathcal{H}_{g_i}$ , respectively. Hence, the learning procedure of  $L_i$  amounts to an optimization problem over the layer hypothesis space  $\mathcal{H}_i = \mathcal{H}_{\sigma_i} \times \mathcal{H}_{g_i}$ .

Usually,  $\sigma_i$  is taken as a non-learnable function; therefore, in the most common scenario  $\mathcal{H}_{\sigma_i}$  is a singleton:  $\mathcal{H}_i = \{\sigma_i\} \times \mathcal{H}_{g_i}$ . For example, for a fully-connected layer from  $\mathbb{R}^{n_i}$  to  $\mathbb{R}^{m_i}$  with ReLU activation we have that  $\mathcal{H}_{g_i}$  is the set of all *affine transformations* from  $\mathbb{R}^{n_i}$  to  $\mathbb{R}^{m_i}$ , i.e.  $\mathcal{H}_i = \{\text{ReLU}\} \times \text{Lin}(\mathbb{R}^{n_i}, \mathbb{R}^{m_i}) \times \text{K}(\mathbb{R}^{m_i})$ , where  $\text{Lin}(\mathbf{A}, \mathbf{B})$  and  $\text{K}(\mathbf{B})$  are the sets of linear maps between  $\mathbf{A}$  and  $\mathbf{B}$ , and the set of translations of  $\mathbf{B}$ , respectively.

In this paper, we introduce two techniques to define learnable activation functions that could be plugged in all hidden layers of a neural network architecture. The two approaches differ in how they define the hypothesis space  $\mathcal{H}_{\sigma_i}$ . Both of them are based on the following idea: (i) select a finite set of activation functions  $\mathcal{F} := \{f_1, \dots, f_N\}$ , whose elements will be used as base elements; (ii) define the learnable activation function  $\sigma_i$  as a linear combination of the elements of  $\mathcal{F}$ ; (iii) identify a suitable hypothesis space  $\mathcal{H}_{\sigma_i}$ ; (iv) optimize the whole network, where the hypothesis space of each hidden layer is  $\mathcal{H}_i = \mathcal{H}_{\sigma_i} \times \mathcal{H}_{g_i}$ .

We will now give some basic definitions used throughout the paper. Note that, hereinafter, all activation functions from  $\mathbb{R}$  to  $\mathbb{R}$  will naturally extend to functions from  $\mathbb{R}^n$  to  $\mathbb{R}^n$  by means of entrywise application.

Given a vector space  $\mathbf{V}$  and a finite subset  $\mathcal{A} \subseteq \mathbf{V}$ , we can define the following two subsets of  $\mathbf{V}$ :

(i) the *convex hull* of  $\mathcal{A}$ , namely:

$$\text{conv}(\mathcal{A}) := \{\sum_i c_i \mathbf{a}_i \mid \sum_i c_i = 1, c_i \geq 0, \mathbf{a}_i \in \mathcal{A}\};$$

(ii) the *affine hull* of  $\mathcal{A}$ , namely:

$$\text{aff}(\mathcal{A}) := \{\sum_i c_i \mathbf{a}_i \mid \sum_i c_i = 1, \mathbf{a}_i \in \mathcal{A}\}.$$

We remark that, neither  $\text{conv}(\mathcal{A})$  nor  $\text{aff}(\mathcal{A})$  are vector subspaces of  $\mathbf{V}$ . Indeed,  $\text{conv}(\mathcal{A})$  is just a generic convex subset in  $\mathbf{V}$  reducing to a  $(|\mathcal{A}| - 1)$ -dimensional simplex whenever the elements of  $\mathcal{A}$  are linearly independent. On the other hand,  $\text{aff}(\mathcal{A})$  is an *affine subspace* of  $\mathbf{V}$  of dimension  $|\mathcal{A}| - 1$ , i.e. for an arbitrary  $\bar{\mathbf{a}} \in \text{aff}(\mathcal{A})$  the set  $\{\mathbf{a} - \bar{\mathbf{a}} \mid \mathbf{a} \in \text{aff}(\mathcal{A})\}$  is a *linear subspace* of  $\mathbf{V}$  of dimension  $|\mathcal{A}| - 1$ . Clearly,  $\text{conv}(\mathcal{A}) \subset \text{aff}(\mathcal{A})$ .

