Regularized Deep Learning in High Energy Physics

Josh Gartman
Northeastern University
Boston, MA
gartman.j@husky.neu.edu

Abstract

The ABSTRACT is to be in fully-justified italicized text, at the top of the left-hand column, below the author and affiliation information. Use the word "Abstract" as the title, in 12-point Times, boldface type, centered relative to the column, initially capitalized. The abstract is to be in 10-point, single-spaced type. Leave two blank lines after the Abstract, then begin the main text. Look at previous CVPR abstracts to get a feel for style and length.

1. Introduction

Recent years have seen a dramatic increase in the popularity of deep neural networks. These networks have shown improved performance over existing methods in diverse areas such as computer vision, speech recognition, and text analysis. Multiple hidden layers and non-linear activations allow deep networks the flexibility to model complex functions more efficiently and with better generalization that their shallow counterparts. The use of shallow neural networks has been common practice in high energy physics since the 1980's. An important application of these networks is in classifying the subatomic particles produced by collisions at particle accelerators.

At particle accelerators like the Large Hadron Collider, located outside Geneva Switzerland, protons are accelerated to nearly the speed of light by powerful magnets and then smashed together with resulting collisions observed by a series of detectors. At high enough energies rare and unstable particles can be produced. The data from these collisions including the speed and trajectories of the resulting particles can then be input to machine learning algorithms to classify the particles themselves. The important benefits of improved machine learning algorithms are two-fold. Firstly, better classification accuracy can improve the chances of correctly classifying a potentially rare or undiscovered particle. Secondly, improved algorithms can learn on smaller training sets. Although training data may be created by computer simulation it can be computationally expensive to produce and unwieldy to use. Smaller training sets allow for models to be trained more quickly and may in some cases be less prone to overfitting. The goal of this paper is to apply regularized deep learning to datasets of limited size while maintaining or improving classification accuracy when compared to shallow networks.

An recent example of the use of machine learning in high energy physics is in the search for decays of the Higgs Boson directly into fermions at the LHC. Evidence consistent with the decay of the a Higgs particle into two fermionic tau leptons has been seen in data collected at the LHC but current methods lack the statistical power to cross the standard threshhold for claims of discovery [1]. Observing the $H \to \tau^+\tau^-$ decay would provide further verification that the Standard Model of particle physics is an appropriate description of the behavior of sub-atomic particles.

1.1. Related Work

Shallow networks have been used for decades in particle classification problems but only recently has the use of deep networks been explored. Baldi et. al. [5][6] apply deep learning to several high energy physics classification tasks. First in [5], the authors investigate the performance of deep classifiers in detecting the production of the Higgs boson and separately their performance in classifying the production of super-symmetric charged particles that decay to W bosons. They observed that deep networks utilizing low level data and high level derived features give improved performance over shallow networks trained with the same features. In the Higgs production task the classifiers performance was not improved by the inclusion of the high level features indicating that the network was able to learn a representation of these features simply from the low level raw input data. However, in the super-symmetry task the inclusion of high level features led to improved performance making it difficult to generalize whether deep learning can learn representations of these high level features for all particle physics datasets. Both datasets were of similar size, containing roughly 10 millions training examples with about 30 features for each example.

In [6] the authors focus on the previously described task of searching for the decay of the Higgs into two tau leptons. Their data set was very large cosisting of 80 million examples. They compare the performance of deep and shallow networks in detecting the decay using combinations of both high and low level features. Results indicate that deep networks outperform shallow networks even in the case where the shallow network was trained with the full feature set and the deep network was only given the low level features.

