Prepared for submission to JCAP

Analyzing γ-rays of the Galactic Center with Deep Learning

Sascha Caron[a,b] **Germán A. Gómez-Vargas**[c] **Luc Hendriks**[a] **Roberto Ruiz de Austri**[d]

[a]Institute for Mathematics, Astrophysics and Particle Physics, Faculty of Science, Mailbox 79, Radboud University Nijmegen, P.O. Box 9010, NL-6500 GL Nijmegen, The Netherlands

[b]Nikhef, Science Park, Amsterdam, The Netherlands

[c]Instituto de Astrofísica, Pontificia Universidad Católica de Chile, Avenida Vicuña Mackenna 4860, Santiago, Chile

[d]Instituto de Física Corpuscular, IFIC-UV/CSIC, Valencia, Spain

E-mail: scaron@cern.ch, ggomezv@uc.cl, luc@interaqt.nl, rruiz@ific.uv.es

Abstract.

We present a new method to interpret the γ-ray data of our inner Galaxy as measured by the *Fermi* Large Area Telescope (*Fermi* LAT). We train and test convolutional neural networks with simulated *Fermi*-LAT images based on models tuned to real data. We use this method to investigate the origin of an excess emission of GeV γ-rays seen in previous studies. Interpretations of this excess include γ rays created by the annihilation of dark matter particles and γ rays originating from a collection of unresolved point sources, such as millisecond pulsars. Our new method allows precise measurements of the contribution and properties of an unresolved population of γ-ray point sources in the interstellar diffuse emission model.

Contents

1 Introduction

Images of the γ-ray sky obtained with the *Fermi*-LAT in the direction of the inner Galaxy contain a diffuse component or diffuse emission and point sources. The diffuse emission originates from the outer Galaxy, true inner Galaxy, foreground, unresolved sources, a possible dark matter (DM) contribution, and a cosmic-ray (CR) instrumental background.

The usual method to disentangle the diffuse emission from the Galactic plane from other sources is the so called *template fitting method*. The normalization of each two- or three-dimensional template[1], based on averaged properties of the emission from a particular diffuse source, is fitted to data energy bin by energy bin. These templates can be generated using GALPROP[2] [1, 2] . This code calculates the propagation of CRs in the interstellar medium, and computes diffuse γ-ray emission due to their interaction with interstellar nucleons, magnetic fields (in the case of CR electrons) and photons in the same framework. Each GALPROP run using specific realistic astrophysical inputs (among others, distribution and properties of CR sources in the Galaxy, its gas and radiation fields content and a model for the CR propagation) will create a different interstellar emission model (IEM). Using the template fitting method an excess of γ-rays from the Inner Galaxy region has been identified: the Galactic Center Excess (GCE) [3–12]. This emission has been interpreted as a possible signal for the annihilation of DM particles or as a signal of unresolved point sources.

The data contain information that can not be captured in template models as they represent average source emission. In particular, the template fitting method can not distinguish if the GCE is completely diffuse in nature, as expected in the DM scenario, or whether it has a granular morphology, as expected if due to a collective emission of point sources too weak to be detected individually[3]. The aim of this work is to present a new method based on convolutional neural networks (ConvNets) to unveil the diffuse vs. granular nature of emissions. We apply our method on the data collected by the *Fermi* LAT in the region of the Galactic Center, we use simulations of this region to train and validate the network. This is a proof-of-principle work and the models for simulations are simplified to a level in which the results on the GCE nature provide valuable information for a more sophisticated implementation of the method in a future publication.

The paper is organized as follows: We start with an introduction to the GCE. Section 3 presents the basics on ConvNets. This technology needs large amounts of labeled images to work and we use realistic simulations for both training and testing the networks; section 4 expands on the setup to create these images. Section 5 presents the ConvNet designed to make predictions on the fraction of granularity present in the GCE, i.e. how much of the GCE is due to a population of unresolved point sources or a truly diffuse source. Results are presented in section 6, conclusions and foreseen applications are discussed in Section 7.

2 The 2017 status of the Galactic Center Excess

After being discussed by several groups, the GCE was also recently studied in depth by the *Fermi*-LAT Collaboration [14, 15]. In particular the latest work presents an updated status of the GCE using the reprocessed pass-8 event data collected in about 6.5 years of observations

[1] two spatial dimensions, but sometimes energy spectral information can be added.

[2] http://galprop.stanford.edu

[3] If the point sources are too dim or too close to each other it is very hard to distinguish them from diffuse radiation. Also, it has been recently argued that if some fraction of the DM has dissipative interactions, it can form dense DM clumps resulting in a population of γ-ray point sources in the GC region [13].

[15]. A large set of systematic sources in the extraction of the GCE properties (spectral shape, magnitude and morphology) were investigated in [15], concluding that GCE is present in the data (at least in the 1-6 GeV range), but that its exact properties are significantly model dependent. The gray band of figure 1 presents the set of spectral energy distributions of the GCE found in [15] derived from the different systematics sources investigated. In the 1-6 GeV energy range the GCE is always present. DM-like signals observed in other regions of the Galactic Plane, where such signals are not expected (control regions), give a handle on the magnitude of systematic uncertainties due to diffuse emission modeling in the Galactic Center. This prevents the unambiguous interpretation of the GCE as a signal for DM annihilation.

An alternative mechanism to produce the GCE is based on the facts that, on the one hand some pulsars (~ 210) have been identified in the γ-ray band of the whole sky[4] and on the other hand, due to the high level of past star formation activity in that region, a population of pulsars is expected in the Galactic bulge [16]. Many groups have investigated the possibility that a population of unresolved γ-ray pulsars could be the reason of the GCE emission, see for instance: [17–22]. In particular reference [22] concludes that about 60 Galactic bulge pulsars should have been seen already by the *Fermi*-LAT, however they may not have been identified as pulsars. Furthermore, using novel statistical methods the authors of [23] and [24] claim evidence for the existence of an unresolved population of γ-ray sources in the inner 20 deg of the Galaxy with a spatial distribution and collective flux compatible with the GCE.

Recently the *Fermi*-LAT Collaboration also investigated the pulsar interpretation of the GCE [25]. Performing a new point source search in 7.5 years of Pass 8 *Fermi*-LAT data in a $40^o \times 40^o$ box around the GC, they confirm the findings of [23] and [24]. In the analysis they detect 400 sources with 66 of them being γ-ray pulsar candidates, and find that they are more likely to be the brighter members of an underlying population of pulsars in the Galactic bulge than Galactic bar pulsars. They find that the collective emission of the pulsar population is compatible with the GCE properties. However, some arguments [26–28] have been raised against this interpretation of the GCE, pointing out that a putative γ-ray millisecond pulsar (MSP) population in the Galactic bulge with similar properties of extant populations, as in globular clusters and the local Galactic disc, is not able to reproduce the whole GCE emission. The weak point of those arguments is the assumption that a MSP population in the Galactic bulge needs to share similarities with populations in different environments and with diverse origin [29]. Therefore, the debate on the nature of the GCE is not yet closed, leaving the possibility of having all or a fraction of the GCE due to DM annihilation or any other diffuse source.

3 Convolutional Neural Networks

ConvNets are a class of deep neural networks (DNNs) that are designed to make predictions on visual data. Neural networks are used in many areas of research to solve classification or regression problems and are well suited when problems are high dimensional. In general, a ConvNet takes an N-dimensional input, transforms it using different layers, and produces an M-dimensional output. This can be a prediction that a specific input belongs to a particular class (e.g. object detection) or a prediction of a regression problem (e.g. this work). The input of a computer vision problem is typically an image, which can be represented by a (w, h, c)-tensor, where w and h represent the width and height of the network and c the number of channels. In this work the number of channels is 1: A single energy-integrated counts

[4] https://confluence.slac.stanford.edu/x/5Jl6Bg

map of the Galactic Center region with a specific fraction of the GCE due to a population of unresolved point sources free to vary between 0 and 1, we call this fraction the $f_{\rm src}$ parameter.

When a DNN is initialized, its weights are set at random following some distribution (typically normal). During training, the weights are modified such that the network can predict the training data without overfitting. When enough samples are given the network can learn to generalize and predict. In this work the network is trained to predict $f_{\rm src}$ from Galactic Center images created with events in the energy band 1-6 GeV. A more thorough introduction to ConvNets is given in Appendix A. $f_{\rm src}$ represents the fraction of point sources of the GCE. $f_{\rm src} = 0$ means that the GCE is composed of only a diffuse soure and $f_{\rm src} = 1$ means that the GCE is composed of only point sources.