Let  $\mathcal{F} := \{f_0, f_1, \dots, f_N\}$  be a finite collection of activation functions  $f_i$  from  $\mathbb{R}$  to  $\mathbb{R}$ . We can define a vector space  $\mathbf{F}$  from  $\mathcal{F}$  by taking all linear combinations  $\sum_i c_i f_i$ . Note that, despite  $\mathcal{F}$  is (by definition) a spanning set of  $\mathbf{F}$ , it is not generally a basis; indeed  $|\mathcal{F}| \geq \dim \mathbf{F}$ .

Since (almost everywhere) differentiability is a property preserved by finite linear combinations, and since  $\text{conv}(\mathcal{F}) \subset$

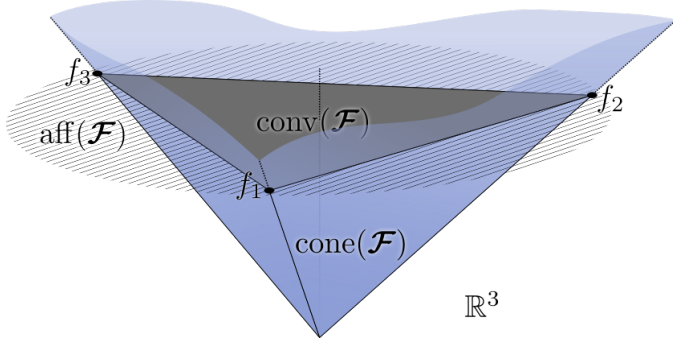


Fig. 1. The figure shows the relationship between affine hull, convex hull, and convex cone for a set  $\mathcal{F}$  made of three linearly independent elements:  $\mathcal{F} := \{f_1, f_2, f_3\}$ . Since  $\dim \mathbf{F} = |\mathcal{F}| = 3$ ,  $\text{conv}(\mathcal{F})$  is a 2-simplex (the gray triangle in the figure), while  $\text{aff}(\mathcal{F})$  is a plane within the three dimensional vector space  $\mathbf{F}$  (the plane identified by the dashed circle in the figure).  $\text{cone}(\mathcal{F})$  is the three dimensional manifold delimited by the three incident straight lines  $l_i := \{x \in \mathbf{F} \mid x = \alpha f_i, \forall 0 \leq \alpha \in \mathbb{R}\}$  for  $i = 1, 2, 3$ , i.e. the cone extremal rays. It can easily be seen that  $\text{conv}(\mathcal{F})$  corresponds to the intersection of  $\text{cone}(\mathcal{F})$  with  $\text{aff}(\mathcal{F})$ .

$\text{aff}(\mathcal{F}) \subseteq \mathbf{F}$ , assuming that  $\mathcal{F}$  contains (almost everywhere) differentiable activation functions,  $\text{conv}(\mathcal{F})$  and  $\text{aff}(\mathcal{F})$  are made of (almost everywhere) differentiable functions, i.e. valid activation functions for a neural network that can be learned by means of gradient descent.

The activations used in real world scenarios are usually monotonic increasing functions. Unfortunately, the monotonicity is not ensured under arbitrary linear combination, meaning that even if all  $f_i \in \mathcal{F}$  are non-decreasing functions, an arbitrary element  $\bar{f} \in \mathbf{F}$  might be neither a non-decreasing nor a non-increasing function. As a matter of fact, considering only non-decreasing differentiable functions ( $f'_i \geq 0 \forall f_i \in \mathcal{F}$ ), all non-decreasing differentiable functions in  $\mathbf{F}$  lie inside the *convex cone*  $\text{cone}(\mathcal{F}) \subset \mathbf{F}$ , i.e.:

$$\text{cone}(\mathcal{F}) := \{\sum_i c_i f_i \mid c_i \geq 0, f_i \in \mathcal{F}\}.$$

