UNIVERSITY OF CALIFORNIA

Los Angeles

Discovering strong gravitational lensing with deep learning

A thesis submitted in partial satisfaction

of the requirements for the degree Master of Science

in Statistics

by

Ablaikhan Akhazhanov

2018

ABSTRACT OF THE THESIS

Discovering strong gravitational lensing

with deep learning

by

Ablaikhan Akhazhanov

Master of Science in Statistics

University of California, Los Angeles, 2018

Professor Chad J Hazlett, Chair

The thesis focuses on deep learning methods applied to discovery of gravitational lensing events in the universe. Publicly available I-band images of the known gravitational lenses were combined with simulated ones and randomly sampled cutouts of the galaxies and stars. Deep convolutional networks outperform the conventional discovery methods and achieve up to 0.9984 mean ROC AUC and 0.9895 mean F1-score on the out-of-sample 7-fold cross-validation. The models demonstrated excellent agreement with the latest list of 92 candidates published in the literature and created with combination of deep learning and manual analysis by professional astronomers.

The thesis of Ablaikhan Akhazhanov is approved.

Ying Nian Wu

Arash Ali Amini

Chad J Hazlett, Committee Chair

University of California, Los Angeles

2018

To my family who gave me love and support

TABLE OF CONTENTS

LIST OF ACRONYMS

Adam	Adaptive Moment Estimation
ANN	Artificial Neural Network
CASTLES	Cfa-Arizona Space Telescope Lens Survey of gravitational lenses
CIFAR-10	Collection of images that are commonly used to train machine learning
CNN	Convolutional Neural Network
CV	Cross-Validation
EM	Expectation-Maximization algorithm
F1	F1-score, a measure of a test's accuracy
FN	False Negatives
FP	False Positives
FPR	False Positive Rate
GD	Gradient Descent algorithm
GMM	Gaussian Mixture Model
ICA	Independent Component Analysis
ImageNet	Large visual database
Inception	Architecture of deep CNN proposed in 2014[1]
MLP	Multilayer Perceptron architecture
Nadam	Nesterov-accelerated Adaptive Moment Estimation
NasNet	Neural Architecture Search Network proposed in 2017[2]
PCA	Principal Component Analysis
ResNet	Deep Residual network architecture proposed in 2015[3]

RMSprop	Adaptive learning rate SGD
ROC AUC	Area Under Receiver Operatic Characteristic Curve
SGD	Stochastic Gradient Descent algorithm
TP	True Positives
TPR	True Positive Rate
VGG16	Very deep CNN architecture of 16 layers proposed in 2014 [4]
VGG19	Very deep CNN architecture of 19 layers proposed in 2014 [4]
VLA	Very Large Array of telescopes
Xception	Architecture of deep CNN proposed in 2016[5]

LIST OF SYMBOLS

β	Hyper-parameter of adaptive SGD algorithms
η	Learning rate of GD and SGD
$\mathcal{F}$	Mapping function between X and Y
g	Gradient of the cost function
m	Exponentially decaying average of past gradients of cost function
N	Size of the training set
∇	Gradient, a multi-variable generalization of the derivative
Q	Cost function
θ	Parameters of a model
v	Exponentially decaying average of squares of past gradients of the cost function
X, x	Input features vector (matrix) or set of input samples
Y, y	Label or set of labels

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Gravitational lensing is a unique phenomenon taking place due to the presence of heavy objects in the universe. It offers unique insights into a number of cosmological and astrophysical questions. For example, we can probe the nature of dark matter via measurements of the substructures of the known gravitational lenses. Despite its importance, only on the order of 100 lenses have been found so far, including only 10-20 of the most valuable kinds like quadruply-imaged systems, or highly variable sources. Since lenses are rare and difficult to find, this justifies the small sample size (considered in this study). The conventional discovery method is based on a manual analysis and comparison of the spectra of the objects. This method, however, takes an excessive amount of time and resources and its results are similar to random guessing in their overall outcome.

Novel data-driven methods are a promising alternative to the manual spectral analysis. With the recent development of accurate simulations, researchers have obtained access to large sets of artificial data. The synergy of the computational power and the development of statistical and machine learning enabled fast and robust discovery of gravitational lensing. The evolution of data-driven methods started from the population mixture models and statistical learning and reached the state of the art algorithms, based on deep convolutional neural networks. Latest publications report models composed of early deep learning models, such as AlexNet, VGG16, and Inception. The main drawbacks include large number of parameters causing overfitting, poor fitting capabilities, high computational cost, and long training time.

In this work, we explore deep convolutional neural networks and apply them to the problem of discovery of gravitational lensing. We exploit the advantage of the transfer learning and novel deep architectures to outperform previously published models, tackle the challenge of overfitting, and achieve the state-of-the-art performance.

Following the introduction, in Chapter 1, we describe the physical phenomenon of gravitational lensing. In Chapter 2, we stress its importance and motivate the need for fast and reliable discovery methods. In Chapter 3, we analyze the evolution of data-driven methods such as mixture models, principal component analysis (PCA), gradient-boosted trees, artificial neural networks (ANN), and the state-of-the-art methods based on deep learning. In Chapter 4, we introduce deep learning techniques, such as convolutional neural networks (CNN), stochastic gradient descent (SGD), dropout, transfer learning, as well as the latest architectures including Xception, NasNet, ResNet, Inception and VGG. In Chapter 5, we describe our manually assembled dataset and proceed with training, cross-validation, and testing. Chapter 6 concludes the thesis by summarizing the findings and discusses the future research directions.

CHAPTER 2

GRAVITATIONAL LENSING IN UNIVERSE

Gravitational lensing is a phenomenon of bending of the light caused by a heavy object, such as a cluster of galaxies, between a distant light source and an observer as shown in Figure 2-1 (A). The distribution of matter (i.e. object) is called a lens and the amount of bending can be found from one of the predictions of Albert Einstein's general theory of relativity. Unlike an optical lens, a gravitational lens produces the maximum deflection of light that passes closest to its center, and the minimum deflection of light that travels furthest from its center. Consequently, a gravitational lens has no single focal point, but a focal line.

2.1. Gravitational lenses

If the light source, the massive lensing object, and the observer lie in a straight line, the original light source will appear as a ring around the massive lensing object. If there is any misalignment, the observer will see an arc segment instead. More commonly, if the lensing mass is complex (e.g. a galaxy group or cluster) and does not cause a spherical distortion of the space-time, the source will resemble partial arcs scattered around the lens. The observer may then see multiple distorted images of the same source.

There are three classes of gravitational lensing:

1) Strong lensing that appear in the form of Einstein rings (Figure 2-1 (B)), arcs, and multiple images.

2) Weak lensing that causes small distortions, which require a lot more observations to draw conclusions

3) Microlensing that brings no observable distortions to light.