4 Images for training and testing the ConvNets

Typically a large training dataset is needed to train a ConvNet, because a ConvNet can have up to millions (or billions) of weights that need to be tuned during the training phase (see Appendix A). Also, because the goal of the network is to predict $f_{\rm src}$ regardless of flux and location of the unresolved point sources as well as the IEM, the network should generalize over all these four unknowns. This can be done by supplying the network with many examples in this 4D parameter space. A set of *1.2 million mock images* of the GC region based on models fitted to real data is created. These images represent different realizations of the GCE fraction in form of a randomized population of unresolved point sources, $f_{\rm src}$[5]. During training of the network we notice that less training images would hurt accuracy of the network and more images would not increase accuracy, therefore a training dataset in the order of a million images is deemed to be optimal.

To model the *Fermi*-LAT observation of the Galactic Center region we include an integrated γ-ray flux at Earth, ϕ_s, expected from dark matter annihilation in a density distribution, $\rho(r)$, given by

$$\phi_s(\Delta\Omega) = \underbrace{\frac{1}{4\pi}\frac{\langle\sigma v\rangle}{2m_{DM}^2}\int_{E_{\rm min}}^{E_{\rm max}}\frac{{\rm d}N_\gamma}{{\rm d}E_\gamma}{\rm d}E_\gamma}_{\Phi_{\rm PP}} \times \underbrace{\int_{\Delta\Omega}\Big\{\int_{\rm l.o.s.}\rho^2(r){\rm d}l\Big\}{\rm d}\Omega'}_{\rm J-factor}. \tag{4.1}$$

Here, the $\Phi_{\rm PP}$ term depends on the particle physics properties of dark matter—i.e., the thermally-averaged annihilation cross section, $\langle\sigma v\rangle$, the particle mass, m_{DM}, and the differential γ-ray yield per annihilation, ${\rm d}N_\gamma/{\rm d}E_\gamma$, integrated over the experimental energy range from $E_{\rm min}$ to $E_{\rm max}$. The J $-$ factor is the line-of-sight integral through the dark matter distribution integrated over a solid angle, $\Delta\Omega$. Qualitatively, the J $-$ factor encapsulates the spatial distribution of the dark matter signal, while $\Phi_{\rm PP}$ sets its spectral character.

We use the gNFW density distribution [30, 31]:

$$\rho(r) = \rho_s\frac{r_s^3}{r^\gamma(1+r_s)^{3-\gamma}}. \tag{4.2}$$

Where r_s is the scale radius (20 kpc) and ρ_s a scale density fixed by the requirement that the local DM density at Galactocentric radius of 8.5 kpc is 0.4 GeV cm^{-3}. To model $\Phi_{\rm PP}$, inspired by results in [10], we use a $m_{DM} = 50$ GeV, and the differential γ-ray yield from WIMP annihilating into $b\bar{b}$ final state. The background for the signal is an IEM and its corresponding extragalactic emission model together with the sources listed in the 3FGL catalog [32]. Regarding the IEM we select five different GALPROP-inspired models from [33], these models were produced with all-sky likelihood fits to Pass8 data, see the parameters of the selected IEMs in table 1.

4.1 Fitting procedure and results

In this work we use the *Fermi*-LAT data collected between 2008 August 4 and 2015 August 2 (*Fermi* Mission Elapse Time 239559568 s - 460166404 s). To avoid bias due to CR contamination we select the events belonging to the Pass 8 UltraCleanVeto class and with zenith $< 100^o$ and bin the events into 21 energy bins from 500 MeV to 100 GeV. As this analysis is

[5] 1.000 realizations of 1.200 different values of $f_{\rm src}$ are used for training the ConvNets (1.2 million images). 3 realizations of 20.000 other $f_{\rm src}$ values were used to test the ConvNets (60.000 images).

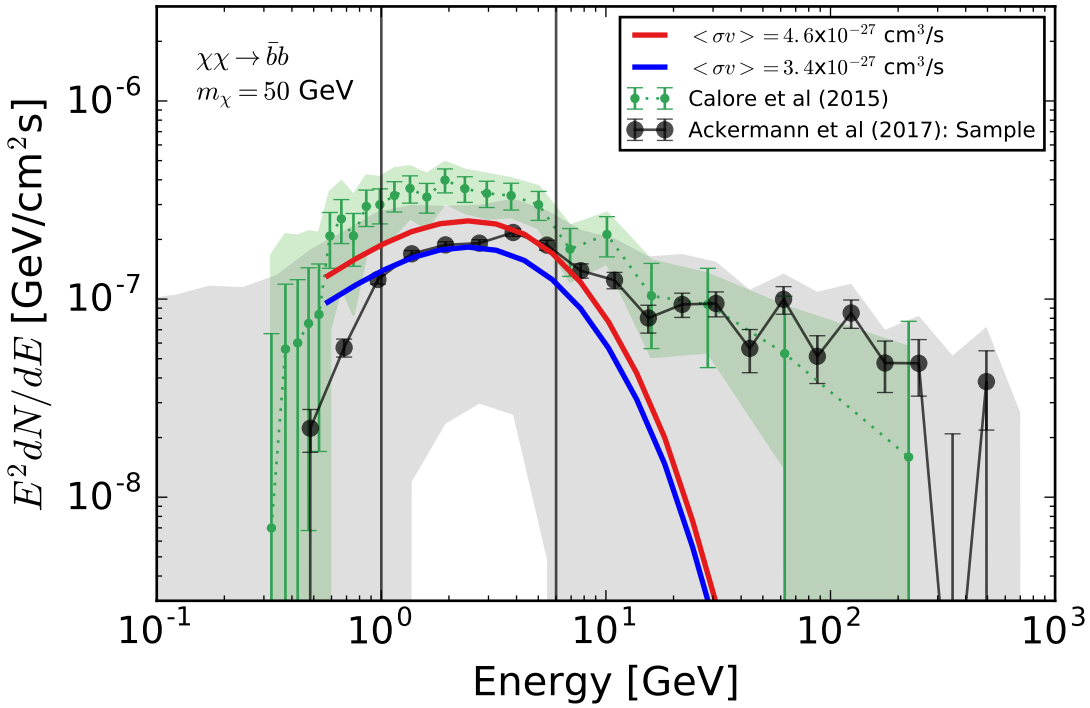


Figure 1: Spectral energy distribution of the GCE. In blue and red the spectra from fits performed in this work to the *Fermi*-LAT photon data using the models Training A and B in table 1, respectively. Following [15] we select the GCE derived using the Sample Model as representative of the GCE spectrum (black points). The variation in the GCE spectra from [15] is represented by the gray shaded area. The vertical black lines point out the energy range where the GCE is always present, we use this band to create all ConvNets analysis images. For comparison the GCE spectrum as found in [10] (green points) is plotted together with the diagonal of the covariance matrix derived there (light green band).

about images only the events converted in the front part of the *Fermi* LAT are used, as this is a good compromise between angular resolution and statistics. We re-fit the five IEMs listed in table 1, including a gNFW template, to *Fermi*-LAT data in the inner $15^o \times 15^o$ about the GC.

Standard *Fermi* tools are used to prepare data and to perform the fits. In this analysis we fit the following set of parameters:

- The normalization of the innermost rings in the IEM.
- The normalization of the brightest 3FGL sources (we do neither vary the position nor the spectral shape of the point sources).
- We model the GCE spatially with a gNFW, spectrally as a WIMP of 50 GeV annihilating into $b\bar{b}$ quarks. In this way the only free parameter for this component is the normalization, that we parametrize as the thermally averaged cross section $\langle\sigma v\rangle$.