Indeed,  $\forall g \in \text{cone}(\mathcal{F})$  we have that  $g' \geq 0$ . Thanks to the definition of  $\text{aff}(\mathcal{F})$ ,  $\text{cone}(\mathcal{F})$ , and  $\text{conv}(\mathcal{F})$ , we can conclude that  $\text{conv}(\mathcal{F}) = \text{cone}(\mathcal{F}) \cap \text{aff}(\mathcal{F})$ , which implies that monotonicity of the elements of  $\mathcal{F}$  is preserved by all the elements of  $\text{conv}(\mathcal{F})$  but not by  $\text{aff}(\mathcal{F})$  (see Figure 1). Nevertheless, in [25] it is shown that even non-monotonic activation functions can approximate arbitrarily complex functions for sufficiently large neural networks. Indeed, in [5] the authors trained a feedforward network using cosine activation functions on the MNIST dataset obtaining an error rate smaller than 1%. Therefore, also  $\text{aff}(\mathcal{F})$  is a proper candidate for  $\mathcal{H}_{\sigma_i}$ .

In addition, the affine subspace  $\text{aff}(\mathcal{F})$  enjoys the following property.

**Proposition 1.** *Let all  $f_i \in \mathcal{F}$  be linearly independent and approximate the identity function near origin (i.e.  $f_i(0) = 0$  and  $f'_i(0) = 1$ ), then,  $\bar{f} \in \mathbf{F}$  approximates the identity if and only if  $\bar{f} \in \text{aff}(\mathcal{F})$ .*

*Proof.* The proof is immediate. Let us expand  $\mathbf{F} \ni \bar{f}$  as  $\bar{f} = \sum_i c_i f_i$  (by hypothesis the  $f_i$  form a basis) and prove the

two-way implication.

( $\Rightarrow$ ) By hypothesis,  $\bar{f}(0) = 0$  and  $\bar{f}'(0) = 1$ . The relation on the derivative reads  $\sum_i c_i f'_i(0) = \sum_i c_i = 1$ . In turns, this implies that  $\bar{f} = \sum_i c_i f_i$  with  $\sum_i c_i = 1$ , namely  $\bar{f} \in \text{aff}(\mathcal{F})$ .  
 ( $\Leftarrow$ ) By hypothesis,  $\bar{f} = \sum_i c_i f_i$  with  $\sum_i c_i = 1$ . Hence,  $\bar{f}(0) = \sum_i c_i f_i(0) = 0$  and  $\bar{f}'(0) = \sum_i c_i f'_i(0) = \sum_i c_i = 1$ , namely  $\bar{f}$  approximates the identity near origin.  $\square$

Since  $\text{conv}(\mathcal{F}) \subset \text{aff}(\mathcal{F})$ , also  $\text{conv}(\mathcal{F})$  enjoys the same property. It is important to underline that, activation functions approximating the identity near origin, have been argued to be desirable to train deep networks randomly initialized with weights near zero [26]. Indeed, for such an activation function the initial gradients will be relatively large, thus speeding up the training phase. Note that, such behavior is enjoyed also by ReLU, which approximates the identity only from one side: i.e.  $x \rightarrow 0^+$ . Furthermore, training deep networks with activation functions not approximating the identity near origin requires a more careful initialization, in order to properly scale the layers input [27].

For the aforementioned reasons: (i) preserved differentiability; (ii) absence of the requirement of monotonicity; (iii) approximation of the identity near origin; both  $\text{aff}(\mathcal{F})$  and  $\text{conv}(\mathcal{F})$  are good candidates for  $\mathcal{H}_{\sigma_i}$ .

Thanks to the previous definitions and the aforementioned properties, we can now formalize the two techniques to build learnable activation functions as follows:

- (i) choose a finite set  $\mathcal{F} = \{f_1, \dots, f_N\}$ , where each  $f_i$  is a (almost everywhere) differentiable activation function approximating the identity near origin (at least from one side);
- (ii) define a new activation function  $\bar{f}$  as a linear combination of all the  $f_i \in \mathcal{F}$ ;
- (iii) select as hypothesis space  $\mathcal{H}_{\bar{f}}$  the sets  $\text{aff}(\mathcal{F})$  or  $\text{conv}(\mathcal{F})$ .