1.2. Overall Approach

The goal of this paper is to replicate the successes of deep learning in particle classification with a training set of limited size. On of the main difficulties with training neural networks is that they can be prone to overfitting. To address this L2 regularization and dropout were explored. In [6] the authors observed that regularization methods did not not improve their results, speculating that because of the size of their training set the main challenge their model faces was learning rather than preventing overfitting. Another challenge of training deep neural networks is optimizing hyper-parameters such as the learning rate, number of hidden layers, and number of units per hidden layer. Because training a neural network is a very computationally expensive procedure it is beneficial to optimize hyper-parameters as efficiently as possible. In grid search the network is trained repeatedly with different pre-determined values of the hyper-parameters. Grid search can be computationally wasteful since it does not focus the search on a space of hyper-parameters likely to give the best results. A Bayesian hyper-parameter optimization procedure is able to use past evaluations of the model to select hyper-parameters most likely to give the best performance which can save valuable computation time.

2. Technical Details of Approach

The following subsections describe in detail the components of the model.

2.1. Hyper-parameter Optimization

The large computational resources that must be spent to train machine learning models such as neural networks can make hyper-parameter optimization difficult. The benefit of a Bayesian hyper-parameter optimization approach is that the choice of hyper-parameters to evaluate can be chosen so as to give the best chance of improving the models performance. In Snoek et. al. [10] the authors describe a Bayesian hyper-parameter optimization scheme base on Guassian Processes. In this scheme the objective function being minimized $f(x)$ is modeled as though it is drawn from a Guassian process prior. After $f(x)$ is evaluated for a given setting of the hyper-parameters an acquisition function is used to try and find a new setting of parameters most likely to minimize the objective. Because of the properties of Gaussian Processes this acquisition function that gives the greatest expected improvement has a closed form. This is given as:

$$a_{EI}(x : x_n, y_n, \theta) = \sigma(x; x_n, y_n, \theta)(\gamma(x)\Phi(\gamma(x)) + N(\gamma(x)))$$
$$\gamma(x) = \frac{f(x_{best}) - \mu(x; x_n, y_n, \theta)}{\sigma(x; x_n, y_n, \theta)}$$

Where $\sigma^2(x; x_n, y_n, \theta)$ is the predictive covariance function, $\mu(x)$ is the predictive mean function, $Phi(x)$ is the cumulative distribution function of the standard normal distribution and $N(x)$ is the probability density function of the standard normal distribution.

In GP it is common to use a squared exponential kernel. However, for this task the authors state that a squared exponential kernel will produce functions that are unrealistically smooth for practical optimization problems. Instead they recommend using the automatic relevance determination Matèrn 5/2 kernel which has the following form:

$$K_{M52}(x, x') = \theta_0(1 + \sqrt{5r^2(x, x')} + \tfrac{5}{3}r^2(x, x'))exp[-\sqrt{5r^2(x, x')}]$$
$$r^2(x, x') = \textstyle\sum_{d=1}^{D}(x_d - x'_d)^2/\theta_d^2$$

The θ are length scale parameters. In [10] this Bayesian hyper-parameter optimization scheme is show to give improved performance over random and grid search on benchmark optimization tasks.

2.2. Gradient Based Optimization Techniques

In it's simplest form gradient descent optimization seeks to minimize an objective function by taking steps in the direction of the functions direction of greatest decrease. One issue with gradient descent algorithms that use a fixed step size is that as the optimization procedure approaches a minima the gradient shrinks and progress towards the minima slows down. This is the motivation for gradient based optimization procedures with variable step sizes. Adam, which stands for adaptive moment estimation is a recently proposed variable step size optimization method [8]. Adam computes individual adaptive learning rates for parameters based on estimates of the first and second moments of the gradient. Some of the advantages of Adam are that it does not require a stationary objective, it's step sizes are bounded and it works well with sparse gradients.

Formally, the Adam optimization method can be described as follows. Let $f(\theta)$ be an objective that is differentiable with respect to its parameters θ. $g_t = \nabla_\theta f_t(\theta)$ is the gradient of f with respect to θ evaluated at time step t. At each iteration the algorithm updates exponential moving averages of the gradient (m_t) and squared gradient (v_t) according to the rule:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1)g_t$$
$$v_t = \beta_2 v_{t-1} + (1 - \beta_2)g_t^2$$

The β_1 and β_2 are hyper-parameters controlling the exponential decay rates of these moving averages. The (m_t) and (v_t) are estimates of the first moment (mean) and second moment (variance) of the gradient. In [8] the authors show that Adam gives improved performance over other gradient based optimization methods such as RMS Prop or AdaGrad.