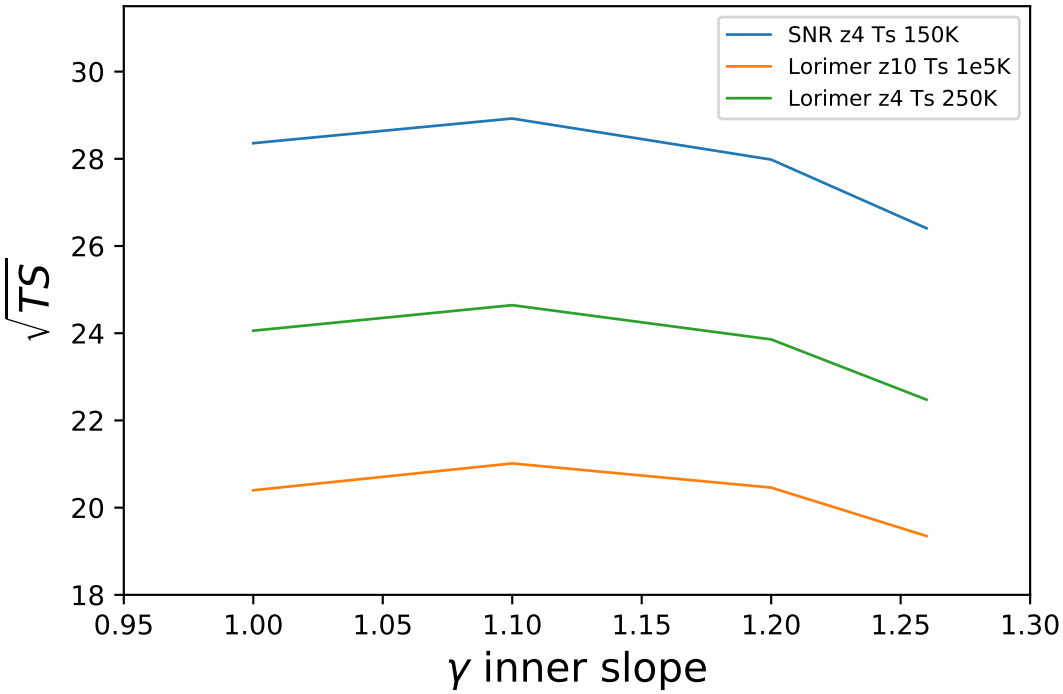


Figure 2: Best fit values for γ using the three different IEMs for training the ConvNet. TS stands for Test Statistics and is defined as $TS = -2\ln\Delta\mathcal{L}$, where $\Delta\mathcal{L}$ is the likelihood ratio between models with and without source, in our case the gNFW component [34]. The fit is always improved after including a gNFW template. In this proof-of-concept work $\gamma = 1.1$ as this value provides the larger TS over the different models tested. Although TS is well defined for point sources and we are testing an extended source, we use this quantity as an approximated proxy for comparing models.

The internal slope of the gNFW template is varied, changing the γ parameter in eq 4.2. For each value of γ we run five different fits corresponding to the five IEMs selected. Inspired by the results of [15] the following values for γ are used: 1.0, 1.1, 1.2, 1.26, and 1.3. In the simulations used for training and validating the ConvNets, γ is set at 1.1, as we find a slightly better fit with this configuration, see figure 2. The likelihood of the five selected IEMs in shown in table 1 together with the resulting $\langle\sigma v\rangle$ when the gNFW with $\gamma = 1.1$ are used as IEMs to generated the training and validation data. It is worth noting that all IEM models were previously fitted to all-sky data making it hard to compare models based on their likelihood value. However, after inspection of spatial and spectral residuals we confirm that all models listed in table 1 describe the data equally well.

4.2 Simulation of *Fermi* -LAT data

Once we get the five models of the GC region tuned to *Fermi*-LAT observations, synthetic data is generated where we modify the texture of the GCE model, keeping its magnitude, spectral shape and distribution aligned with the fits, see figure 3.

The spectral shape of the GCE is highly dependent on the assumptions of the underlying

Table 1: Results of fitting the $15^o \times 15^o$ region around the GC using 5 different IEMs. The gNFW template is created with $\gamma = 1.1$. All the likelihoods and both, spatial and spectral residuals are at the same level, implying that all models provide a similar description of the data. We assume the sources of CRs are supernova remnants (SNRs) considering two CR source distributions, one traced by the measured distribution of pulsars, Lorimer [35], and other tracing SNRs observed [36]. T_S stands for spin temperature of the atomic hydrogen for the derivation of gas column densities from the 21-cm line data.

Usage	CR distribution	Halo height z (kpc)	T_S (K)	Log$\mathcal{L}$	$\langle\sigma v\rangle \times 10^{-27}$ cm^3/s
Training A	SNR	10	150	-442855	45.59
Training B	Lorimer	10	1×10^5	-442304	33.61
Training C	Lorimer	4	150	-442357	39.32
Testing A	Lorimer	10	150	-442539	39.63
Testing B	SNR	4	1×10^5	-442664	42.67

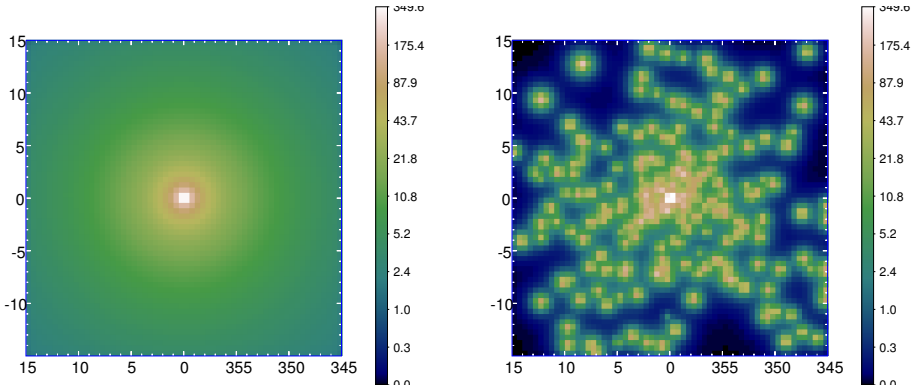


Figure 3: Counts maps in the 1-6 GeV energy range comparing the granular vs diffuse nature of the GCE, these templates are indistinguishably for the template method. Both maps have the same total emission and follow a gNFW distribution.

model, however in the 1-6 GeV energy band the excess is always present, see figure 1. Images for training and testing ConvNets are created with events in the energy band 1-6 GeV. The left panel of figure 3 shows an image made with the gNFW template ($\gamma = 1.1$) in the 1-6 GeV band. The right hand side of figure 3 shows the image of a simulated population of point sources that produces the same amount of total flux as the gNFW template. The density of sources in the population is also spatially distributed following equation 4.2 with $\gamma = 1.1$. To generate such kind of populations that mimic the GCE we make the following assumptions: All sources have the same spectral shape and are distributed randomly (see figure 4 for a sample plot of a randomized population of point sources). Besides following equation 4.2, the flux distribution in the population is modeled as a power law;

$$\frac{dN_{src}}{ds} = As^{\alpha}, \tag{4.3}$$

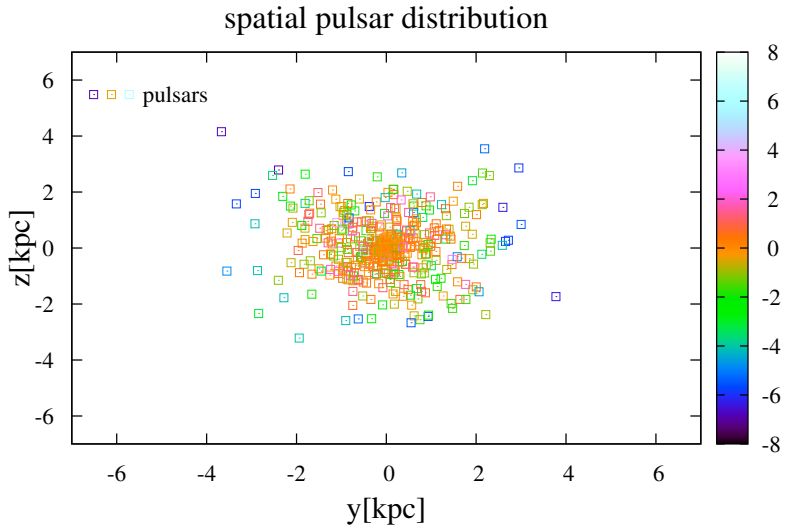


Figure 4: 3D distribution of a point source distribution following an gNFW profile with $\gamma = 1.1$. Each time such a population is generated the sources are placed at different positions, also the flux s of each source is randomly assigned. The (0,0,0) point is the GC, the color code stands for x axis in kpc.

with the normalization A determined by the total emission of the GCE as obtained in the corresponding fit and the constraint that all sources are below the 3FGL detection threshold (1×10^{-9} cm^{-2} s^{-1} for fluxes above 1 GeV [32]). s represents the source flux and $\frac{dN_{src}}{ds}$ the number of sources for a specific flux interval. The spectral index α determines the number of point sources per flux below the 3FGL detection threshold. A value of $\alpha = 1.05$ implies that almost all sources have a flux close to the detection threshold. In this analysis we let α range from -1.05 to 1.05[6].

In figure 5 we present some of the models for training, these are extreme models where the GCE is made completely by DM annihilation or by a point source population. It is worth noting that the data contains Poisson noise, as the *Fermi*-LAT is a photon-counting experiment, but the tool to generate maps only convolves flux models with the *Fermi*-LAT response and exposure to produce counts maps. Therefore, Poisson noise is added to the maps in order to make them more realistic.