In Section IV, we present results using the following choices for  $\mathcal{F}$ :

$$\begin{aligned} \mathcal{F} &:= \{\text{id}, \text{ReLU}\}, & \mathcal{F} &:= \{\text{id}, \tanh\}, \\ \mathcal{F} &:= \{\text{ReLU}, \tanh\}, & \mathcal{F} &:= \{\text{id}, \text{ReLU}, \tanh\}, \end{aligned} \quad (1)$$

where  $\text{id}$  is the identity function. Clearly, other choices of  $\mathcal{F}$  may be available, provided the requirements in (i) are satisfied.

Since  $\text{conv}(\mathcal{F}) \subset \text{aff}(\mathcal{F})$ , the convex hull-based technique can to be understood as a *regularized version* of the affine hull-based one, where the corresponding hypothesis space has been explicitly constrained to be *compact*. Such a regularization, in addition to restrict the complexity of the hypothesis space, guarantees that the final activation function is monotonic (provided all  $f_i \in \mathcal{F}$  are monotonic as well). Moreover, the convex hull-based technique together with  $\mathcal{F} := \{\text{id}, \text{ReLU}\}$  recovers the learnable LReLU activation function, i.e.  $\text{LReLU}_\alpha(x) = x$  if  $x \geq 0$ , while  $\text{LReLU}_\alpha(x) = \alpha x$  otherwise, where  $0 \leq \alpha \ll 1$  (usually  $\alpha = 10^{-2}$ ). Indeed,  $\text{conv}(\mathcal{F}) = \{\bar{f} := p \cdot \text{id} + (1-p) \cdot \text{ReLU} \text{ with } 0 \leq p \leq 1\}$  and since  $\text{ReLU} = \text{id}$  for  $x \geq 0$  and  $\text{ReLU} = 0$  otherwise,

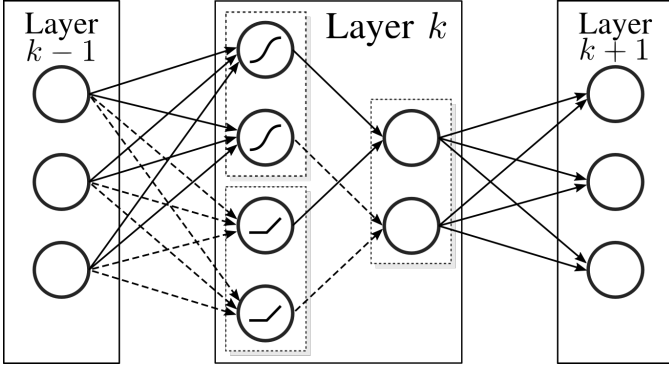


Fig. 2. The figure shows how a neural network layer with convex/affine hull-based activation can be seen as a two-stage pipeline. Here the  $k$ -th layer is a two neurons layer, with set  $\mathcal{F}$  composed of two base functions. The first stage of the pipeline is made of two stacked layers, each one featuring only one activation belonging to  $\mathcal{F}$ . The weights between the two stacked layers are shared. The second stage of the pipeline is a 1D-convolution with kernel of size 1 between the stacked layers, which play the role of  $n$  channels for the convolution. The weights of the convolution are constrained to sum to one, and also to be positive when using the convex hull-based technique.

we have that  $\bar{f} = p \cdot \text{id} + (1-p) \cdot \text{id} = \text{id}$  for  $x \geq 0$  and  $\bar{f} = p \cdot \text{id} + (1-p) \cdot 0 = p \cdot \text{id}$  otherwise, i.e.  $\text{LReLU}_p$ .

It is worth mentioning that, as shown in Figure 2, layers using convex hull-based and affine hull-based activations with  $n$  base functions can also be seen as the following two-stage pipeline: (i)  $n$  stacked fully connected layers, each of them featuring one of the base functions, and all of them sharing weights; (ii) a 1D-convolution with kernel of size 1 between the  $n$  stacked layers (which are treated by the convolution as  $n$  separate channels), whose weights are constrained to sum to one (and to be positive in case of the convex hull-based technique)<sup>1</sup>.

#### IV. RESULTS

We tested the convex hull-based and the affine hull-based approaches by evaluating their effectiveness on three publicly available datasets used for image classification, greatly differing in the number of input features and examples. Moreover, each dataset was evaluated using different network architectures. These networks were trained and tested using as activation functions (for all their hidden layers) those learned by the convex hull-based and the affine hull-based approaches combining the base activations reported in Equation (1). In addition, the base activation functions alone and  $\text{LReLU}$  were also employed in order to compare the overall performance.