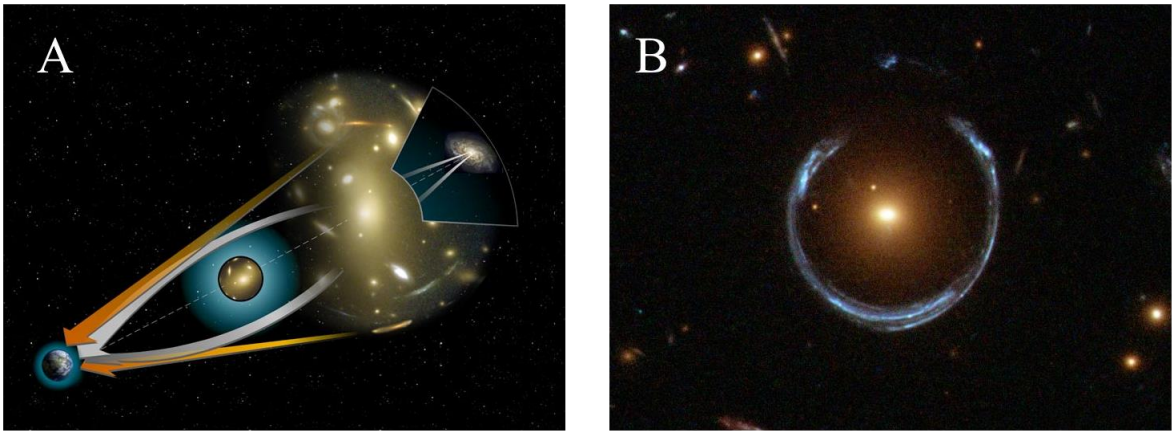

Figure 2-1. Gravitational lensing in the universe: (A) working principle and (B) the Einstein's ring lensing

2.2. Discovery methods

Gravitational lenses in space help astronomers to tackle the most important questions in cosmology, including the discovery of dark matter, imaging the deep space, and exploring gravitational interactions between extremely heavy objects in the universe[6]. Gravitational lensing is directly sensitive to the amount and distribution of dark matter. This means that, to measure the amount of lensing on a patch of sky, we do not need to know anything about what type of galaxies we are observing, how they form and behave or what color light they emit. This makes gravitational lensing a very clean and reliable cosmological probe since it relies on few assumptions or approximations. Therefore, gravitational lensing helps astronomers build accurate models of the dark matter distribution, such as the map shown in Figure 2-2[7].

Most of the gravitational lenses in the past have been discovered accidentally. A search for gravitational lenses in the northern hemisphere, performed in the range of radio frequencies using the Very Large Array (VLA) in New Mexico, led to the discovery of 22 new lensing systems, a major milestone. This has opened a completely new avenue for research, ranging from finding

very distant objects to finding values for cosmological parameters, which can help better understand the universe.

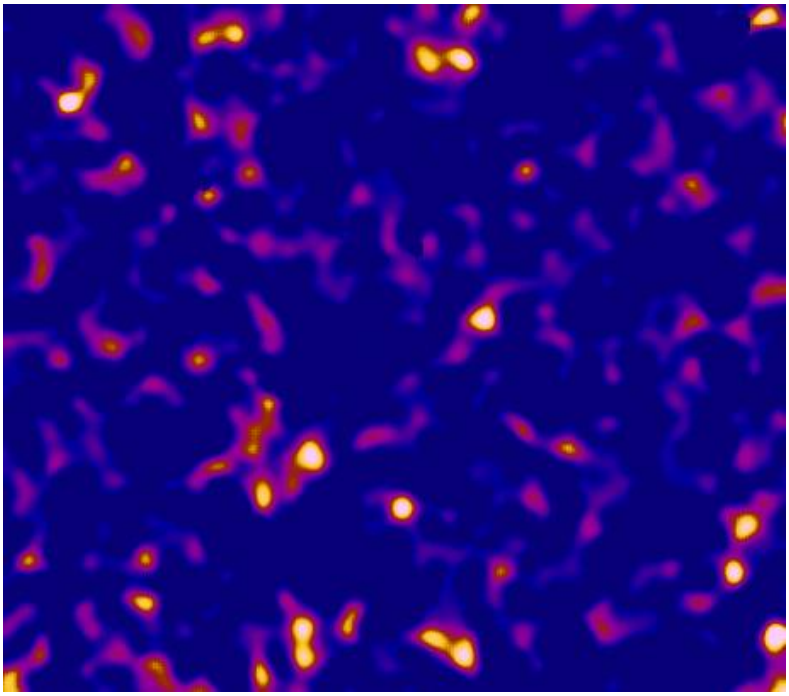

Figure 2-2. Dark matter map by CFHTLenS Collaboration[7]

Conventional techniques for discovering gravitational lenses are based on the analysis of spectral characteristics of the objects. Gravitational lenses have an equal effect on all kinds of electromagnetic radiation, not just visible light. If two neighboring objects have similar spectrums, it is likely that they are coming from the same source, which is a good indication of lensing effect (see Figure 2-3). Their relative locations can be further used to infer the accurate model of the lens. For these purposes, researchers use radiofrequency and infrared telescopes.

Conventional methods require a meticulous analysis of thousands of terabytes of images and expensive spectroscopy measurements. It is of extreme importance to develop computational tools that enable well-informed predictions of potential candidate objects. In the next chapter, we will

cover statistical modeling approaches and novel machine learning and deep learning methods demonstrating promising results.

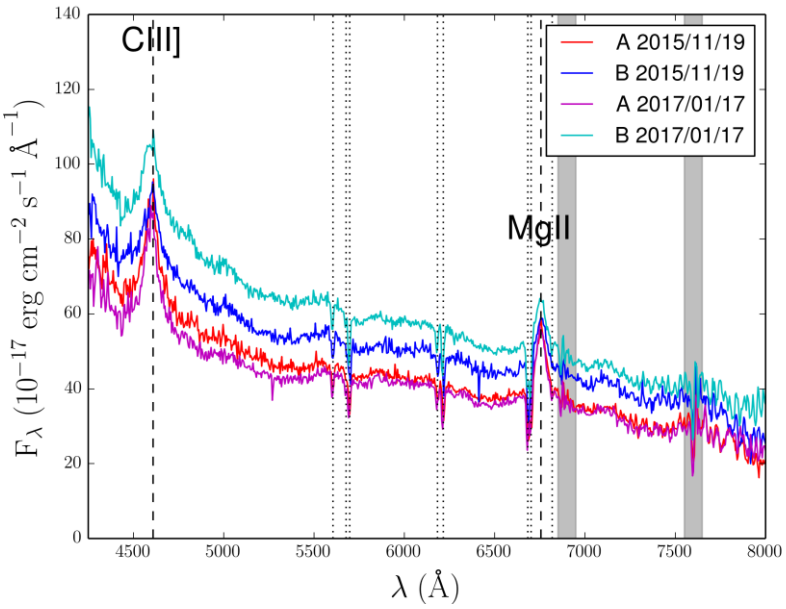

Figure 2-3. SPRAT spectroscopy of QSO B0957+561 on 19 November 2015 and 17 January 2017[8]

CHAPTER 3

DATA-DRIVEN METHODS

Despite significant importance for cosmology and physics, little progress is achieved in discovery gravitational lenses. Among the biggest challenges are the rare nature of the phenomenon, the lack of high-resolution and low noise astronomical images, and the absence of numerical methods for fast and reliable detection in large-scale astronomical surveys. Data-driven methods is a promising alternative to manual spectral analysis. With recent development of accurate models of lensing effects[9,10], researchers have access to large sets of simulated data. The synergy of computational power and development of statistical and machine learning enables fast and robust discovering of gravitational lensing.