2.3. Dropout

Although neural networks have been in use for decades, difficulties with overfitting in deep networks meant that until recently only single layer networks were viable. Dropout is a method that has shown great promise in ameliorating the problems with training deep networks. In dropout hidden units are randomly dropped with some probability p during training [11]. The model is described as:
$r_j^{(l)}\ Bernoulli(p)$
$\tilde{y}^{(l)} = r^{(l)} * y^{(l)}$
$z_i^{(l+1)} = w_i^{(l+1)}\tilde{y}^l + b_i^{(l+1)}$
$y_i^{(l+1)} = f(z_i^{(l+1)})$
f is an activation function and $*$ denotes an element wise product. Dropout can be viewed as a form of model averaging. During each training iteration a different set of neurons are randomly dropped out so essentially a new network is being used. At test time the models weights can be rescaled essentially resulting in an averaging of all the training networks. Another view of dropout is that it prevents units from becoming co-adapted with one another. Since each hidden unit must learn to work with random subsets of the other hidden units it is driven toward creating useful features on its own without relying on other hidden units.

2.4. Weight Decay

In addition to dropout, another method of preventing overfitting in neural networks is weight decay. When a model begins to overfit its parameters become tuned to noise in the training data and weights have a tendency to take on extremely large or small values. Weight decay is a way of limiting overfitting by promoting weights that are smaller in magnitude. For an error function $E_0(w)$ L2 weight decay is described as follows:
$E(w) = E_0(w) + \frac{1}{2}\lambda\sum_i w_i$
In [9] the authors show that an L2 weight decay is able to improve the generalization performance of a neural network with non-linear hidden units. Weight decay can be viewed as a manifestation of Occam's razor. It chooses the simplest model that is capable of solving a problem. It is noted in [11] that a combination of weight decay and dropout can lead to better performance than either method used in isolation since using weight decay allows the learning rate to be dramatically increased without worrying ...

2.5. Activation Functions

Traditionally neural networks were training with sigmoid or tangent hyperbolic activation functions of the hidden units. A drawback of training networks with these activation functions is that gradients computed during backpropagation can quickly vanish making it difficult to update the weights in the networks initial layers. Eventually it was observed that better results could be achieved by unsupervised pre-training of the network [12]. Further investigation has revealed that the unsupervised pre-training can be replaced by a change of activation function [7]. A rectified linear unit or ReLU is an activation function described as:
$f(x) = max(0, x)$

One important advantage of ReLU when compared with sigmoid activations is that ReLU encourages sparsity in the model. As mentioned in [7] sparse models are more likely to be linearly separable and may better disentangle the factors explaining variations in the data.

2.6. Paper length

Papers, excluding the references section, must be no longer than eight pages in length. The references section will not be included in the page count, and there is no limit on the length of the references section. For example, a paper of eight pages with two pages of references would have a total length of 10 pages. **There will be no extra page charges for CVPR 2017.**

Overlength papers will simply not be reviewed. This includes papers where the margins and formatting are deemed to have been significantly altered from those laid down by this style guide. Note that this LaTeX guide already sets figure captions and references in a smaller font. The reason such papers will not be reviewed is that there is no provision for supervised revisions of manuscripts. The reviewing process cannot determine the suitability of the paper for presentation in eight pages if it is reviewed in eleven.