One of the main concerns of this approach is the quality of the simulations versus the actual images. This is often referred to as the 'Reality Gap': the distance between simulations and actual data. A lot of effort has been taken to ensure that the simulations represent the actual data. However, there are still unknowns that go into the simulations. These are:

- Which IEM is correct? The computation of IEMs requires input data that are highly uncertain, the models of our Galaxy represent quite well the whole γ-ray sky, but in the

[6] In [25] the best-fit α value found for the Bulge pulsar population is 1.2. We do not extend our α range to cover that value as we only work with simulations, a follow-up work will be devoted to extract physical quantities from actual *Fermi*-LAT images.

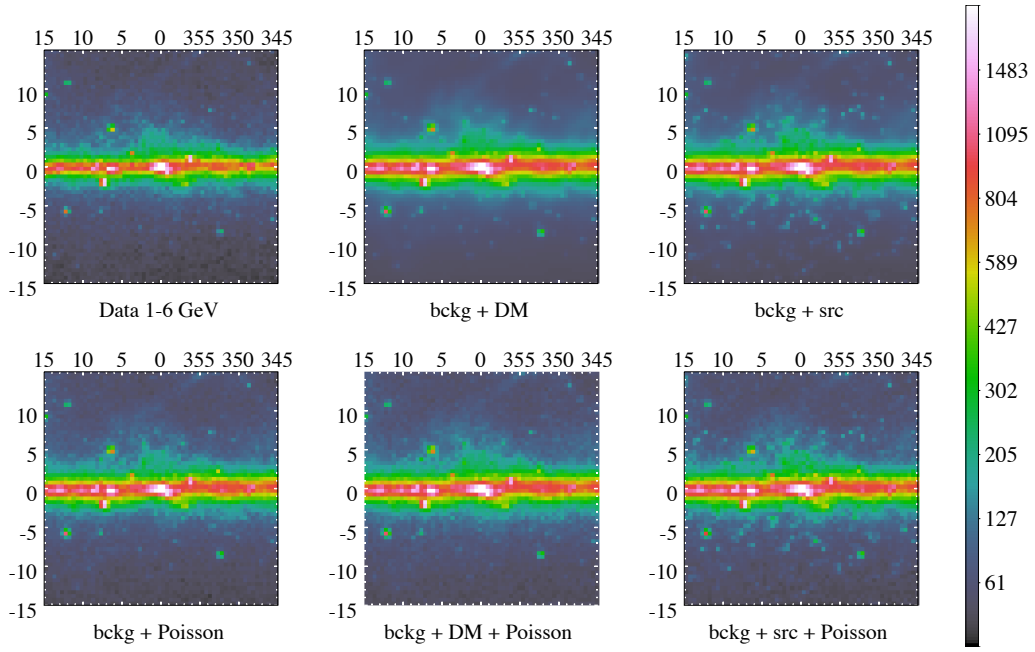


Figure 5: Images of the Galactic Center in the 1-6 GeV γ-ray band. Color code in counts/pixel. Upper left, real data. Upper center, model of the GC region as result of the fit to the Training A model, in upper right the same model but the gNFW template is replaced by a distribution of point sources. The lower row presents the same models of the first row but with poissonian noise added to make them more realistic. All the maps are generated with the gtmodel code of *Fermi* Science Tools version 11-05-00 except the data.

inner region the complexity of the environment is significantly higher. In particular, as already mentioned GALPROP-generated templates are 'smoothed' versions of reality as many smaller molecular clouds are not present in the GALPROP gas maps, i.e. the small scale structure of IEM is lacking in the training set, see appendix F of [24].

- The list of detected point sources. The 3FGL catalog is used, which is created using four years of data, while we use about seven years of data to generate the simulations. As new sources are expected to appear with more data, the results of the network are going to be biased towards higher $f_{\rm src}$: the already detected sources that are not in the 3FGL are counted towards the $f_{\rm src}$ value while they are part of the background[7].

- The positions of point sources part of the unresolved population (only their distribution is fixed in the current simulation).

- The internal distribution of the gNFW template parametrized by γ in equation 4.2, which is here fixed to 1.1.

To account for these unknowns, the following actions were taken:

[7]The 2FIG catalog presents about 200 more sources detected in the GC region than the 3FGL and uses seven years of data. In a follow up study updated catalogs will be used to remove this bias.

Training the network was done using simulations of three different background models (Training A, B, and C in table 1). The validation of the ConvNets was done on models that were generated using two background models that were not in the training set (Testing A and B in table 1). In total five background models were used. This ensures that the network cannot overfit on one particular background model and has to generalize over different background models.

To ensure that the network cannot overfit on the actual location of the point sources in the unresolved population, many initializations of the sources were taken and in every model simulation their actual positions are different. Therefore the network cannot rely on fixed positions of the point sources, but only on their distribution which is determined by the parameter α (see section 4).

5 Neural network setup

This section describes the analysis setup of the convolutional neural networks that were used to make the predictions of f_{src}. The input data of the network is a (w, h, c)-tensor, with w and h the width and height of the image and c the number of channels (colors in case of a color image). As the pixels of the image represent photon counts in the band between 1 and 6 GeV, the number of channels is one (meaning the image is monochromatic). The output of the network is a number between 0 and 1, representing the value of $f_{\rm src}$. In total five networks were trained with the same network architecture, see table 2. One network is trained on the full image, and four networks are trained on half the image. After training, the output values of the half networks can be compared to check if their predictions are in agreement and their predictions can be averaged. Combining multiple network results in this way typically improves accuracy [37] (also called *ensemble learning*).

Table 2: The five different networks trained on the simulations.

Network name	**Tensor size**
Full image	$(120, 120, 1)$
Left half of image	$(60, 120, 1)$
Right half of image	$(60, 120, 1)$
Top half of image	$(120, 60, 1)$
Bottom half of image	$(120, 60, 1)$

After training, the results of the ConvNets that were trained on half the image were averaged to get an averaged network result. The training data consists of images with a random value of α between -1.05 and 1.05. γ is always 1.1 (see section 4.1). To be sure the network does not overfit on the background model used, the training data contains images that are generated using three different background models. The validation data consists of 60.000 images from two background models that are not part of the training data (Training A and B in table 1, the images are split evenly) and with randomized α and point source locations. This means that in order to get a high accuracy on the 60.000 images in the validation set, the trained network has to predict the correct value of $f_{\rm src}$ regardless of α, BG model and point source locations, in the whole range of $f_{\rm src} = 0$ to $f_{\rm src} = 1$.

As a preprocessing step, the pixel values of an image are normalized to values between 0 and 1 (this improves training speed). The last layer of the network architecture (see Appendix

A) has a sigmoid activation function. This function goes asymptotically to 1, making an $f_{\text{src}} = 1$ prediction impossible (the input of the last layer has to be infinite in this case). To improve network accuracy for very high f_{src} predictions, the output value of a prediction is normalized with a factor of $\frac{1}{\max(f_{\text{src}})}$ to map the highest predicted output to one. Here $\max(\text{f}_{\text{src}})$ is the highest predicted value of the validation data.

5.1 Network architecture

This section explains the architecture of the ConvNet used for this work. For a general high-level introduction to ConvNets, see Appendix A. The ConvNet architecture is identical for all five networks and is visualized in figure 6. The activation functions of all inner layers are ReLU functions [38]: $f(x) = x$ if $x > 0$ and $f(x) = 0$ otherwise. The last layer has a sigmoid activation function.

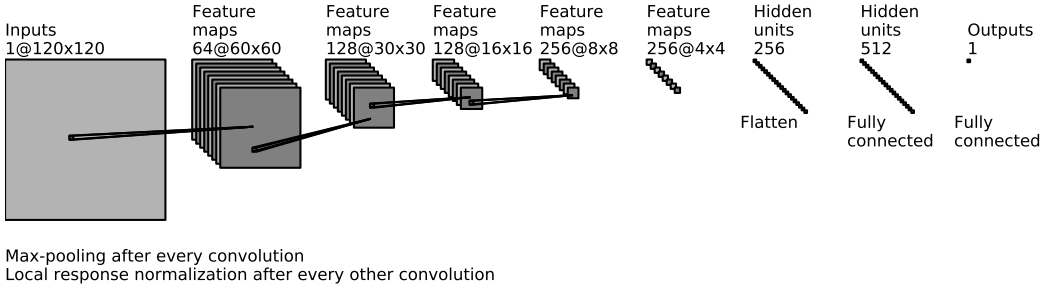


Figure 6: Visualization of the convolutional neural network. The network consists of an input layer, 5 convolutional + pooling layers, 2 fully connected layers and finally an output layer.