Specifically, the datasets we used are:

**Fashion-MNIST:** it is a dataset of Zalando’s article images, composed by a training and test sets of 60k and 10k examples, respectively. Each example is a 28x28 grayscale image, associated with a label that belongs to 10 classes [28]. We divided each pixel value by 255, and we augmented the resulting dataset during training phase by means of random horizontal flip and image shifting;

<sup>1</sup>The 1D-convolution with kernel of size 1 can also be seen as a weighted average of the stacked fully connected layers (with possibly negative weights in case of the affine hull-based technique).

TABLE I  
KERASNET ARCHITECTURE

Id	Layers	Properties
conv_1	2D convolution Activation	$32 \times (3, 3)$
conv_2	2D convolution Activation	$32 \times (3, 3)$
	Max pooling Dropout	$(2, 2)$ 25%
conv_3	2D convolution Activation	$64 \times (3, 3)$
conv_4	2D convolution Activation	$64 \times (3, 3)$
	Max pooling Dropout	$(2, 2)$ 25%
fc_5	Fully connected Activation	512
	Dropout	50%
fc_6	Fully connected Activation	10 softmax

**CIFAR-10:** it is a balanced dataset of 60k 32x32 colour images belonging to 10 classes. There are 50k training images and 10k test images [29]. Also in this case we divided each pixel value by 255, and we augmented the resulting dataset during training phase by means of random horizontal flip and image shifting;

**ILSVRC-2012:** a 1k classes classification task with 1.2M training examples and 50k validation examples. The examples are colour images of various sizes [30]. We resized each image to 256x256 pixels, we subtracted pixel mean values, and we augmented the dataset during training phase by randomly cropping to 227x227 pixels and horizontally flipping the images. Note that, we did not train our network with the relighting data-augmentation proposed in [31].

The considered architectures are the following:

**LeNet-5:** a convolutional network made of two convolutional layers followed by two fully connected layers [32]. The convolutional layers have respectively 20 and 50 filters of size  $5 \times 5$ , while the hidden fully connected layer is made of 500 neurons. Max pooling with size  $2 \times 2$  is used after each convolutional layer, without dropout. We assessed LeNet-5 on Fashion-MNIST and CIFAR-10 datasets, resulting in two networks with 431k and 657k parameters, respectively;

**KerasNet:** a convolutional neural network included in the Keras framework [33]. It is made of 4 convolutional layers and 2 fully connected layers, and it employs both max pooling and dropout. This architecture is presented in Table I. We tested KerasNet on both Fashion-MNIST and CIFAR-10 datasets, resulting in two networks with 890k and 1.2M parameters, respectively;

**ResNet-56:** a residual network made of 56 total layers, employing pooling and skip connection [34]. Its performance has been evaluated on CIFAR-10, corresponding to a network with 858k parameters;

**AlexNet:** a convolutional network made of 5 convolutional and 3 fully connected layers [31]. We tested it against ILSVRC-2012 dataset, resulting in a network with 61M pa-

