Multi-task Convolution Neural Networks for the Chips Experiment

Josh Chalcraft Tingey
of University College London

A dissertation submitted to University College London
for the degree of Doctor of Philosophy

Abstract

LHCb is a b-physics detector experiment which will take data at the 14 TeV LHC accelerator at CERN from 2007 onward…

Declaration

This dissertation is the result of my own work, except where explicit reference is made to the work of others, and has not been submitted for another qualification to this or any other university. This dissertation does not exceed the word limit for the respective Degree Committee.

Andy Buckley

Acknowledgements

Of the many people who deserve thanks, some are particularly prominent, such as my supervisor…

Preface

This thesis describes my research on various aspects of the LHCb particle physics program, centred around the LHCb detector and LHC accelerator at CERN in Geneva.

For this example, I'll just mention Chapter **??** and Chapter 1.

Contents

“Writing in English is the most ingenious torture ever devised for sins committed in previous lives.”

— James Joyce

Chapter 1.

Neutrino oscillations: theoretical background and current status

Consider a simple two body decay Neutrino physics covers the widest possible range of Proposal of a mysterious undetector particle to explain beta decays in the 1930s through to the resolutions of a 30-year problem with the confirmation of oscillations in the early 2000s and onto the precision era. Neutrino oscillations first discoveed in 1957 when Bruno Pornecorvo proposed a model in which neutrinos oscilate to antineutrinos and back, similar to the kain. It was actually shown that neutrinos iscilate from one flavour to another. The field of neutrino physics is ever expanding with a new generation of experiments planned for the coming years. This chapter aims to provide an introduction to neutrino

1.1. A history of neutrino oscillations

In the early 20th century, beta decays were assumed to follow the simple two-body process, $A \rightarrow B + e$, where a nuclei spontaneously an electron, and only an electron. To conserve both energy and angular momentum the ejected electron must have a discrete kinetic energy defined by the different in binding energies between the initial and final nuclei. However, in 1914, J. Chadwick instead measured a continuous energy spectrum for the electron [1], placing this theory in doubt.

W. Pauli finally proposed a 'desperate solution' to this paradox in 1930 [2]. If a light, neutrally charged, spin 1/2 particle was also produced in the interaction, the continuous energy distribution could be explained. Initially this mysterious new particle was named the 'neutron'. But, to avoid confusion with the heavy baryon of the same name discovered in 1932, E. Fermi renamed it the 'neutrino' when he formalised beta decay in 1934 [3].

The following month, H. Bethe and R. Peierls used Fermi's work to estimate the cross-section of the inverse beta decay process $\nu + p^+ \rightarrow n + e^+$ [4]. They calculated a value of less than the very small $10^{-44} cm^2$ and declared 'there is no practically possible way of observing the neutrino.' Although extensive neutrino detection has proved possible, it hinted at the huge difficulties experimentalists would face hunting down the neutrino and measuring it's properties in the years to come.

After an initial tentative identification if 1953, F. Reines and C. Cowan made the first confirmed observation of the neutrino in 1956 [5]. Electron antineutrinos produced in the Savannah River Plan nuclear reactor were detected via the inverse decay process outlined in the previous paragraph. A 'club-sandwich' detector of three 1500 litre liquid scintillator tanks and two 200 litre cadmium doped water target tanks, was constructed in an underground room of the reactor building. A total of 330 photomultiplier tubes were then able to measure the prompt positron annihilation signal followed by the gamma ray burst from the neutron capture in cadmium, the signature identification for the interaction.

Brookhaven two kinds of neutrinos in Ref. [6] - Muon neutrino discovered by the 'long track' from the decays of pions from a reactor in 1988, got a nobel prize. - In 1962 at the alternating gradient synchrotron at Brookhaven, neutrinos created from pion decays together with muons were observed t produce only muons not electrons, this then confirmed the existence of the muon neutrino. - Neutrinos originating from pion decays primary produce muons, not electron. Detected as single long tracks in a spark chamber. - Got the 1988 nobel prize for this discovery of the second neutrino. - Distinct from the previously known electron neutrino - The neutrinos produced by the pions decay from a accelerator beam, were not the same as the neutrinos observed in beta decay - Did so by observing that it was far more likely that the neutrinos pro-duced in the decay of pions would interact to create muons, as opposed to electron - First experiment to construct and use an artificial neutrino beam - 34 identified muon events in total no electrons.