The loss function of the network is mean square error, the optimizer is chosen to be the Adam optimizer [39]. The network is trained in two steps: first with a learning rate of 10^{-5} and then with a learning rate of 10^{-9}, both using 20 epochs. The TensorFlow[8] library is used and training was ran on two Nvidia GTX1080 cards.

[8] https://www.tensorflow.org

6 Results

6.1 Neural network results

In figure 7 a visualization is shown of the different activations within the ConvNet layers. The input image is shown on the left. Each column represents a convolutional layer and the image passes through the network from left to right. Only four convolutional filters are shown per column for clarity, in the actual network the number of filters varies between 64 and 256. Also, because there is a max pooling layer between the convolutional layers, the images are smaller in subsequent convolutional layers. They have been zoomed to identical sizes in the visualization. The figure shows that some filters tend to filter out the diffuse background, while others seem to only take the diffuse background into account. Internally the network seems to learn to decompose images into its diffuse and point source components. For this particular simulated sample, the difference between the ConvNet prediction and the actual value of f_{src} of the sample was 0.04.

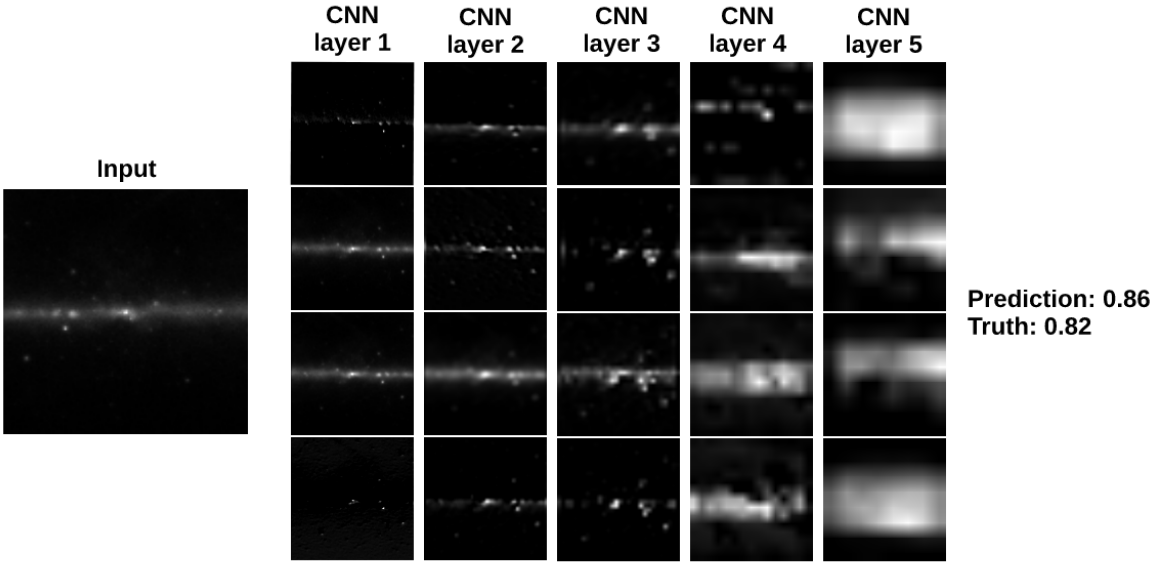
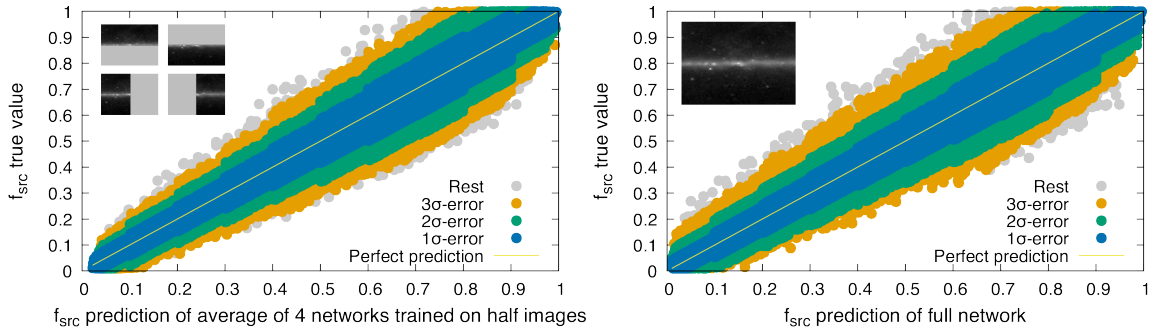


Figure 7: Activations of the different convolutional layers on a simulation. Each column of images represents a convolutional layer. The data flows from left to right (from input to prediction). Each convolutional + pooling layer accepts the image as input and outputs n smaller images, where n is the number of convolutional filters in that layer (n ranges from 64 to 256, see figure 6). Only four filters per layers are shown for clarity.

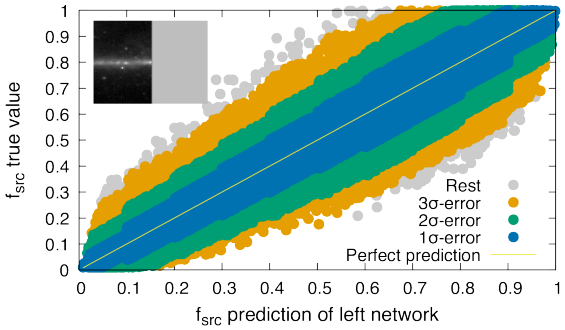
The results of the network output on the validation data is plotted in figure 8 for three different networks: the full network, the averaged network and the left network. Note that these predictions are done on two different models that were not in the training data (Testing A and B models in table 1). The accuracy of the network is calculated in 10 equally spaced bins based on this validation set. This is done because the network is more accurate for low f_{src}-predictions than for high f_{src}-predictions.

To minimize the researcher bias for the future implementation of this method only the left network was used to make a prediction on the real *Fermi* data. The output of the left



(a) Prediction of the average network versus true values.

(b) Prediction of the full network versus true values.



(c) Prediction of the left network versus true values.

Figure 8: Network results on the validation set. The different colors represent the 1σ, 2σ and 3σ bands. The diagonal line represents a perfect prediction.

Table 3: Left network properties at the predicted value on the real *Fermi* result.

Name	**Value**
Predicted output	0.887
1σ error	0.105
2σ error	0.210
3σ error	0.324
Maximum error	0.416

network is 0.887, putting the network value in the bin with the errors shown in table 3. The right image, all other half-images and the full image are still blinded and will be used in a follow up paper.

The predicted value for $f_{\rm src}$ is 0.887 ± 0.105. This favors an interpretation in terms of point sources for the GCE. However, it is worth noticing that we must not use this ConvNet on real data as the images used for training and testing the ConvNets use the 4-years 3FGL catalog while we use about 7 years of data. The newer catalog 2FIG contains about double the sources in the same region [25]. Also, the network was trained on α values between -1.05 and 1.05, while it is already showed that the simulations should at least consider values up to 1.2 [25]. For this proof-of-concept work this is not a problem, and both issues will be

addressed in the follow-up work.

The value of γ is set to 1.1 in the simulations used to generate the training data. To see if changing γ negatively impacts the predictions a few simulations were run with varying γ values. As can be seen in figure 9 this had a sizable impact on the result. For $\gamma = 1.1$ and $\gamma = 1$ the network predictions are not negatively impacted, but for higher values of γ the network under-predicts $f_{\rm src}$. Because a higher γ means that the point sources are closer to the GC, it is harder to distinguish them from a diffuse source. It is therefore expected that the network under-predicts $f_{\rm src}$ for higher values of γ. Since the most likely value of γ we obtained in our fit is between 1 and 1.1, it is expected that the network is still accurate on the real *Fermi*-LAT data. However, the next iteration of the network will be trained on multiple values of γ to ensure that the $f_{\rm src}$ predictions are accurate regardless of different γ values.