2.7. The ruler

The LaTeX style defines a printed ruler which should be present in the version submitted for review. The ruler is provided in order that reviewers may comment on particular lines in the paper without circumlocution. If you are preparing a document using a non-LaTeX document preparation system, please arrange for an equivalent ruler to appear on the final output pages. The presence or absence of the ruler should not change the appearance of any other content on the page. The camera ready copy should not contain a ruler. (LaTeX users may uncomment the `\cvprfinalcopy` command in the document preamble.) Reviewers: note that the ruler measurements do not align well with lines in the paper — this turns out to be very difficult to do well when the paper contains many figures and equations, and, when done, looks ugly. Just use fractional references (e.g. this line is 095.5), although in most cases one would expect that the approximate location will be adequate.

2.8. Mathematics

Please number all of your sections and displayed equations. It is important for readers to be able to refer to any particular equation. Just because you didn't refer to it in the text doesn't mean some future reader might not need to refer to it. It is cumbersome to have to use circumlocutions like "the equation second from the top of page 3 column 1". (Note that the ruler will not be present in the final copy, so is not an alternative to equation numbers). All authors will benefit from reading Mermin's description of how to write mathematics: `http://www.pamitc.org/documents/mermin.pdf`.

2.9. Blind review

Many authors misunderstand the concept of anonymizing for blind review. Blind review does not mean that one must remove citations to one's own work—in fact it is often impossible to review a paper unless the previous citations are known and available.

Blind review means that you do not use the words "my" or "our" when citing previous work. That is all. (But see below for techreports.)

Saying "this builds on the work of Lucy Smith [1]" does not say that you are Lucy Smith; it says that you are building on her work. If you are Smith and Jones, do not say "as we show in [7]", say "as Smith and Jones show in [7]" and at the end of the paper, include reference 7 as you would any other cited work.

An example of a bad paper just asking to be rejected:

> An analysis of the frobnicatable foo filter.
>
> In this paper we present a performance analysis of our previous paper [1], and show it to be inferior to all previously known methods. Why the previous paper was accepted without this analysis is beyond me.
>
> [1] Removed for blind review

An example of an acceptable paper:

> An analysis of the frobnicatable foo filter.
>
> In this paper we present a performance analysis of the paper of Smith *et al.* [1], and show it to be inferior to all previously known methods. Why the previous paper was accepted without this analysis is beyond me.
>
> [1] Smith, L and Jones, C. "The frobnicatable foo filter, a fundamental contribution to human knowledge". Nature 381(12), 1-213.

If you are making a submission to another conference at the same time, which covers similar or overlapping material, you may need to refer to that submission in order to explain the differences, just as you would if you had previously published related work. In such cases, include the anonymized parallel submission as additional material and cite it as

> [1] Authors. "The frobnicatable foo filter", F&G 2014 Submission ID 324, Supplied as additional material `fg324.pdf`.

Finally, you may feel you need to tell the reader that more details can be found elsewhere, and refer them to a technical report. For conference submissions, the paper must stand on its own, and not *require* the reviewer to go to a techreport for further details. Thus, you may say in the body of the paper "further details may be found in". Then submit the techreport as additional material. Again, you may not assume the reviewers will read this material.

Sometimes your paper is about a problem which you tested using a tool which is widely known to be restricted to a single institution. For example, let's say it's 1969, you have solved a key problem on the Apollo lander, and you believe that the CVPR70 audience would like to hear about your solution. The work is a development of your celebrated 1968 paper entitled "Zero-g frobnication: How being the only people in the world with access to the Apollo lander source code makes us a wow at parties", by Zeus *et al.*

You can handle this paper like any other. Don't write "We show how to improve our previous work [Anonymous, 1968]. This time we tested the algorithm on a lunar lander [name of lander removed for blind review]". That would be silly, and would immediately identify the authors. Instead write the following:

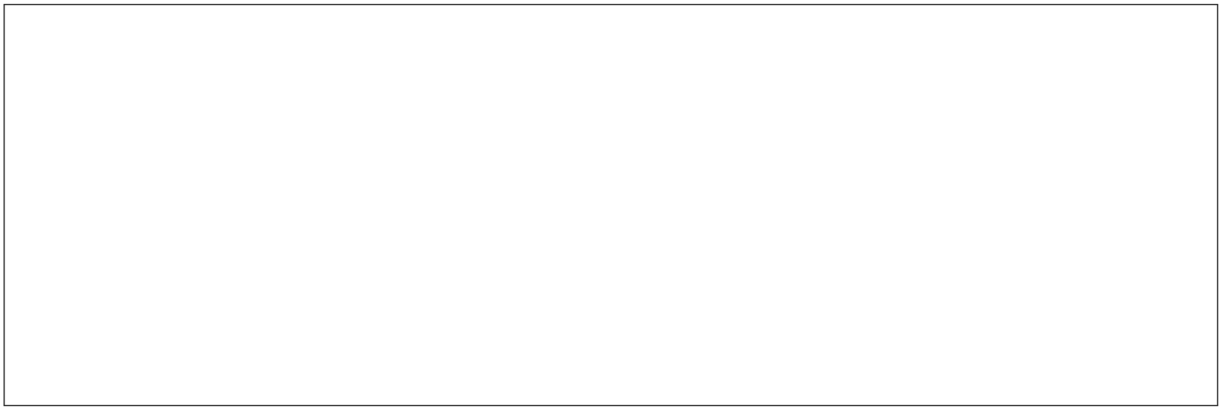

Figure 1. Example of caption. It is set in Roman so that mathematics (always set in Roman: $B \sin A = A \sin B$) may be included without an ugly clash.

> We describe a system for zero-g frobnication. This system is new because it handles the following cases: A, B. Previous systems [Zeus et al. 1968] didn't handle case B properly. Ours handles it by including a foo term in the bar integral.
>
> ...
>
> The proposed system was integrated with the Apollo lunar lander, and went all the way to the moon, don't you know. It displayed the following behaviours which show how well we solved cases A and B: ...

As you can see, the above text follows standard scientific convention, reads better than the first version, and does not explicitly name you as the authors. A reviewer might think it likely that the new paper was written by Zeus *et al.*, but cannot make any decision based on that guess. He or she would have to be sure that no other authors could have been contracted to solve problem B.

FAQ: Are acknowledgements OK? No. Leave them for the final copy.

2.10. Miscellaneous

Compare the following:

`$conf_a$`	$conf_a$
`$\mathit{conf}_a$`	conf_a

See The TeXbook, p165.

The space after *e.g.*, meaning "for example", should not be a sentence-ending space. So *e.g.* is correct, *e.g.* is not. The provided `\eg` macro takes care of this.

When citing a multi-author paper, you may save space by using "et alia", shortened to "*et al.*" (not "*et. al.*" as "*et*" is a complete word.) However, use it only when there are three or more authors. Thus, the following is correct: " Frobnication has been trendy lately. It was introduced by Alpher [2], and subsequently developed by Alpher and Fotheringham-Smythe [3], and Alpher *et al.* [4]."

This is incorrect: "... subsequently developed by Alpher *et al.* [3] ..." because reference [3] has just two authors. If you use the `\etal` macro provided, then you need not worry about double periods when used at the end of a sentence as in Alpher *et al.*

For this citation style, keep multiple citations in numerical (not chronological) order, so prefer [3, 2] to [2, 3].