3.1. Mixture models

Because of the diverse nature of sources and lenses in the universe, distribution of light intensity along the spectrum can be modelled as a mixture of models. *Williams et al.* applied Gaussian mixture model (GMM) to separate point-like quasars, quasars with an extended host, and strongly lensed quasars[11]. To optimize the model, authors use the expectation-maximization (EM) algorithm. At E-step, they compute log-likelihood function (1) and membership probabilities α_k for each class k with current parameters θ_t. On M-step, they find θ_{t+1} that maximizes the log-likelihood function. EM algorithm stops when the model converges to the optimal θ_{OPT}. To rule out a local minimum, the EM algorithm is executed several times with randomly chosen initial parameters for the models.

$$l(\theta) = \log p(X|\theta) = \log \prod_i^N \sum_k^K \frac{\alpha_k}{(2\pi)^{P/2}|\Sigma_k|^{1/2}} e^{-\frac{1}{2}(x_i-\mu_k)^T \Sigma_k^{-1}(x_i-\mu_k)} \quad (1)$$

GMM has several practical advantages including high speed, applicability to multiple different data sources (many astronomical surveys use different equipment), small dimensionality of the data (uses only 9-features vector), and naturalistic representation of normally distributed light sources in space. However, its simplicity is achieved by sacrificing fitting power. In particular, it only uses aggregated statistics, while ignoring pixel-wise analysis.

3.2. Machine learning approaches

Similarly to GMM, researchers explored other types of machine learning, including kernel principal component analysis (PCA), gradient boosted trees, and artificial neural networks (ANN). In comparison to GMM, these methods are able to fit non-linear data and work "out of the box" (one needs to find an optimal number of components in mixture models).

A good example is presented in *Agnello et al.*[12] The authors broke down the problem into two stages: target selection and candidate selection. In the target selection stage, promising systems were selected based solely on information available at astronomical surveys. In the candidate selection stage, they returned to the images of the targets in order to narrow down the search: they used *10 arcsecond* (*25-by-25 pixels*) cutout images, and reduced the dimensionality via kernel-PCA to *200 features*. Reduced data was used to train artificial ANN and gradient boosted trees.

3.3. Deep learning approaches

State of the art in data-driven methods of discovery of gravitational lensing are based on deep convolutional networks[13–17]. Authors combine advances in deep learning and gravitational lensing simulations to collect large datasets and train sophisticated models. Because deep convolutional networks perform pixel-wise feature extraction, these models are inherently applicable to a wide range of data sources and do not require additional dimensionality reduction steps. Another

advantage is its speed and scalability, which is extremely important as the number of astronomical images increases every year.

One of the latest works on deep-learning-based lens detection is CMU DeepLens[17] (see Figure 3-1), a model for detection of Einstein rings (particular types of strong lens show in Figure 2-1 (B)). Authors built deep neural network based on ResNet units, previously introduced by Microsoft[3]. They trained and validated the models on a set of 20,000 simulated observations, including a range of lensed systems of various sizes and signal-to-noise ratios. Although the reported performance of the proposed model is promising, authors do not include realistic images that usually have lower signal-to-noise ratio, contain contaminations and additional objects, and have larger diversity in nature.

Research group from Stanford reported the use of a deep neural network composed of eight convolutional layers and two fully connected layers to estimate lensing parameters in an extremely fast and automated way, circumventing the difficulties that are faced by maximum likelihood methods. They also showed that the removal of lens' light could be made fast and automated using independent component analysis (ICA) of multi-filter imaging data. Proposed convolutional neural networks can recover the parameters of the "singular isothermal ellipsoid" density profile, which is commonly used to model strong lensing systems, with an accuracy comparable to the uncertainties of sophisticated models but about ten million times faster: 100 systems in approximately one second on a single graphics processing unit[13]. Despite impressive speed up and performance, authors could train more sophisticated models by employing pre-trained feature maps. Although they justify use of random Xavier initialization by different nature of the data comparing to ImageNet, pre-trained weights could be helpful at intermediate- and high-level

feature extraction. Another possible improvement is use of other pre-trained models, besides AlexNet, Inception, and OverFeat.

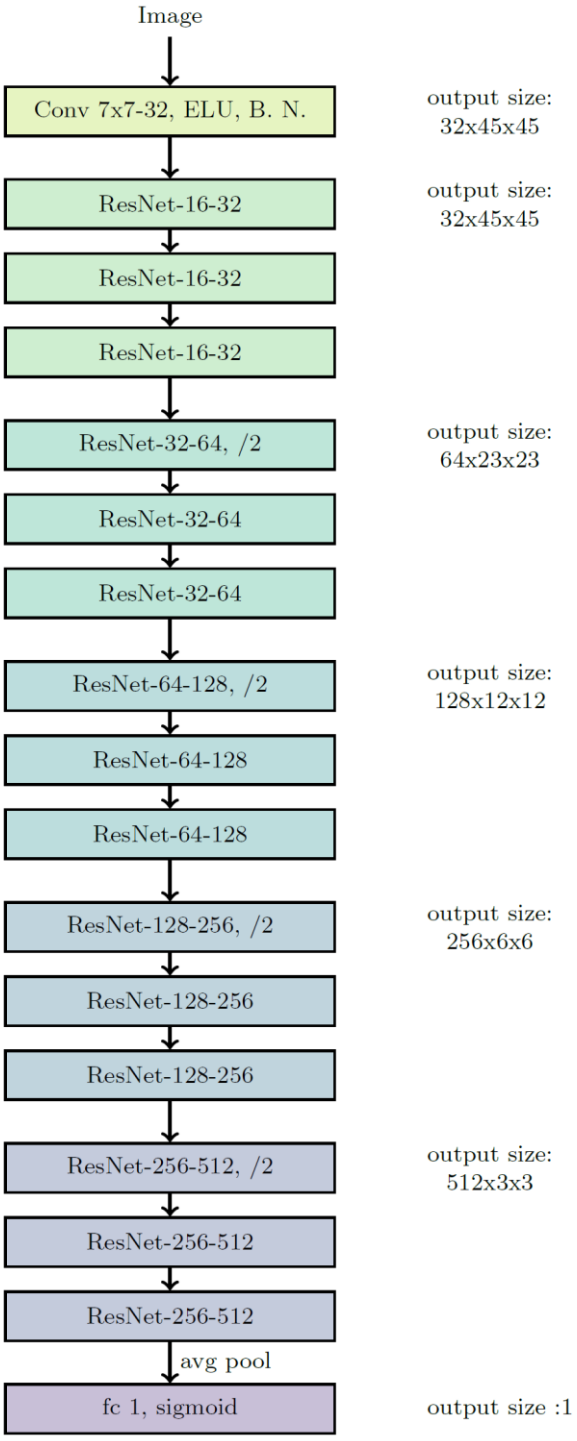

Figure 3-1. Architecture of CMU DeepLens

The recently published LensFlow model is based on an early convolutional neural network architecture (see Table 3-1)[15]. Authors combine simulated images and images of known gravitational lenses to train the model and produce *92* real candidates. The developed model is more computationally efficient and complimentary to the classical lens identification algorithms, and it is ideal for discovering such events across wide areas from current and future. The combination of simulated and real images prevents overfitting and makes the model applicable to real astronomical survey data. However, the use of more complex nets such as Inception, ResNet and others might improve the results.