Discovery of the tau lepton [7] Also precise z-resonance measurement at lep in the 1990's [8] Finally measured by DONUT in 2000 [9] - evidence for the third neutrino, finally discovered at DONUT in 2000 - After the discovery of the tau lepton in 1975 [7], this suggested the existence of the third neutrino which DONUT found in 2000. - DONUT finally found the tau neutrino in 2000 using 800GeV protons from the Tevatron.

- In 2000 the DONUT experiment at the tevatron collider in fermilab performned a direct detection of the tau neutrino completing the three flavour picture. - Experiments at the LEP e+e- collider in the 1990s made precision measurmnets of the z decay width, from a fir to the data in showed there are exactly three active generations of neutrinos. - This indicates the number of active neutrino states can only be 1.984+-0.008. Therefore, any as yet undiscovered neutrinos must be sterile, in that they do not couple to the weak interaction.

Homestake deficit observation in Ref. [10] first SSU predictions used to compare against homestake in Ref. [11] Kamiokande II deficit in Ref. [12] SAGE experiment deficit in Ref. [13] GALLEX experiment deficit in Ref. [14] SSM Prediction for Ga in Ref. [15] - As the standard model of particle physics was developed, neutrinos were presumed to be massless and occur only in the three flavour eigenstates. - Various hints that this was not the case kept appearing, leading to neutrino oscillations, by witch one neutrino can oscillate to another flavour and the non-zero masses that follow as a direct consiquence from this. - In the solar neutrino sector there is the "solar anomaly" noting a deficit of electron neutrino compared to predications made by the standard solar model (SSM) - First observed at the Homestack experiment, neutrinos ineracted with the chlorine creating radioactive argon atoms, beause it is a noble gas it does not bing to the perchloroethylene and it can be extracted by purging the liquid with gaseous helium and then extracted from the helium with a cooled carbon trap. - Gallium was also used by other experiments and kamiokande also observed the deficit. - Also the fluxes measured where not consistent, depending on the energy range probed. Hinting at oscillations dependent on energy, - 400 000 litresof perchloroethylene (a dry-cleaning fluid), containing520 tof chlorine, placed in the Homestake Mine,1.5 kmunderground [24]. - he reported experimental rate was about two thirds less than what was expected from theStandard Solar Model (SSM). This large discrepancy, known as thesolar neutrino problem, wasinitially believed to be an experimental flaw. - This is where the future DUNE detector will be housed, nice full circle - This is in the solar sector

SNO oscillation measurement in Ref. [16] - neutrino oscillations were one way of explaining this deficit if some of the electron neutrinos converted flavour in flight. - SNO finally answered the question when it was able to measure three channels with different relation between te flyx or electron neutrinos and the other neutrinos. SNO could prove that the electron neutrinos are changing flavour. WHile the total flux of all neutrinos remains constant and in agrrement with the SSM. - 1kton tank of heavy (D2O deuterium) water, able to detect three different channels of neutrino interaction -

Cherenkov experiment, with 9500 8inch photomultiplier tubes detectro the light from neutrino interactions. - Since each of the rates for the three channels has a different relation between the flux of electron neutrinos and the others, SNO could confirm electron neutrinos are changing flavour, with the total flux being constant and in agreement with the SSM. - electron neutrino CC, NC and elastic scattering also. - total rate was consistent but less electron neutrinos than expected as they had oscillated. - However, only electron neutrinos canundergo CC interactions, as solar neutrinos do not have enough energy to produce muonor tau leptons. -

Atmospheric kamiokande deficit in Ref. [17] IMD detector atmospheric deficit in Ref. [18] Superkamiokande direction atmospheric neutrinos in Ref. [18] - This is in the atmospheric sector

1.2. Neutrino oscillation theory

blah blah blah

1.3. Current status and the future

1.3.1. Atmospheric

1.3.2. Accelerator experiments

1.3.3. Reactor experiments

1.3.4. The future

blah blah blah

Chapter 2.