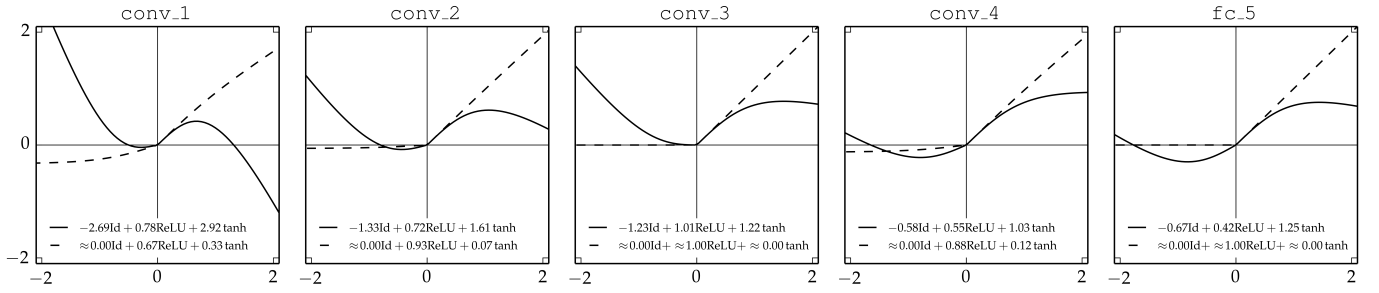


Fig. 3. The figure shows the hidden affine hull-based and convex hull-based activation functions learned during the training of the KerasNet architecture on CIFAR-10 dataset. Activations learned on  $\text{aff}(\{\text{id}, \text{ReLU}, \text{tanh}\})$  are represented with solid lines, while those learned on  $\text{conv}(\{\text{id}, \text{ReLU}, \text{tanh}\})$  are dashed.

rameters. Note that, as shown in [31], ReLU-based activation functions significantly outperform networks based on other activations. Therefore, in this context ReLU-based networks only have been built and tested for comparison.

All networks were implemented from scratch using the Keras framework [33] on top of TensorFlow [35]. Notice that, our AlexNet implementation is a porting to Keras of the *Caffe* architecture<sup>2</sup>. We trained LeNet-5, KerasNet, and ResNet-56 using RMSprop with learning rate  $10^{-4}$  and learning rate decay over each mini-batch update of  $10^{-6}$ . For AlexNet we used SGD with starting learning rate  $10^{-2}$ , step-wise learning rate decay, weight-decay hyper-parameter  $5 \cdot 10^{-4}$ , and momentum 0.9.

Table II shows the top-1 accuracy for all the run experiments. The best configurations (shaded cells in the table) are always achieved using our techniques, where in 5 out of 6 experiments the affine hull-based approach outperformed convex hull-based ones. The uplift in top-1 accuracy using our approaches compared to customary activations goes from 0.69 percentage points (pp) for LeNet-5 on Fashion-MNIST up to 4.59 pp for KerasNet on CIFAR-10. It is worth mentioning that, even in deep neural networks, such as AlexNet, a substantial increase is observed, i.e. 3.01 pp on ILSVRC-2012.

Moreover, the proposed techniques usually achieve better results than their corresponding base activation functions (boldface in the table). Note that, the novel activations work well for very deep networks and also with various architectures involving different types of layer (e.g. residual unit, dropout, pooling, convolutional, and fully connected).

Furthermore, our experiments show how the learning of the leakage parameter achieved by the activation based on  $\text{conv}(\{\text{id}, \text{ReLU}\})$  allows to outperform or to achieve the same results of LReLU.

Figure 3 shows activations learned by KerasNet on CIFAR-10 when using convex hull-based and affine hull-based activations with  $\mathcal{F} = \{\text{id}, \text{ReLU}, \text{tanh}\}$ . It is possible to notice that, as already theoretically proved in Section III, convex combinations preserved monotonicity of the base activation functions while affine ones did not.

<sup>2</sup>See [https://github.com/BVLC/caffe/tree/master/models/bvlc\\_alexnet](https://github.com/BVLC/caffe/tree/master/models/bvlc_alexnet).

## V. CONCLUSION

In this paper we introduced two novel techniques able to learn new activations starting from a finite collection  $\mathcal{F}$  of base functions. Both our ideas are based on building an arbitrary linear combination of the elements of  $\mathcal{F}$  and on defining a suitable hypothesis space where the learning procedure of the linear combination takes place. The hypothesis spaces for the two techniques are  $\text{conv}(\mathcal{F})$  and  $\text{aff}(\mathcal{F})$ . We showed that, provided all the elements of  $\mathcal{F}$  approximate the identity near origin,  $\text{aff}(\mathcal{F})$  is the only set where it is possible to find combined activations that also approximate  $\text{id}$  near origin. Moreover,  $\text{aff}(\mathcal{F})$  allows to explore non-monotonic activation functions, while  $\text{conv}(\{\text{id}, \text{ReLU}\})$  may be seen as a learnable LReLU activation function.

We tested the two techniques on various architectures (LeNet-5, KerasNet, ResNet-56, AlexNet) and datasets (Fashion-MNIST, CIFAR-10, ILSVRC-2012), comparing results with LReLU and single base activation functions.

In all our experiments, the techniques proposed in this paper achieved the best performance and the combined activation functions learned using our approaches usually outperform the corresponding base components. The effectiveness of the proposed techniques is further proved by the increase in performance achieved using networks with different depths and architectures.