Table 3-1. Tabulated Architecture of the LensFlow network

Layer	**Type**	**Data directionality**
	input	1-by-100-by-100
1	convolution + tanh	30-by-96-by-96
2	max-pooling	30-by-48-by-48
3	convolution + tanh	60-by-44-by-44
4	max-pooling	60-by-22-by-22
5	convolution + tanh	90-by-18-by-18
6	max-pooling	90-by-9-by-9 (7290 features)
7	fully connected + ReLU	1000
8	fully connected + ReLU	800
9	fully connected + ReLU	600
10	softmax	2

CHAPTER 4

DEEP CONVOLUTIONAL NEURAL NETWORKS

Introduced by *LeCun et al.*[18] in 1998, convolutional neural networks (CNN) gained global popularity after triumph of *AlexNet*[19]. Inspired by the visual cortex of the brain, CNN can achieve super-human performance in computer vision task such as recognition, reconstruction, restoration, and motion analysis[20]. It is therefore intuitive to employ CNN to discover new gravitational lenses.

4.1. Deep convolutional neural networks

The core component of CNN is a convolutional layer (see Figure 4-1), which applies a convolution operation to the input, passing the result to the next layer. It emulates the response of an individual neuron to visual stimuli of a particular pattern or color. The advantage of the convolutional layer is sparsity, as it has a small window size and requires much less memory to store the weights. For instance, if we use fully connected dense layer, then an image of *100-by-100* pixels would lead to *10000* weights for each neuron in the second layer, while few *5-by-5* convolutional filters would require *25* weights each.

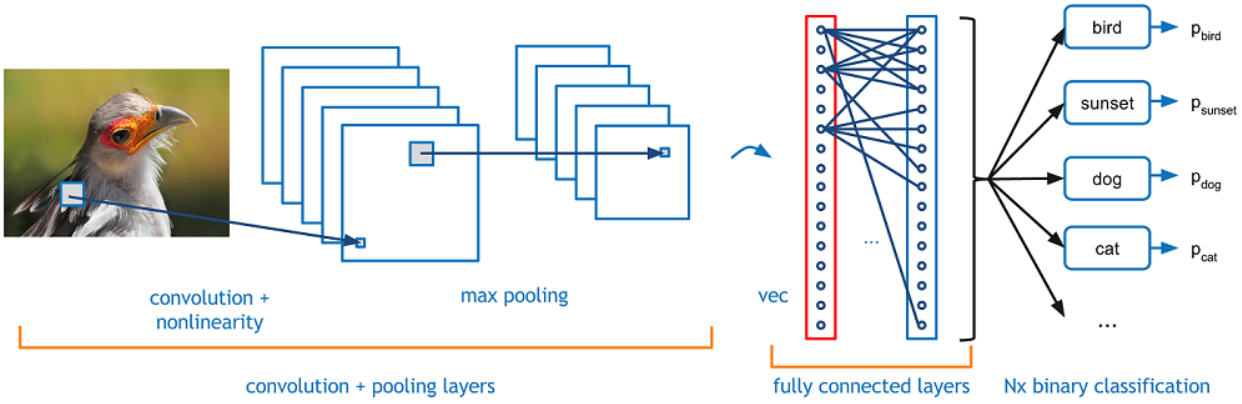

Figure 4-1. Convolutional neural network

Two other components of CNN are the pooling layer and the fully connected layer (see Figure 4-1). Pooling combines the outputs of neuron clusters at one layer into a single neuron in the next layer. It is usually represented by a simple fixed operation such as $\text{mean}\{x\}$ or $\max\{x\}$. Fully connected layers connect every neuron in one layer to every neuron in another layer. It is, in principle, the same as the traditional multi-layer perceptron neural network (MLP).

Over the last 5 years, CNNs have achieved impressive performance in numerous applications. They evolved into sophisticated architectures, also called "deep learning", which allow the models to learn representations of data with multiple levels of abstraction. Deep learning discovers an intricate structure in large data sets by using the backpropagation algorithm to indicate how a machine should change its internal parameters, which are used to compute the representation in each layer from the representation in the previous layer[20].

4.2. Gradient descent optimization

The engine of backpropagation algorithm is gradient decent (GD) optimization. It is a popular numerical method for minimization of the cost function $Q(\theta)$ by updating parameters θ in the direction opposite to that of the gradient of the cost function $\overrightarrow{\nabla_\theta Q(\theta_t)}$ as shown in equation (2).

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta Q(\theta_t) \tag{2}$$

Depending on how many samples of data we use at each update step, GD is divided into batch GD, mini-batch GD, and stochastic GD (SGD). The latter attracted far the most attention due to its faster speed and inherent capability to avoid local minima of the cost function[21]. At each step SGD stochastically chooses a single entry from new data and updates the parameters.

One of the major challenges of SGD is choosing the optimal learning rate η. Depending on the data and model, cost function can become extremely nonlinear with numerous local minimums.

To solve this problem, several adaptive variants of SGD were introduced. The most popular are RMSProp, Adam, and Nadam[22,23]. RMSProp divides the learning rate by an exponentially decaying average of squared gradients:

$$\begin{cases} g_t = \nabla_\theta Q(\theta_t) \\ E(g^2)_t = 0.9E(g^2)_{t-1} + 0.1g_t^2 \\ \theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{E(g^2)_t + 10^{-8}}} g_t \end{cases} \tag{3}$$

In addition to storing an exponentially decaying average of the past squared gradients v_t, Adaptive Moment Estimation (Adam) also keeps an exponentially decaying average of past gradients m_t, similar to momentum as shown in (4). Whereas momentum can be seen as a ball running down a slope, Adam behaves like a heavy ball with friction, which therefore prefers flat minima in the cost function space.

$$\begin{cases} m_t = \frac{\beta_1^t m_{t-1} + (1-\beta_1^t) g_t}{1-\beta_1^t} \\ v_t = \frac{\beta_2^t v_{t-1} + (1-\beta_2^t) g_t^2}{1-\beta_2^t} \\ \theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{v_t + 10^{-8}}} m_t \end{cases} \tag{4}$$

A combination of Adam and Nesterov momentum gave birth to Nesterov-accelerated Adaptive Moment Estimation (Nadam):

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{v_t + 10^{-8}}} \left(\beta_1^t \frac{\beta_1^t m_{t-1} + (1-\beta_1^t) g_t}{1-\beta_1^t} + \frac{(1-\beta_1^t) g_t}{1-\beta_1^t} \right) \tag{5}$$

4.3. Regularization techniques

Another common challenge for machine learning is overfitting. In a practical scenario, it is likely that without prior knowledge of the data one can build a model that contains more parameters than can be justified by the data. One way to overcome this issue is regularization, a technique

used in an attempt to solve the overfitting problem in statistical models. Most popular and effective regularization methods in deep learning are the dropout and batch normalization.