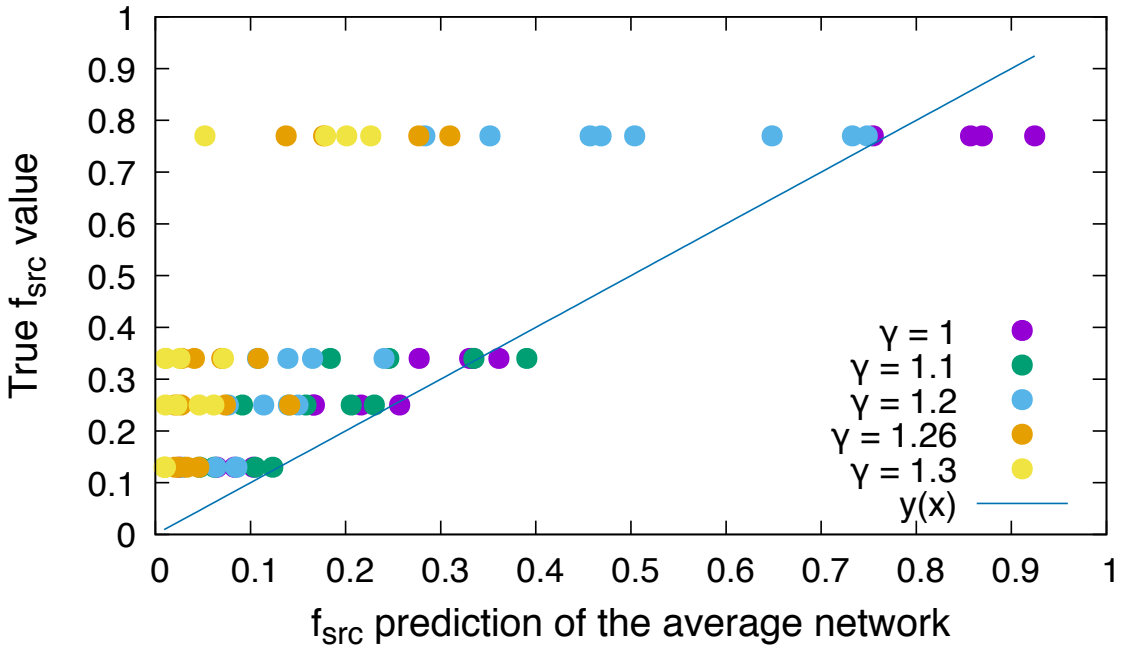


Figure 9: Predictions of the average network using different values for γ. For $\gamma = 1$ and $\gamma = 1.1$ the prediction are still accurate. However for higher values of γ the network starts to under-predict. This shows the network does not generalize over higher than $\gamma = 1.1$ values, but does for values between 1 and 1.1.

7 Conclusion

Training a convolutional neural network on simulated data of the *Fermi*-LAT yields promising results for characterize the contribution to diffuse emission from dim point sources. We find that the ConvNet method gives predictions beyond classification of either 100% point sources or diffuse source: it can also predict a mixture. The error on the prediction of f_{src} is usually in the order of 10% for the averaged network. Furthermore, this method is not limited to predicting f_{src}. One can also make predictions on the properties of the source population, i.e. α and γ, as well as the most likely background model by training networks on those particular values.

When applied to the real data, the left network prediction of the point source fraction f_{src} of the Galactic Center excess is 0.887 ± 0.105. This strongly disfavors a diffuse source only explanation of the Galactic Center excess. This is comparable to other studies [23–25]. However, this is a proof-of-concept paper and our results are biased towards higher f_{src} by the use of a list of detected sources (3FGL) created using three-years less *Fermi*-LAT data than used in this analysis. New catalogs, such as 2FIG, already contains twice the amount of sources in the same region. Also, the α parameter range should be extended from ± 1.05 to at least ± 1.2 to account for best-fit values found by other studies.

Currently the training data of the network is a value per pixel which represents the amount of photon events between 1 and 6 GeV. However, more information can be fed into the network in order to make more accurate predictions and allow for more generalizations. This includes multiple values of γ, as discussed in the previous section. Instead of a $(w, h, 1)$-tensor representing the counts of 1-6 GeV photons, the next iteration of this network will use a (w, h, c)-tensor, where c is the number of bins. Each bin will contain the pixel count of some specific energy bin. With this extra dimension of the data, the network can learn between correlations in any direction of the tensor. This extra information may lead to improved performance and will allow more complex architectures. The full analysis will yield the following improvements:

- Increase sensitivity of α to realistic values from ± 1.05 to at least ± 1.2
- Generalize over multiple values of γ instead of assuming $\gamma = 1.1$
- Use the latest point source catalog instead of 3FGL
- Use a richer data structure to improve accuracy (use multiple energy bins as channels instead of one channel)

Appendices

A ConvNets

Recently the field of deep learning has received a lot of attention because of their predictive power. Mainly in the field of computer vision deep learning has made many breakthroughs, which can be mainly attributed to convolutional neural networks. These networks can be used to detect objects in images, create text-to-speech algorithms and many more applications. This novel approach utilizes the many advances in machine learning of the past years and can lead to better predictions using less assumptions, as ConvNets typically require the raw data as input. Using DNNs over conventional data analysis methods has up- and downsides:

- DNNs work with raw data and learn to recognize correlations between the data automatically. This has two advantages: 1) there is no need to prune the data and 2) the raw data contains all possible information. Pruned data does not. A conventional method that uses preprocessed data cannot access all the information that is contained in the raw data. The quality of the pruned data is entirely dependent on the human understanding of the data and it might happen one unknowingly removes correlations when preprocessing data.
- DNNs are very general. The network architecture changes from problem to problem, but the layer types and methods can be used in many different problem areas like computer vision, regression, speech analysis and many more domains. This means the advances in one area of research (for example face detection) also benefit all the other research areas.
- DNNs can generalize over randomness. When a DNN needs to learn magnitudes of stars in a 2D image, the randomized location of the star carries no information. By using convolutional layers and enough training samples the network can learn to 'ignore' the location of the star. This is an important feature in the analysis in this work: the unresolved point sources have a random location in our simulations.
- DNNs are hard to train. There are many hyperparameters that need to be tuned to specific values in order for the network to make accurate predictions. These hyperparameters need to be set before training and requires knowledge of the problem, the network and trail-and-error.
- DNNs require a lot of data and processing power. This is one of the main reasons DNNs are only becoming popular in recent years. The hardware required to train networks that can make accurate predictions on real world data is only available since the advent of the GPU. Also, deep neural networks typically require a lot of training examples to train well.

In particular the need of large amount of data may be seen as an issue in applying DNNs to the *Fermi*-LAT GC data, as there is only one γ-ray image of the GC region. However, the advanced knowledge of the *Fermi*-LAT instrument and the availability of state-of-the-art interstellar diffuse models make it possible to create realistic mock *Fermi*-LAT data for training and testing examples. Our interest is to design a ConvNet able to predict if the GCE is due to a truly diffuse source, a population of unresolved point sources or a combination, therefore we must create a set of images covering the different possibilities.

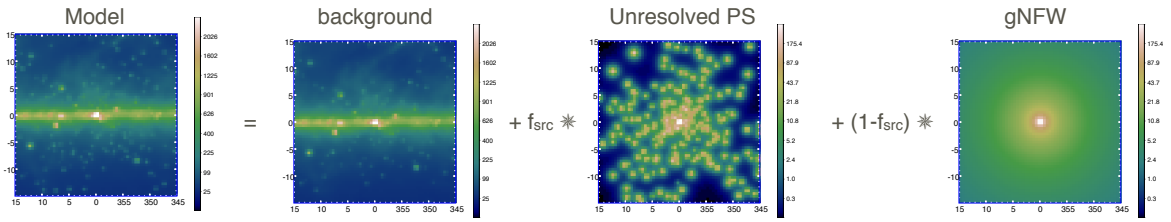


Figure 10: Visualization of how a model is constructed. The model (that is used as training data for the ConvNets) is generated by adding a fraction f_{src} of point sources and a fraction of $1 - fsrc$ of diffuse source to the background. The sum of the diffuse and point source fraction that is added on top on the background always equals the GC excess.

A.1 ConvNet data pipeline

In general, a ConvNet takes an N-dimensional input, transforms it using different layers, and produces an M-dimensional output. This can be a prediction that a specific input belongs to a particular class (e.g. object detection) or a prediction of a regression problem (e.g. this research). The input of a computer vision problem is typically an image. The image can be represented by a (w, h, c)-tensor, where w and h represent the width and height of the network and c the number of channels. For instance: a 120×120 color image has $120 \cdot 120 \cdot 3 = 43.200$ different values embedded into it. In this analysis the number of channels is one and it represents the photon count in the energy bin 1-6 GeV. The input is the first layer of a neural network and in the example consists of 43.200 neurons. When an image is fed into the network, each input neuron consists of one color of one pixel value. In left hand side of figure 10 we present an example of an image for training. It is composed of a background emission, and the GCE, whose granularity is controlled by the *fraction of point sources* f_{src} parameter, which varies between 0 and 1. A value of $f_{\mathrm{src}} = 0$ means the GCE is composed of only a diffuse source, and $f_{\mathrm{src}} = 1$ means the GCE is composed of only point sources. With ConvNets we transform such images into predictions of the parameter f_{src}.