A convolutional neural network for CHIPS

2.1. Deep neural network history and theory

Neuron equation

$$z^{(i)} = \boldsymbol{w}^{(i)} \cdot \boldsymbol{x} + b^{(i)} \tag{2.1}$$

With activation function applied

$$a_i(\boldsymbol{x}) = \sigma_i(z^{(i)}) \tag{2.2}$$

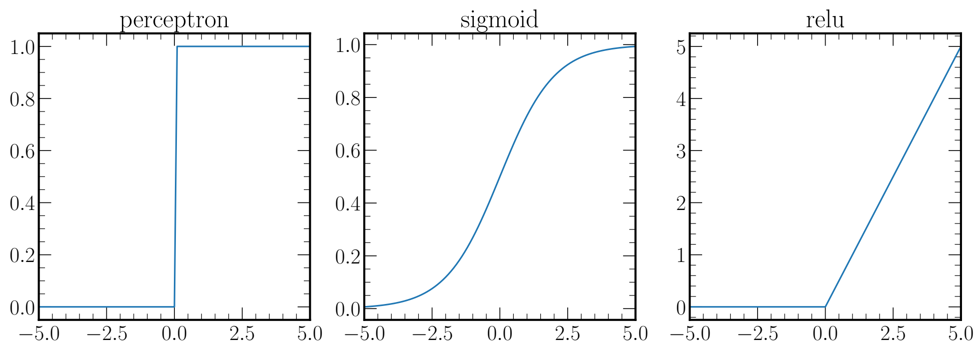

Figure 2.1.: CKM Fitter constraints on α from combined $\mathrm{B} \to \pi\pi$, $\mathrm{B} \to \rho\pi$ and $\mathrm{B} \to \rho\rho$ decay analyses.

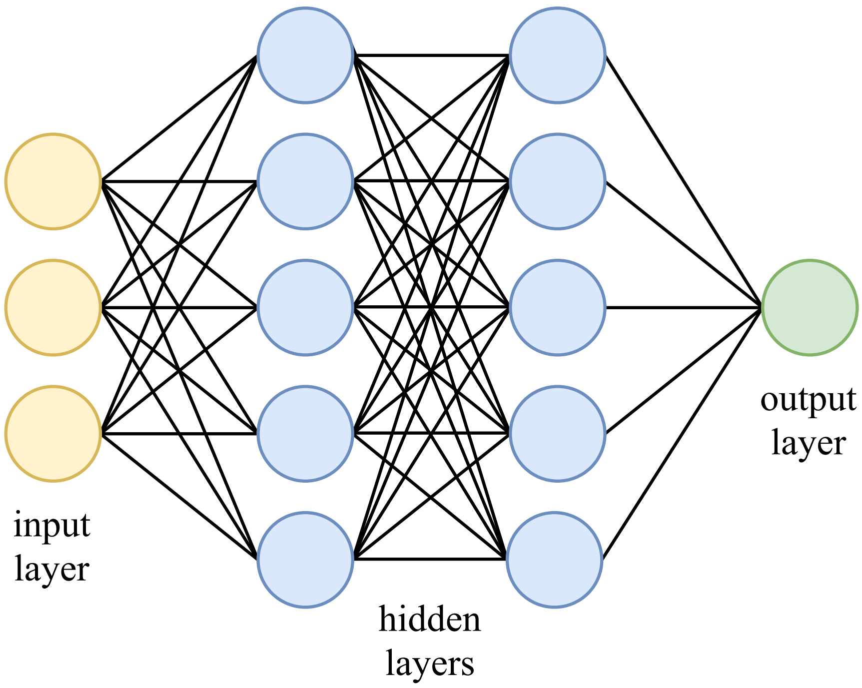

Figure 2.2.: Classic neural network architecture with neurons arrange in layers. The output of one layer acts as the input to the next layer until the output is reached.