The key idea behind dropout is to randomly drop units (along with their connections) from the neural network during training. This prevents units from co-adapting too much. During training, dropout samples from an exponential number of different "thinned" networks. At test time, it is easy to approximate the effect of averaging the predictions of all these thinned networks by simply using a single "thick" network that has smaller weights. This significantly reduces overfitting and gives major improvements over other regularization methods. It has been shown that dropout improves the performance of neural networks on supervised learning tasks in vision, speech recognition, document classification and computational biology, obtaining state-of-the-art results on many benchmark data sets[24].

Batch normalization was initially proposed as a method for accelerated training since it leads to faster and better performance[25]. However, it is commonly used together with dropout to achieve more flexible regularization effect without sacrificing model complexity. The core idea is that in deep networks the distribution of each layer's inputs changes during training, as the parameters of the previous layers change. This slows down the training by requiring lower learning rates and a careful parameter initialization, and makes it notoriously hard to train models with saturating nonlinearities. Authors call this phenomenon internal covariate shift and address the problem by normalizing layer inputs. Batch normalization is applied for each training mini-batch and becomes an integral part of the model. It allows use of much higher learning rates and simplifies hyper parameter optimization.

4.4. State of the art architectures

In 2014 *Simonyan and Zisserman* of the University of Oxford created *19-layers* and *16-layers* CNN that strictly used *3-by-3* filters with stride and pad of *1*, along with *2-by-2* max-pooling layers with stride *2*. These architectures became widely known as VGG16 and VGG19[4]. Authors also replaced large convolutions with two consecutive small convolutional layers in order to achieve higher flexibility and a decrease in the number of parameters. VGG was the first CNN architecture that reinforces the idea of shrinking spatial dimensions, but growing depth.

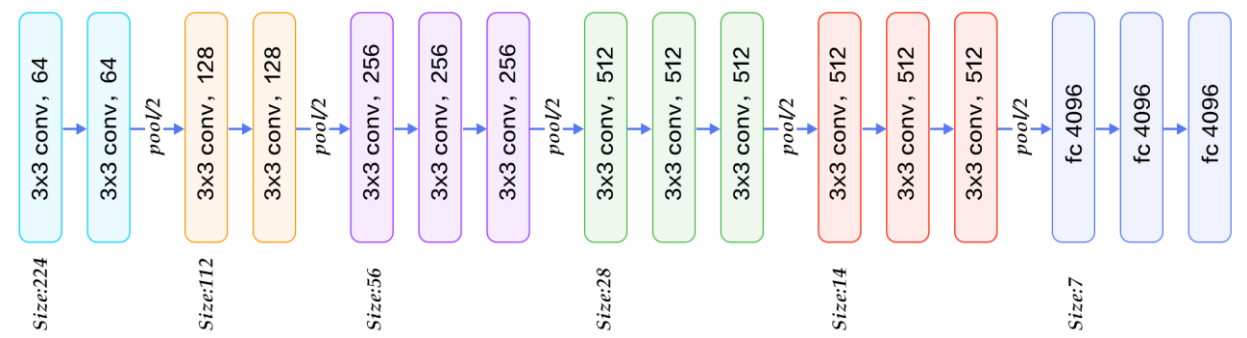

Figure 4-2. Architecture of VGG16

The winner of ILSVRC in 2014 was Google's Inception net[1]. It is *22-layers* CNN that is built out of "inception" building block (see Figure 4-3 (A)). It contains input (bottom green box), intermediate parallel convolutions (blue boxes), pooling (red box) and dimensionality reduction (yellow boxes) operations, and concatenated output (top green box). The idea behind inception module is to perform multiple operations in parallel and increase fitting power of the model. Besides inception module, the authors showed that a creative structuring of layers could lead to improved performance. Moreover, by avoiding use of fully connected layers, they make Inception net computationally efficient. In the next few years, the field of deep learning witnessed truly amazing architectures based on these ideas.

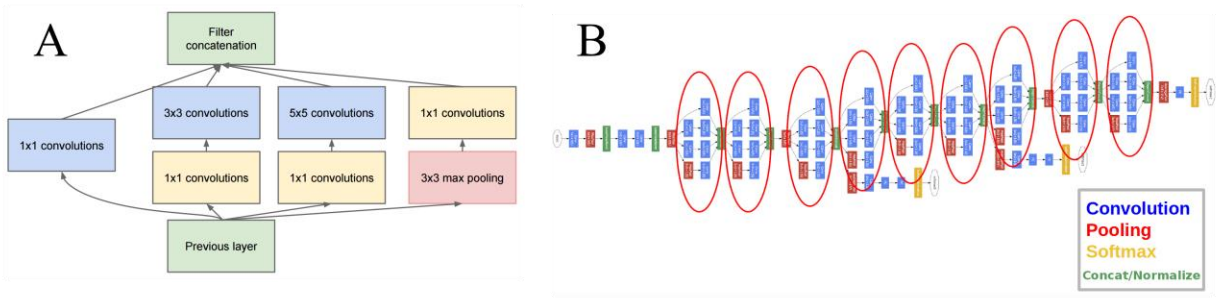

Figure 4-3. Inception net: (A) its building block and (B) 22-layers architecture

In 2015, Microsoft set a new world record in classification, detection, and localization on ILSVRC with even deeper ResNet architecture[3]. The model has depth of *152* layers and is based on a residual block, which tries to optimize the residual mapping of the data. As shown in Figure 4-4, we pass the data (x) through two layers of convolution and then add it to itself. The resulting function $\mathcal{F}(x)$ will learn the required residual. Stacking these modules on top of each other achieves super-human performance in traditional computer vision tasks. The main argument is that it is easier to optimize the residual mapping than to optimize the original, unreferenced mapping. Moreover, such scheme inherently overcomes the vanishing gradient problem, as we evenly distribute the gradient through regular addition operations.