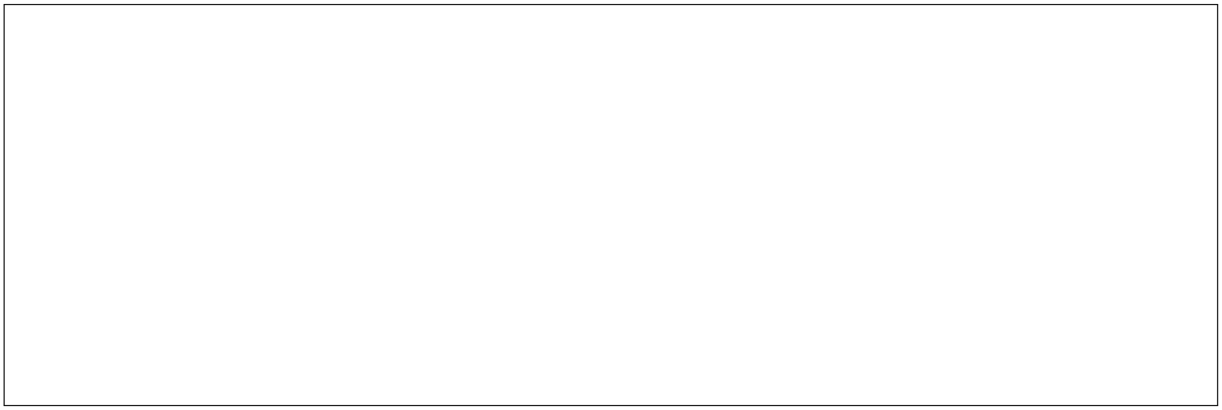

Figure 2. Example of a short caption, which should be centered.

3. Formatting your paper

All text must be in a two-column format. The total allowable width of the text area is $6\frac{7}{8}$ inches (17.5 cm) wide by $8\frac{7}{8}$ inches (22.54 cm) high. Columns are to be $3\frac{1}{4}$ inches (8.25 cm) wide, with a $\frac{5}{16}$ inch (0.8 cm) space between them. The main title (on the first page) should begin 1.0 inch (2.54 cm) from the top edge of the page. The second and following pages should begin 1.0 inch (2.54 cm) from the top edge. On all pages, the bottom margin should be 1-1/8 inches (2.86 cm) from the bottom edge of the page for 8.5×11-inch paper; for A4 paper, approximately 1-5/8 inches (4.13 cm) from the bottom edge of the page.

3.1. Margins and page numbering

All printed material, including text, illustrations, and charts, must be kept within a print area 6-7/8 inches (17.5 cm) wide by 8-7/8 inches (22.54 cm) high. Page numbers should be in footer with page numbers, centered and .75 inches from the bottom of the page and make it start at the correct page number rather than the 4321 in the example. To do this fine the line (around line 23)

```
%\ifcvprfinal\pagestyle{empty}\fi
\setcounter{page}{4321}
```

where the number 4321 is your assigned starting page.

Make sure the first page is numbered by commenting out the first page being empty on line 46

```
%\thispagestyle{empty}
```

3.2. Type-style and fonts

Wherever Times is specified, Times Roman may also be used. If neither is available on your word processor, please use the font closest in appearance to Times to which you have access.

MAIN TITLE. Center the title 1-3/8 inches (3.49 cm) from the top edge of the first page. The title should be in Times 14-point, boldface type. Capitalize the first letter of nouns, pronouns, verbs, adjectives, and adverbs; do not capitalize articles, coordinate conjunctions, or prepositions (unless the title begins with such a word). Leave two blank lines after the title.

AUTHOR NAME(s) and AFFILIATION(s) are to be centered beneath the title and printed in Times 12-point, non-boldface type. This information is to be followed by two blank lines.

Method	Frobnability
Theirs	Frumpy
Yours	Frobbly
Ours	Makes one's heart Frob

Table 1. Results. Ours is better.

The ABSTRACT and MAIN TEXT are to be in a two-column format.

MAIN TEXT. Type main text in 10-point Times, single-spaced. Do NOT use double-spacing. All paragraphs should be indented 1 pica (approx. 1/6 inch or 0.422 cm). Make sure your text is fully justified—that is, flush left and flush right. Please do not place any additional blank lines between paragraphs.

Figure and table captions should be 9-point Roman type as in Figures 1 and 2. Short captions should be centred.

Callouts should be 9-point Helvetica, non-boldface type. Initially capitalize only the first word of section titles and first-, second-, and third-order headings.

FIRST-ORDER HEADINGS. (For example, **1. Introduction**) should be Times 12-point boldface, initially capitalized, flush left, with one blank line before, and one blank line after.