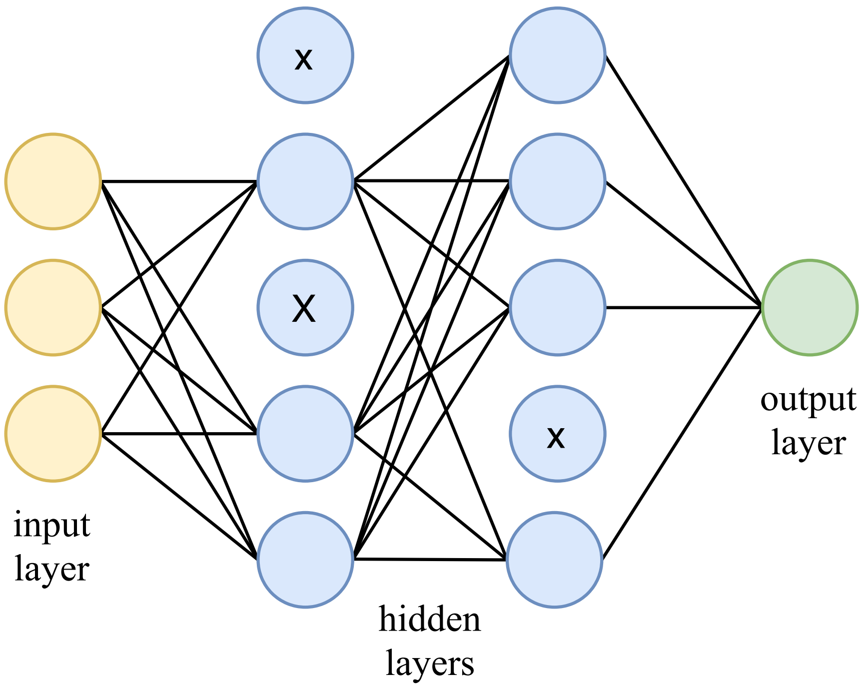

Figure 2.3.: CKM Fitter constraints on α from combined $B \to \pi\pi$, $B \to \rho\pi$ and $B \to \rho\rho$ decay analyses.

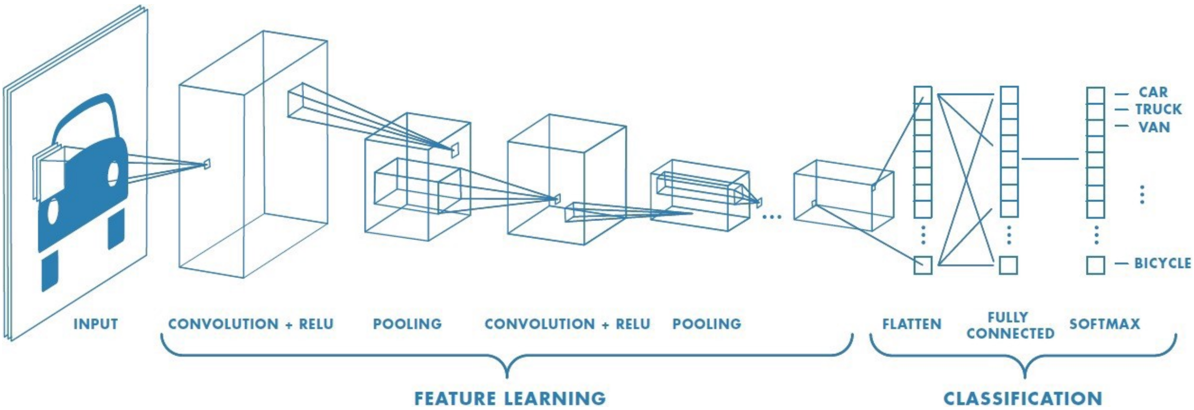

Figure 2.4.: CKM Fitter constraints on α from combined $\mathrm{B} \to \pi\pi$, $\mathrm{B} \to \rho\pi$ and $\mathrm{B} \to \rho\rho$ decay analyses.