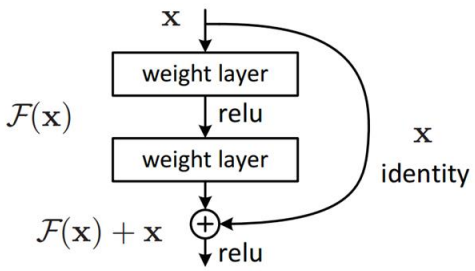

Figure 4-4. Residual block of ResNet

Extreme version of Inception net was proposed in 2017 by researchers from Google[5]. Authors hypothesize that the mapping of cross-channels correlations and spatial correlations in the feature maps of convolutional neural networks can be entirely decoupled. Xception architecture has 36

convolutional layers forming the feature extraction base of the network. Convolutional layers are structured into 14 modules, all of which have linear residual connections around them, except for the first and last modules. In other words, the Xception architecture is a linear stack of depth-wise separable convolution layers with residual connections.

Recently developed NasNet learns the model architectures directly on the dataset of interest. Since this approach is expensive when the dataset is large, authors propose to search for an architectural building block on a small dataset and then transfer the block to a larger dataset. NasNet creates a new search space (NasNet search space) which enables transferability[2]. Similarly to Inception module, NasNet builds on parallelizing multiple operations and varies their number. For example, one may build a deep network from "normal" and "reduction" cells (see Figure 4-5) by stacking them on each other.

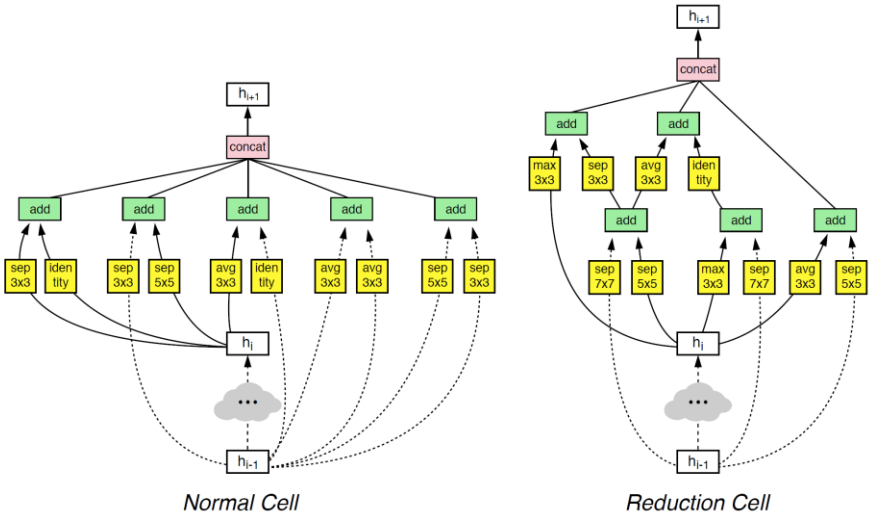

Figure 4-5. Example of NasNet's building blocks: normal cell (left) and reduction cell (right)

4.5. Transfer learning

A major assumption in many machine learning and data mining algorithms is that the training and future data must be in the same feature space and have the same distribution. However, in

many real-world applications, this assumption may not hold, which makes a previously trained model inapplicable in many tasks. Transfer learning is a research problem in machine learning that focuses on storing knowledge gained while solving one problem and applying it to a different but related problem.

Beauty of transfer learning is that it allows us to use previously trained models and their feature maps. For example, we sometimes have a classification task in one domain of interest, but we only have sufficient training data in another domain of interest, where the latter data may be in a different feature space or follow a different data distribution. In such cases, knowledge transfer, if done successfully, would greatly improve the performance of learning by avoiding much expensive data-labeling efforts. In recent years, transfer learning has emerged as a new learning framework to address this problem. *Pan and Yang* published a thorough review of modern transfer learning approaches[26].

Transfer learning is a powerful technique that lets us to employ pre-trained models, including state-of-the-art ones such as Inception, ResNet, VGG16, VGG19, NasNet, and Xception.

CHAPTER 5

TRAINING AND VALIDATION

In the previous chapter, we introduced deep convolutional networks that are used in computer vision tasks such as recognition, reconstruction, restoration and motion analysis. Each one of these models achieved the state-of-the-art performance. Given demonstrated success of deep learning in discovery of gravitational lensing, we hypothesize that such models can be of immediate use in this field.

We train models in transfer learning paradigm by taking advantage of pre-trained feature mappings. In particular, we add a few deconvolutional layers to upsample the input image to appropriate dimensions. Then we replace the last few layers in the pre-trained model with randomly initialized equivalent ones that produce the output of length two. Finally, we apply the softmax operation to produce probability-akin output. To avoid overfitting, we introduce dropout and batch normalization layers.

In this work, we use the following pre-trained models:

1) Inception[1]
2) ResNet[3]
3) NasNet[2]
4) Xception[5]
5) VGG16[4]
6) VGG19[4]

5.1. Training dataset

To build our discovery algorithm for gravitational lensing we collect data from Cfa-Arizona Space Telescope Lens Survey (CASTLES)[27] of gravitational lenses. CASTLES is an open database of known lenses built on hundreds of published papers. It provides a list of objects with observed statistics and images at different wavelengths. To be consistent with LensFlow[15] data and to be able to evaluate the latest proposed candidates, we use only I-band images cleaned from noise and artifacts. In addition to this, we collect simulated true positives from the LensFlow paper.

Assuming that gravitational lensing is an extremely rare event, we randomly sample I-band images from Hubble Legacy Archive[28] measured with the same technical specifications as in CASTLES data. These random images of galaxies and stars are used as true negatives. In addition, we collect simulated false positives from the LensFlow paper.

Finally, we employ data augmentation to combat the imbalanced classes problem. We apply the following random transformations to augment the data and assemble a dataset of a total of 15,000 positive and 13,239 negative samples:

7) Rotation around center $[-45°, 45°]$
8) Zooming $[0.9,1.0]$
9) Shear mapping $[-15°, 15°]$
10) Horizontal flipping
11) Vertical flipping

The overall collected dataset is shown in Figure 5-1 and a random sample is illustrated in Figure 5-2. Each image is a single-channel *100-by-100*-pixel array of single precision floating numbers (32 bits).