SECOND-ORDER HEADINGS. (For example, **1.1. Database elements**) should be Times 11-point boldface, initially capitalized, flush left, with one blank line before, and one after. If you require a third-order heading (we discourage it), use 10-point Times, boldface, initially capitalized, flush left, preceded by one blank line, followed by a period and your text on the same line.

3.3. Footnotes

Please use footnotes[1] sparingly. Indeed, try to avoid footnotes altogether and include necessary peripheral observations in the text (within parentheses, if you prefer, as in this sentence). If you wish to use a footnote, place it at the bottom of the column on the page on which it is referenced. Use Times 8-point type, single-spaced.

3.4. References

List and number all bibliographical references in 9-point Times, single-spaced, at the end of your paper. When referenced in the text, enclose the citation number in square brackets, for example. Where appropriate, include the name(s) of editors of referenced books.

3.5. Illustrations, graphs, and photographs

All graphics should be centered. Please ensure that any point you wish to make is resolvable in a printed copy of the paper. Resize fonts in figures to match the font in the body text, and choose line widths which render effectively in print. Many readers (and reviewers), even of an electronic copy, will choose to print your paper in order to read it. You cannot insist that they do otherwise, and therefore must not assume that they can zoom in to see tiny details on a graphic.

When placing figures in LaTeX, it's almost always best to use `\includegraphics`, and to specify the figure width as a multiple of the line width as in the example below

[1] This is what a footnote looks like. It often distracts the reader from the main flow of the argument.

```
\usepackage[dvips]{graphicx} ...
\includegraphics[width=0.8\linewidth]
                 {myfile.eps}
```

3.6. Color

Please refer to the author guidelines on the CVPR 2017 web page for a discussion of the use of color in your document.

4. Final copy

You must include your signed IEEE copyright release form when you submit your finished paper. We MUST have this form before your paper can be published in the proceedings.

References

[1] C. Adam-Bourdarios, G. Cowan, C. Germain, I. Guyon, B. Kgl, and D. Rousseau. The Higgs boson machine learning challenge. In G. Cowan, C. Germain, I. Guyon, B. Kgl, and D. Rousseau, editors, *Proceedings of the NIPS 2014 Workshop on High-energy Physics and Machine Learning*, volume 42 of *Proceedings of Machine Learning Research*, pages 19–55, Montreal, Canada, 13 Dec 2015. PMLR.

[2] A. Alpher. Frobnication. *Journal of Foo*, 12(1):234–778, 2002.

[3] A. Alpher and J. P. N. Fotheringham-Smythe. Frobnication revisited. *Journal of Foo*, 13(1):234–778, 2003.

[4] A. Alpher, J. P. N. Fotheringham-Smythe, and G. Gamow. Can a machine frobnicate? *Journal of Foo*, 14(1):234–778, 2004.

[5] P. Baldi, P. Sadowski, and D. Whiteson. Searching for exotic particles in high-energy physics with deep learning. *Nature Commun.*, 5:4308, 2014.

[6] P. Baldi, P. Sadowski, and D. Whiteson. Enhanced higgs boson to $\tau^+\tau^-$ search with deep learning. *Phys. Rev. Lett.*, 114(11):111801, 2015.

[7] X. Glorot, A. Bordes, and Y. Bengio. Deep sparse rectifier neural networks. In G. J. Gordon and D. B. Dunson, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS-11)*, volume 15, pages 315–323. Journal of Machine Learning Research - Workshop and Conference Proceedings, 2011.

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

[9] A. Krogh and J. A. Hertz. A simple weight decay can improve generalization. In *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 4*, pages 950–957. Morgan Kaufmann, 1992.

[10] J. Snoek, H. Larochelle, and R. P. Adams. Practical bayesian optimization of machine learning algorithms. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*, NIPS'12, pages 2951–2959, USA, 2012. Curran Associates Inc.

[11] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958, Jan. 2014.

[12] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.*, 11:3371–3408, Dec. 2010.