In our opinion, an interesting extension of this work would be to analyze other sets ( $\mathcal{F}$ ) of base functions.

## ACKNOWLEDGMENT

The authors would like to thank Enrico Deusebio for the many insightful discussions.

## REFERENCES

- [1] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig, "Achieving human parity in conversational speech recognition," *arXiv preprint arXiv:1610.05256*, 2016.
- [2] D. C. Ciresan, U. Meier, J. Masci, and J. Schmidhuber, "A committee of neural networks for traffic sign classification," in *IJCNN*. IEEE, 2011, pp. 1918–1921.
- [3] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE transactions on neural networks*, vol. 5, no. 2, pp. 157–166, 1994.

TABLE II

EXPERIMENT RESULTS. THE TABLE SHOWS THE TOP-1 ACCURACY RESULTS FOR ALL THE ANALYZED NETWORKS ON FASHION-MNIST TEST SET, CIFAR-10 TEST SET, AND ILSVRC-2012 VALIDATION SET. CONVEX HULL-BASED AND AFFINE HULL-BASED ACTIVATIONS ACHIEVING TOP-1 ACCURACY RESULTS GREATER THAN THEIR CORRESPONDING BASE ACTIVATION FUNCTIONS ARE HIGHLIGHTED IN BOLDFACE. THE BEST RESULT FOR EACH NETWORK/DATASET IS SHADED.

Activation	Fashion-MNIST		CIFAR-10			ILSVRC-2012
	LeNet-5	KerasNet	LeNet-5	KerasNet	ResNet-56	AlexNet
id	90.50%	90.51%	75.27%	75.23%	42.60%	—
ReLU	91.06%	90.79%	80.37%	79.97%	89.18%	56.75%
tanh	92.33%	93.43%	78.96%	82.86%	82.65%	—
LReLU	91.03%	91.13%	80.94%	80.07%	89.38%	57.03%
conv({id, ReLU})	<b>91.87%</b>	<b>92.39%</b>	<b>80.94%</b>	<b>84.74%</b>	<b>90.51%</b>	<b>57.54%</b>
conv({id, tanh})	<b>92.36%</b>	<b>93.64%</b>	<b>79.45%</b>	78.53%	<b>86.46%</b>	—
conv({tanh, ReLU})	<b>92.56%</b>	92.04%	80.21%	<b>84.80%</b>	88.31%	<b>58.13%</b>
conv({id, ReLU, tanh})	92.21%	92.94%	<b>80.48%</b>	<b>85.21%</b>	<b>89.60%</b>	56.55%
aff({id, ReLU})	<b>92.83%</b>	<b>93.37%</b>	<b>80.52%</b>	<b>84.93%</b>	88.92%	<b>58.48%</b>
aff({id, tanh})	<b>92.65%</b>	<b>94.41%</b>	<b>78.97%</b>	<b>86.05%</b>	<b>82.97%</b>	—
aff({tanh, ReLU})	<b>93.02%</b>	<b>93.48%</b>	<b>81.23%</b>	<b>84.83%</b>	<b>89.44%</b>	<b>60.04%</b>
aff({id, ReLU, tanh})	<b>92.80%</b>	<b>94.41%</b>	80.13%	<b>87.45%</b>	88.62%	<b>57.20%</b>