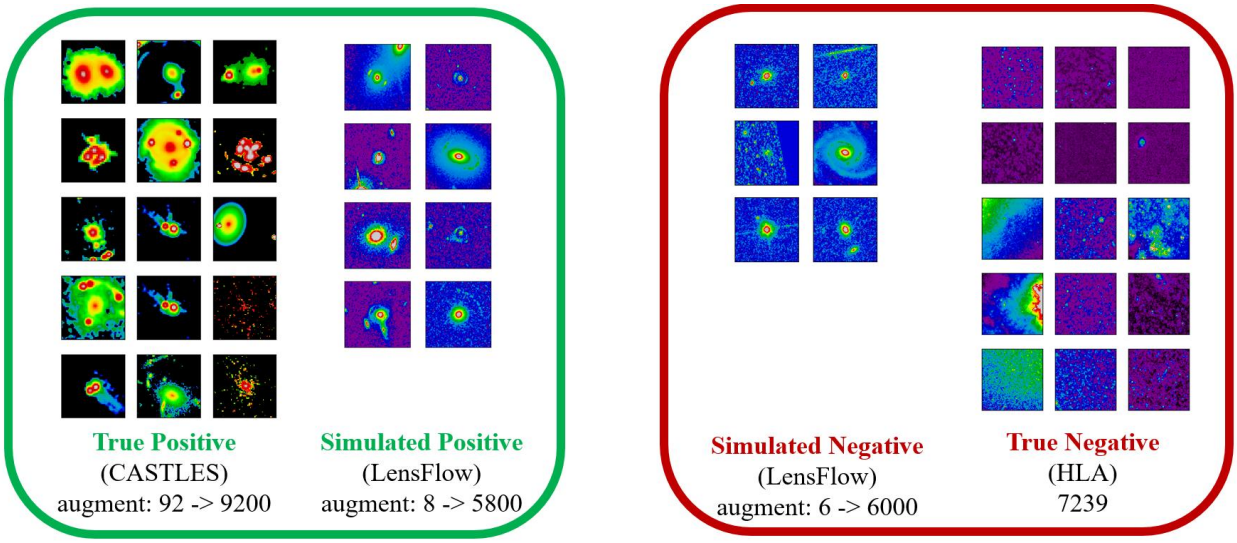

Figure 5-1. Training dataset: gravitational lenses are in a green box (left) and non-lenses are in a red box (right)

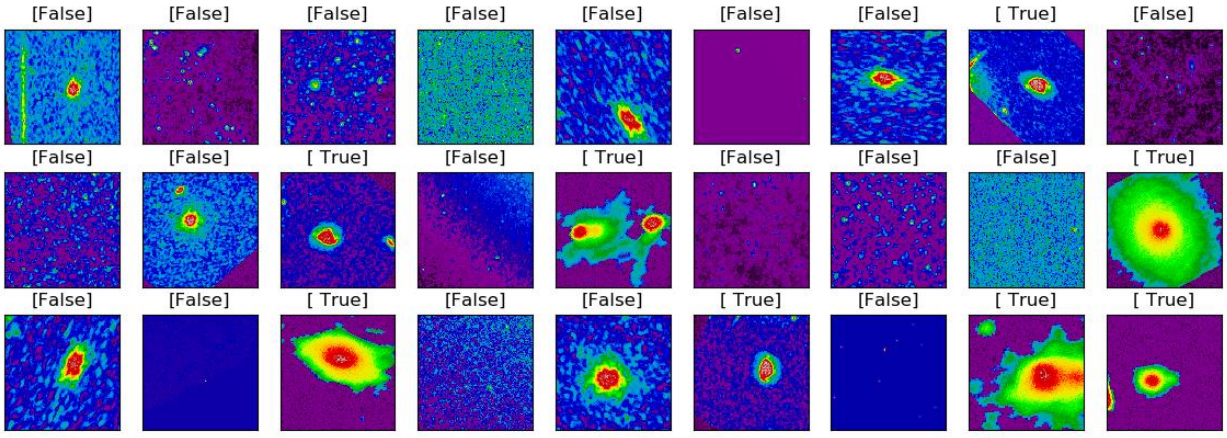

Figure 5-2. Random sample from augmented training dataset with labels: [True] and [False] correspond to lenses and non-lenses respectively

5.2. Cross validation training

In the model selection step, we use a 7-fold cross-validation (CV) procedure, which splits the entire dataset into seven subsets. At each iteration, the models are randomly initialized and then trained on six shuffled subsets. After an iteration, each model is evaluated on the last remaining subset. After seven iterations, we calculate the mean of the performance metrics and report them.

We use the area under the receiver operating characteristic curve (ROC AUC) and F1-score given by (6) and (7) respectively.

$$ROC\ AUC = \int_{-\infty}^{+\infty} TPR(T)FPR'(T)dT = 1 - \frac{1}{2}\sum_{\mathcal{F}(X_k)>\mathcal{F}(X_{k-1})}^{N}(X_k - X_{k-1})(Y_k - Y_{k-1}) \quad (6)$$

$$F1 = \frac{2TP^2}{2TP+FP+FN} \quad (7)$$

Since we adjust the original models by adding deconvolutional layers in the beginning and replacing fully connected layers in the end, we have a choice of whether to keep the intermediate layers intact. In addition, we use random Xavier initializations, that was suggested in literature[13], for all layers to test if it yields better performance. The highest average CV results (out-of-sample) is obtained when the intermediate layers are kept unchanged during the training (see Table 5-1). On average, training the full network leads to significant overfitting and degradation of ROC AUC below 0.7 (70%). Random Xavier initialization also causes overfitting, as the models cannot generalize on the validation set and yield ROC AUC similar to random guesses.

Table 5-1. Out-of-sample cross-validation performance

Model name	**# of parameters**	**ROC AUC**	**F1-score**
Xception	23M	99.84%	98.95%
Inception	24M	92.69%	80.00%
ResNet	26M	78.49%	54.33%
NasNet	93M	74.40%	15.49%
VGG16	139M	71.72%	56.91%
VGG19	144M	61.54%	38.23%

As one can see from the table above, the best performance was achieved by the models with smallest number of fitting parameters such as Xception and Inception. This is a strong indication of overfitting in ResNet, NasNet, VGG16 and VGG19 models. Although we used dropout and batch normalization, the corresponding ROC AUC and F1-score significantly drop when we cross the 25 millions of parameters. Nevertheless, Xception and Inception models achieved the state-of-the-art ROC AUC and F1-scores. They by far outperform the conventional methods for discovery of gravitational lenses as well as early data-driven methods.

5.3. Predicting new candidates

To test our best-fitting models against previously published works, we produce predictions on the latest 92 candidates published in the LensFlow paper[15] (see Table 5-2). The candidates passed meticulous manual analysis by experts and were divided into 3 classes: A (most likely a gravitational lens), B (there is chance that it is not a lens) and C (most likely not a lens).

Table 5-2. Predictions on 92 candidates reported in LensFlow paper[15]

Predicted Probability		**Xception**		**Inception**		**ResNet**	
		mean	**std dev**	**mean**	**std dev**	**mean**	**std dev**
LensFlow Grade	A*	0.982	0.041	0.872	0.314	0.767	0.385
	B**	0.618	0.297	0.852	0.290	0.775	0.356
	C***	0.553	0.336	0.639	0.354	0.657	0.447

* Grade A corresponds to images that are clearly a strong gravitational lens.

** Grade B lenses correspond to images that are most likely a lens, but there is a chance they could also be artifacts, noise, structures in elliptical galaxies, satellite galaxies, tidally interacting galaxies, etc.

*** Grade C lenses consist of images that are most likely not a lens, but there is a chance they might be gravitationally lensed.