In ConvNets, between the input and output layers there are so-called *hidden layers*. In this work we use the following kinds of hidden layers: fully connected, convolutional, max pooling and local response normalization.

A.1.1 Layer Architectures

A fully connected layer connects all neurons of a layer to the next one. A weight w is assigned to each connection (see figure 11). The value $n_{y,j}$ of a particular neuron j in layer y, is defined as $n_{y,j} = f\left(\sum_i n_{x,i} \cdot w_{x,i;y,j} + \mathrm{bias}\right)$, where $f(x)$ is called the activation function. In this research the internal layers have a Rectified Linear Unit (ReLU) activation function ($f(x > 0) = x$ and $f(x < 0) = 0$). The output layer has an activation function of a sigmoid function, to map all input values to the range (0,1), which represents the output range of f_{src} parameter.

During training, an image is fed into the network and an f_{src} output value is predicted. This output can be compared to the actual f_{src} used to create the image. The error is backpropagated into the network which *adjusts* all w and biases [40]. Before training all weights and biases are set to normal distributed random numbers (the network initialization).

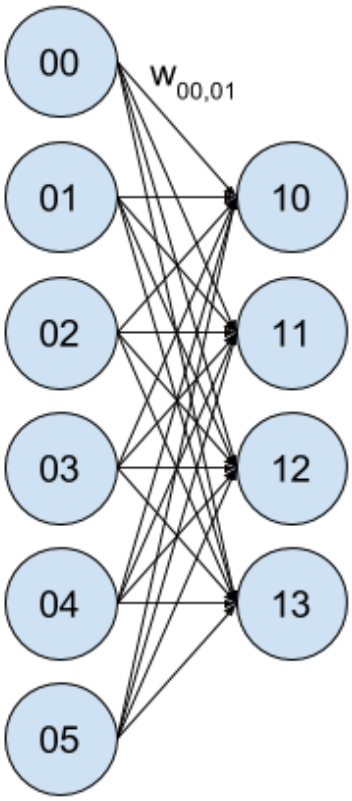


Figure 11: Schematic view of a fully connected layer with 6 input and 4 output neurons. The 24 lines represent the weights of the connections between the the neurons from the two layers.

A DNN with only fully connected layers has many parameters (weights and biases) to train. For instance: a 120x120 color image has 43.200 input neurons. If the neural network has one hidden layer with 100 neurons and one output neuron, the amount of weights in the network is $43.200 \cdot 100 + 100 \cdot 1 = 4.320.100$. This makes fully connected networks very hard to train.

A convolutional layer is a layer that has less weights than a fully connected layer by utilizing the fact that there is a correlation between neighboring pixels on an image [41]. Max pooling layers take a patch of an image and returns the maximum of this value [42]. A max pooling layer is used to reduce the dimensions of an image during network propagation. After multiple convolutional and pooling layers, fully connected layers are added to make a final prediction of the image. A graphical representation of a typical convolutional neural network can be found in figure 2 of [41].

A.1.2 Choice of activation function

The choice of the activation functions of the neurons is very important, as this introduces a non-linearity in the network and a good choice of the activation function speeds up training.

If the activation function would be linear, for example $f(x) = x$, the neural network would just be a linear combination of linear functions. If this is the case, it is impossible to make predictions about non-linear problems. Until a few years ago the sigmoid and hyperbolic tangent (tanh) function were used a lot, because they are bound between 0 and +1 for the sigmoid function and -1 and +1 for the tanh function. However, they suffer from the so-called vanishing gradient problem [43]: the gradient at large positive or negative values go to zero. As the gradient of the activation function is a used in the backpropagation algorithm to update the weights, the network effectively stops learning when the gradient is (almost) zero.

The ReLU activation function is currently a popular choice, because it has no vanishing gradient, is non-linear. On top of that, the derivative of the ReLU function is computationally trivial, while calculating the derivative of a sigmoid or tanh function is quite expensive. However, the ReLU function has no gradient for negative values, leading to so-called 'dying ReLUs': if the update of a weight makes the output of the ReLU negative, the particular neuron is effectively dead as there is no gradient anymore. This means the weights will not be updated anymore during training. There are many more activation functions that can be chosen and each of them have advantages and disadvantages. In this work the ReLU activation function was chosen because of its speed and non-vanishing gradient.

A.2 Training a neural network

A neural network has many weights in it, which are initialized randomly according to some distribution (typically normal distributed). To make accurate predictions, the weights need to be set to very particular values. This is done during the training phase of the network. During training, the network receives an image and calculates an (incorrect) output. The error is backpropagated into the network and the weights are set accordingly. There are many backpropagation algorithms, of which the Adam optimizer is a popular one and the one used in this research[39].

A typical difficulty of training neural networks is overfitting. In this case the network can predict every example in the training data perfectly, while it fails to generalize over the important features. Instead of learning those, it 'memorized' the training data and cannot be used for other cases. To determine whether a network is overfitted to the training data, one can set aside a part of the dataset and not use it during training. Then after the network is trained, calculate the accuracy of the network on this validation set. It is a very good indication that the network has overfitted if the training set accuracy is much higher than the validation set accuracy. There are a number of methods to prevent overfitting. In this research, the convolutional layers have an L2 regularizer to suppress really small weights [44]. Adding dropout has been considered as well, but just using L2 regularizers was enough to prevent overfitting and adding a dropout layer did not improve the network.

In order to force the network that it cannot overfit to particular values that are not known (the actual positions of the point sources, α, or the particular background model used for example), many images are generated that span the possible space of values. In order to achieve good accuracy, the network has to generalize over these values. This means many realizations of the simulations are needed. As long as the real data is somewhere inside the training box spanned by f_{src}, randomized locations of unresolved point source population and background model, and the ConvNet generalizes over all these values, the network output is reliable.

Acknowledgements

The *Fermi* LAT Collaboration acknowledges generous on-going support from a number of agencies and institutes that have supported both the development and the operation of the LAT as well as scientific data analysis. These include the National Aeronautics and Space Administration and the Department of Energy in the United States; the Commissariat a l'Energie Atomique and the Centre National de la Recherche Scientifique/Institut National de Physique NuclÃľaire et de Physique des Particules in France; the Agenzia Spaziale Italiana and the Istituto Nazionale di Fisica Nucleare in Italy; the Ministry of Education, Culture, Sports, Science and Technology (MEXT), High Energy Accelerator Research Organization (KEK), and Japan Aerospace Exploration Agency (JAXA) in Japan; and the K. A. Wallenberg Foundation, the Swedish Research Council, and the Swedish National Space Board in Sweden. Additional support for science analysis during the operations phase is gratefully acknowledged from the Istituto Nazionale di Astrofisica in Italy and the Centre National d'Etudes Spatiales in France.

The authors thanks Mattia di Mauro, Manuel Meyer and Gabriela Zaharijas for fruitful comments on the manuscript. R. RdA, is supported by the Ramón y Cajal program of the Spanish MICINN and also thanks the support of the Spanish MICINN's Consolider-Ingenio 2010 Programme under the grant MULTIDARK CSD2209-00064, the Invisibles European ITN project (FP7-PEOPLE-2011-ITN, PITN-GA-2011-289442-INVISIBLES, the "SOM Sabor y origen de la Materia" (FPA2011-29678), the "Fenomenologia y Cosmologia de la Fisica mas alla del Modelo Estandar e lmplicaciones Experimentales en la era del LHC" (FPA2010-17747) MEC projects and the Spanish MINECO Centro de Excelencia Severo Ochoa del IFIC program under grant SEV-2014-0398. The work of GAGV was supported by Programa FONDECYT Postdoctorado under grant 3160153.

References

[1] I. V. Moskalenko and A. W. Strong, "Production and propagation of cosmic ray positrons and electrons," *Astrophys. J.*, vol. 493, pp. 694–707, 1998.

[2] A. W. Strong, I. V. Moskalenko, and O. Reimer, "Diffuse continuum gamma-rays from the galaxy," *Astrophys. J.*, vol. 537, pp. 763–784, 2000. [Erratum: Astrophys. J.541,1109(2000)].

[3] L. Goodenough and D. Hooper, "Possible Evidence For Dark Matter Annihilation In The Inner Milky Way From The Fermi Gamma Ray Space Telescope," *ArXiv e-prints*, Oct. 2009.