- [4] X. Glorot and Y. Bengio, “Understanding the difficulty of training deep feedforward neural networks,” in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010, pp. 249–256.
- [5] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.
- [6] K. Jarrett, K. Kavukcuoglu, Y. LeCun *et al.*, “What is the best multi-stage architecture for object recognition?” in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 2146–2153.
- [7] X. Glorot, A. Bordes, and Y. Bengio, “Deep sparse rectifier neural networks,” in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 2011, pp. 315–323.
- [8] A. L. Maas, A. Y. Hannun, and A. Y. Ng, “Rectifier Nonlinearities Improve Neural Network Acoustic Models,” in *ICML Workshop on Deep Learning for Audio, Speech and Language Processing*, 2013.
- [9] K. Konda, R. Memisevic, and D. Krueger, “Zero-bias autoencoders and the benefits of co-adapting features,” *arXiv preprint arXiv:1402.3337*, 2014.
- [10] K. He, X. Zhang, S. Ren, and J. Sun, “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification,” in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ser. ICCV ’15. Washington, DC, USA: IEEE Computer Society, 2015, pp. 1026–1034. [Online]. Available: <http://dx.doi.org/10.1109/ICCV.2015.123>
- [11] C. Dugas, Y. Bengio, F. Bélisle, C. Nadeau, and R. Garcia, “Incorporating second-order functional knowledge for better option pricing,” in *Advances in neural information processing systems*, 2001, pp. 472–478.
- [12] I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio, “Maxout networks,” in *Proceedings of the 30th International Conference on International Conference on Machine Learning-Volume 28*. JMLR.org, 2013, pp. III–1319.
- [13] W. Sun, F. Su, and L. Wang, “Improving deep neural networks with multilayer maxout networks,” *2014 IEEE Visual Communications and Image Processing Conference, VCIP 2014*, pp. 334–337, dec 2015.
- [14] C. Gulcehre, K. Cho, R. Pascanu, and Y. Bengio, “Learned-norm pooling for deep feedforward and recurrent neural networks,” in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2014, pp. 530–546.
- [15] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, “Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs),” *Proceedings of ICLR 2016*, nov 2016.
- [16] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, “Self-normalizing neural networks,” in *Advances in neural information processing systems*, 2017.
- [17] Y. Liu and X. Yao, “Evolutionary design of artificial neural networks with different nodes,” in *Evolutionary Computation, 1996., Proceedings of IEEE International Conference on*. IEEE, 1996, pp. 670–675.
- [18] R. Poli, “Parallel distributed genetic programming, new ideas in optimization,” 1999.
- [19] D. Weingaertner, V. K. Tatai, R. R. Gudwin, and F. J. Von Zuben, “Hierarchical evolution of heterogeneous neural networks,” in *Evolutionary Computation, 2002. CEC’02. Proceedings of the 2002 Congress on*, vol. 2. IEEE, 2002, pp. 1775–1780.
- [20] A. J. Turner and J. F. Miller, “Cartesian genetic programming encoded artificial neural networks: a comparison using three benchmarks,” in *Proceedings of the 15th annual conference on Genetic and evolutionary computation*. ACM, 2013, pp. 1005–1012.
- [21] M. M. Khan, A. M. Ahmad, G. M. Khan, and J. F. Miller, “Fast learning neural networks using cartesian genetic programming,” *Neurocomputing*, vol. 121, pp. 274–289, 2013.
- [22] A. J. Turner and J. F. Miller, “Neuroevolution: evolving heterogeneous artificial neural networks,” *Evolutionary Intelligence*, vol. 7, no. 3, pp. 135–154, 2014.
- [23] F. Agostinelli, M. Hoffman, P. Sadowski, and P. Baldi, “Learning activation functions to improve deep neural networks,” in *ICLR Workshop*, 2015.
- [24] M. Lin, Q. Chen, and S. Yan, “Network in network,” *arXiv preprint arXiv:1312.4400*, 2013.
- [25] G. Cybenko, “Approximation by superpositions of a sigmoidal function,” *Mathematics of Control, Signals, and Systems (MCS)*, vol. 2, no. 4, pp. 303–314, 1989.
- [26] H. H. Aghdam and E. J. Heravi, *Guide to Convolutional Neural Networks*. Springer, 2017.
- [27] D. Sussillo and L. Abbott, “Random walk initialization for training very deep feedforward networks,” *arXiv preprint arXiv:1412.6558*, 2014.
- [28] H. Xiao, K. Rasul, and R. Vollgraf. (2017) Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms.
- [29] A. Krizhevsky and G. Hinton, “Learning multiple layers of features from tiny images,” Technical report, University of Toronto, Tech. Rep., 2009.
- [30] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, “ImageNet Large Scale Visual Recognition Challenge,” *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.
- [31] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
- [32] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- [33] F. Chollet *et al.*, “Keras,” <https://github.com/fchollet/keras>, 2015.
- [34] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [35] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, “Tensorflow: A system for large-scale machine learning,” in *OSDI*, vol. 16, 2016, pp. 265–283.