As the results suggest, Xception network has the best conformity with the suggested grades. It predicts a mean of 98% for “grade A”, 62% for “grade B” and 55% for “grade C”, which is an extremely good result given that the model was trained on a completely different dataset and did not require manual analysis from professional astronomers. Interestingly, that for grades B and C Xception has much larger standard deviation, which confirm that these candidates, as their description states, might not be lenses.

CHAPTER 6

CONCLUDING REMARKS

Gravitational lensing is an exceptionally rare event predicted by Einstein's general theory of relativity. Caused by the heaviest objects in the universe, it is inherently important for cosmology and astrophysics. Nonetheless, we have only discovered a little portion of these events predicted by our current understanding of the dark matter. The main bottleneck is the conventional discovery method, which is based on manual analysis and comparison of the spectrum of the objects. It takes infeasible amounts of time and resources. Synergy of theoretical models of gravitational lensing, statistical analysis, and numerical techniques led to novel data-driven approaches and enabled fast and automated predictions of the new gravitational lenses.

Data-driven discovery methods evolved from the population mixture models and statistical learning into the state of the art algorithms based on deep convolutional neural networks. The latest publications report models composed of early deep learning models such as AlexNet, VGG16, and Inception. The main drawback of these models is the large number of parameters that causes overfitting, poor fitting capabilities, high computational cost, and long training time.

In this work, we demonstrated that deep convolutional neural networks, commonly applied to the most challenging problems in computer vision, have a prominent potential in this field. The proposed models achieved the state-of-the-art performance on single channel (I-band) *100-by-100-*pixel images.

6.1. Predicting new candidates

Fine-tooth comparison of the 92 latest proposed candidates revealed a strong conformity with a mean prediction of 98% for “grade A” candidates, 62% for “grade B” and 55% for “grade C”.

The key concluding points:

1) Deep convolutional neural networks seem to catch important features of gravitational lenses and may be extremely helpful in the discovery of new objects in the next decade.
2) Transfer learning (use of pre-trained models) eliminates long training time and yields higher performance. This implies that despite a different domain, astronomical images have common patterns with general-purpose datasets such as CIFAR-10 and ImageNet.
3) Data augmentation (zooming, rotating, shear mapping, flipping) is a simple yet powerful method that works well for this problem.
4) Sophisticated models suffer from significant overfitting. This implies opportunities for higher performance with a better dataset and more efficient model architectures.

6.2. Future research directions

Among future research directions, the most promising ones include:

1) Further improvement in the training dataset almost certainly guarantees better performance. This include larger amount of diverse data, additional multimodal information (e.g. location, additional colors or bands, measurement metadata), and high-quality simulated data.
2) Generative models (e.g. generative adversarial networks) can be used either to generate new data and learn the feature space of the problem, or to train discriminative models and improve false positive rates.

3) Advanced feature extraction can lead to higher performance and faster computations. These methods include but not limited to novel deep learning (e.g. capsule networks, special transformer networks), kernel learning, and image processing (e.g. the phase stretch transform).

REFERENCES

1. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the Inception Architecture for Computer Vision. *ArXiv151200567 Cs* (2015).
2. Zoph, B., Vasudevan, V., Shlens, J. & Le, Q. V. Learning Transferable Architectures for Scalable Image Recognition. *ArXiv170707012 Cs Stat* (2017).
3. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. *ArXiv151203385 Cs* (2015).
4. Simonyan, K. & Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *ArXiv14091556 Cs* (2014).
5. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. *ArXiv161002357 Cs* (2016).
6. Quimby, R. M. *et al.* Detection of the Gravitational Lens Magnifying a Type Ia Supernova. *Science* **344,** 396–399 (2014).
2. The Canada France Hawaii Lensing Survey | CFHTLenS. Available at: http://www.cfhtlens.org/public/canada-france-hawaii-lensing-survey.
8. Gil-Merino, R., Goicoechea, L. J., Shalyapin, V. N. & Oscoz, A. New database for a sample of optically bright lensed quasars in the northern hemisphere. *ArXiv180510336 Astro-Ph* (2018).
9. Schneider, P. Gravitational lensing statistics. in *Gravitational Lenses* (eds. Kayser, R., Schramm, T. & Nieser, L.) **406,** 196–208 (Springer Berlin Heidelberg, 1992).
10. Treu, T. Strong Lensing by Galaxies. *Annu. Rev. Astron. Astrophys.* **48,** 87–125 (2010).
11. Williams, P., Agnello, A. & Treu, T. Population mixtures and searches of lensed and extended quasars across photometric surveys. *Mon. Not. R. Astron. Soc.* **466,** 3088–3102 (2017).

12. Agnello, A., Kelly, B. C., Treu, T. & Marshall, P. J. Data Mining for Gravitationally Lensed Quasars. *Mon. Not. R. Astron. Soc.* **448,** 1446–1462 (2015).

13. Hezaveh, Y. D., Levasseur, L. P. & Marshall, P. J. Fast automated analysis of strong gravitational lenses with convolutional neural networks. *Nature* **548,** 555–557 (2017).

14. Petrillo, C. E. *et al.* Finding Strong Gravitational Lenses in the Kilo Degree Survey with Convolutional Neural Networks. *Mon. Not. R. Astron. Soc.* **472,** 1129–1150 (2017).

15. Pourrahmani, M., Nayyeri, H. & Cooray, A. LensFlow: A Convolutional Neural Network in Search of Strong Gravitational Lenses. *Astrophys. J.* **856,** 68 (2018).

16. Pourrahmani, M., Nayyeri, H. & Cooray, A. LensExtractor: A Convolutional Neural Network in Search of Strong Gravitational Lenses. *Astrophys. J.* **856,** 68 (2018).

17. Lanusse, F. *et al.* CMU DeepLens: Deep Learning For Automatic Image-based Galaxy-Galaxy Strong Lens Finding. *Mon. Not. R. Astron. Soc.* **473,** 3895–3906 (2018).

18. Lecun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **86,** 2278–2324 (1998).

19. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. in *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1* 1097–1105 (Curran Associates Inc., 2012).

20. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521,** 436–444 (2015).

21. Bottou, L. Large-Scale Machine Learning with Stochastic Gradient Descent. in *Proceedings of COMPSTAT'2010* 177–186 (Physica-Verlag HD, 2010). doi:10.1007/978-3-7908-2604-3_16

22. Ruder, S. An overview of gradient descent optimization algorithms. *ArXiv160904747 Cs* (2016).

23. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. *ArXiv14126980 Cs* (2014).

24. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. 30

25. Ioffe, S. & Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *ArXiv150203167 Cs* (2015).

26. Pan, S. J. & Yang, Q. A Survey on Transfer Learning. *IEEE Trans. Knowl. Data Eng.* **22,** 1345–1359 (2010).

27. C.S. Kochanek, E.E. Falco, C. Impey, J. Lehar, B. McLeod, H.-W. Rix. Gravitational Lens Data Base. Available at: https://www.cfa.harvard.edu/castles/.

28. The Hubble Legacy Archive. Available at: https://hla.stsci.edu/.