[4] V. Vitale and A. Morselli, "Indirect Search for Dark Matter from the center of the Milky Way with the Fermi-Large Area Telescope," in *Fermi gamma-ray space telescope. Proceedings, 2nd Fermi Symposium, Washington, USA, November 2-5, 2009*, 2009.

[5] D. Hooper and L. Goodenough, "Dark Matter Annihilation in The Galactic Center As Seen by the Fermi Gamma Ray Space Telescope," *Phys. Lett.*, vol. B697, pp. 412–428, 2011.

[6] C. Gordon and O. Macias, "Dark Matter and Pulsar Model Constraints from Galactic Center Fermi-LAT Gamma Ray Observations," *Phys. Rev.*, vol. D88, no. 8, p. 083521, 2013. [Erratum: Phys. Rev.D89,no.4,049901(2014)].

[7] D. Hooper and T. Linden, "On The Origin Of The Gamma Rays From The Galactic Center," *Phys.Rev.*, vol. D84, p. 123005, 2011.

[8] T. Daylan, D. P. Finkbeiner, D. Hooper, T. Linden, S. K. N. Portillo, N. L. Rodd, and T. R. Slatyer, "The characterization of the gamma-ray signal from the central Milky Way: A case for annihilating dark matter," *Phys. Dark Univ.*, vol. 12, pp. 1–23, 2016.

[9] A. Boyarsky, D. Malyshev, and O. Ruchayskiy, "A comment on the emission from the Galactic Center as seen by the Fermi telescope," *Physics Letters B*, vol. 705, pp. 165–169, Nov. 2011.

[10] F. Calore, I. Cholis, and C. Weniger, "Background model systematics for the Fermi GeV excess," *JCAP*, vol. 3, p. 038, Mar. 2015.

[11] K. N. Abazajian, N. Canac, S. Horiuchi, and M. Kaplinghat, "Astrophysical and Dark Matter Interpretations of Extended Gamma-Ray Emission from the Galactic Center," *Phys. Rev.*, vol. D90, no. 2, p. 023526, 2014.

[12] B. Zhou, Y.-F. Liang, X. Huang, X. Li, Y.-Z. Fan, L. Feng, and J. Chang, "GeV excess in the Milky Way: The role of diffuse galactic gamma-ray emission templates," *Phys. Rev.*, vol. D91, no. 12, p. 123010, 2015.

[13] P. Agrawal and L. Randall, "Point Sources from Dissipative Dark Matter," 2017.

[14] **Fermi/LAT** Collaboration, "Fermi-LAT Observations of High-Energy Gamma-Ray Emission toward the Galactic Center," *ApJ*, vol. 819, p. 44, Mar. 2016.

[15] M. Ackermann *et al.*, "The Fermi Galactic Center GeV Excess and Implications for Dark Matter," *Astrophys. J.*, vol. 840, no. 1, p. 43, 2017.

[16] J.-P. Macquart and N. Kanekar, "On Detecting Millisecond Pulsars at the Galactic Center," *Astrophys. J.*, vol. 805, no. 2, p. 172, 2015.

[17] A. McCann, "A stacked analysis of 115 pulsars observed by the Fermi LAT," *Astrophys. J.*, vol. 804, no. 2, p. 86, 2015.

[18] N. Mirabal, "Dark matter vs. Pulsars: Catching the impostor," *Mon. Not. Roy. Astron. Soc.*, vol. 436, p. 2461, 2013.

[19] R. M. O'Leary, M. D. Kistler, M. Kerr, and J. Dexter, "Young Pulsars and the Galactic Center GeV Gamma-ray Excess," *ArXiv:1504.02477*, Apr. 2015.

[20] Q. Yuan and B. Zhang, "Millisecond pulsar interpretation of the Galactic center gamma-ray excess," *Journal of High Energy Astrophysics*, vol. 3, pp. 1–8, Sept. 2014.

[21] J. Petrović, P. D. Serpico, and G. Zaharijas, "Millisecond pulsars and the Galactic Center gamma-ray excess: the importance of luminosity function and secondary emission," *JCAP*, vol. 1502, no. 02, p. 023, 2015.

[22] I. Cholis, D. Hooper, and T. Linden, "Challenges in Explaining the Galactic Center Gamma-Ray Excess with Millisecond Pulsars," *JCAP*, vol. 1506, no. 06, p. 043, 2015.

[23] S. K. Lee, M. Lisanti, B. R. Safdi, T. R. Slatyer, and W. Xue, "Evidence for Unresolved γ -Ray Point Sources in the Inner Galaxy," *Physical Review Letters*, vol. 116, p. 051103, Feb. 2016.

[24] R. Bartels, S. Krishnamurthy, and C. Weniger, "Strong Support for the Millisecond Pulsar Origin of the Galactic Center GeV Excess," *Physical Review Letters*, vol. 116, p. 051102, Feb. 2016.

[25] M. Ajello *et al.*, "Characterizing the population of pulsars in the Galactic bulge with the *Fermi* Large Area Telescope," *Submitted to: Astrophys. J.*, 2017.

[26] D. Hooper and G. Mohlabeng, "The Gamma-Ray Luminosity Function of Millisecond Pulsars and Implications for the GeV Excess," *JCAP*, vol. 1603, no. 03, p. 049, 2016.

[27] D. Hooper and T. Linden, "The Gamma-Ray Pulsar Population of Globular Clusters: Implications for the GeV Excess," *JCAP*, vol. 1608, no. 08, p. 018, 2016.

[28] D. Haggard, C. Heinke, D. Hooper, and T. Linden, "Low Mass X-Ray Binaries in the Inner Galaxy: Implications for Millisecond Pulsars and the GeV Excess," *JCAP*, vol. 1705, no. 05, p. 056, 2017.

[29] H. Ploeg, C. Gordon, R. Crocker, and O. Macias, "Consistency Between the Luminosity Function of Resolved Millisecond Pulsars and the Galactic Center Excess," *JCAP*, vol. 2017, no. 08, p. 015, 2017.

[30] J. F. Navarro, C. S. Frenk, and S. D. M. White, "A Universal Density Profile from Hierarchical Clustering," *ApJ*, vol. 490, pp. 493–508, Dec. 1997.

[31] J. F. Navarro, C. S. Frenk, and S. D. White, "The Structure of cold dark matter halos," *Astrophys.J.*, vol. 462, pp. 563–575, 1996.

[32] F. Acero *et al.*, "Fermi Large Area Telescope Third Source Catalog," *Astrophys. J. Suppl.*, vol. 218, no. 2, p. 23, 2015.

[33] M. Ackermann *et al.*, "The Third Catalog of Active Galactic Nuclei Detected by the Fermi Large Area Telescope," *Astrophys. J.*, vol. 810, no. 1, p. 14, 2015.

[34] A. A. Abdo, M. Ackermann, M. Ajello, A. Allafort, E. Antolini, W. B. Atwood, M. Axelsson, L. Baldini, J. Ballet, G. Barbiellini, and et al., "Fermi Large Area Telescope First Source Catalog," *Astrophysical Journal, Supplement*, vol. 188, pp. 405–436, June 2010.

[35] D. Lorimer, A. Faulkner, A. Lyne, *et al.*, "The Parkes multibeam pulsar survey: VI. Discovery and timing of 142 pulsars and a Galactic population analysis," *Mon. Not. Roy. Astron. Soc.*, vol. 372, pp. 777–800, 2006.

[36] G. L. Case and D. Bhattacharya, "A new sigma-d relation and its application to the galactic supernova remnant distribution," *Astrophys. J.*, vol. 504, p. 761, 1998.

[37] R. Polikar, "Ensemble based systems in decision making," *IEEE Circuits and Systems Magazine*, vol. 6, no. 3, p. 021, 2006.

[38] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)* (J. FÃijrnkranz and T. Joachims, eds.), pp. 807–814, Omnipress, 2010.

[39] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.

[40] Y. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, "Effiicient backprop," in *Neural Networks: Tricks of the Trade, This Book is an Outgrowth of a 1996 NIPS Workshop*, (London, UK, UK), pp. 9–50, Springer-Verlag, 1998.

[41] 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.

[42] M. D. Zeiler and R. Fergus, *Visualizing and Understanding Convolutional Networks*, pp. 818–833. Cham: Springer International Publishing, 2014.

[43] S. Hochreiter, Y. Bengio, P. Frasconi, and J. Schmidhuber, "Gradient flow in recurrent nets: the difficulty of learning long-term dependencies," 2001.

[44] A. Y. Ng, "Feature selection, l1 vs. l2 regularization, and rotational invariance," in *Proceedings of the Twenty-first International Conference on Machine Learning*, ICML '04, (New York, NY, USA), pp. 78–, ACM, 2004.