MSE

$$E(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i(\boldsymbol{w}))^2 \tag{2.3}$$

Binary cross-entropy

$$E(\boldsymbol{w}) = -\sum_{i=1}^{n} y_i \log \hat{y}_i(\boldsymbol{w}) + (1 - y_i)\log[1 - \hat{y}_i(\boldsymbol{w})] \tag{2.4}$$

categorical cross-entropy

$$E(\boldsymbol{w}) = -\sum_{i=1}^{n}\sum_{m=0}^{M-1} y_{im} \log \hat{y}_{im}(\boldsymbol{w}) + (1 - y_{im})\log[1 - \hat{y}_{im}(\boldsymbol{w})] \tag{2.5}$$

2.2. Applications to HEP problems

2.3. Story of the chapter

There are many standard ways that event classification and paramter (mainly energy) estimation are usually done in HEP. These mainly include the reconstruction of objects and their associated parameters wether these are clusters, tracks, jets or cherenkov rings. Typically these along with other constructed features are then passed through a simple

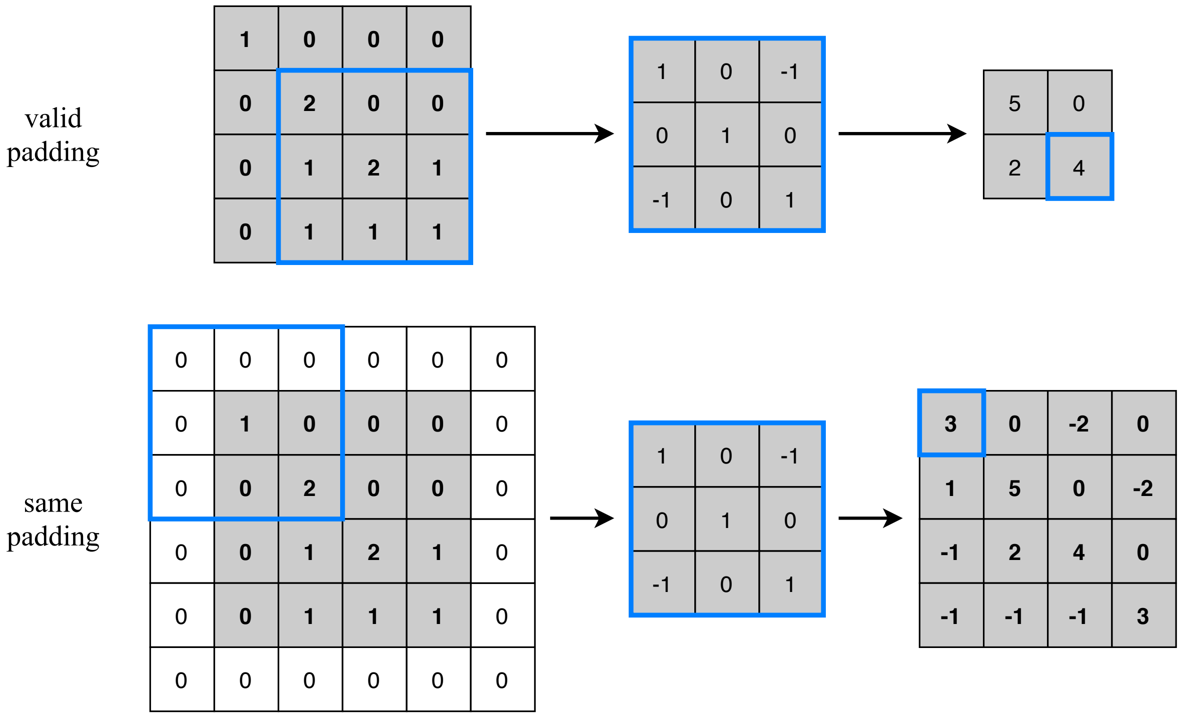

Figure 2.5.: CKM Fitter constraints on α from combined $\mathrm{B} \to \pi\pi$, $\mathrm{B} \to \rho\pi$ and $\mathrm{B} \to \rho\rho$ decay analyses.

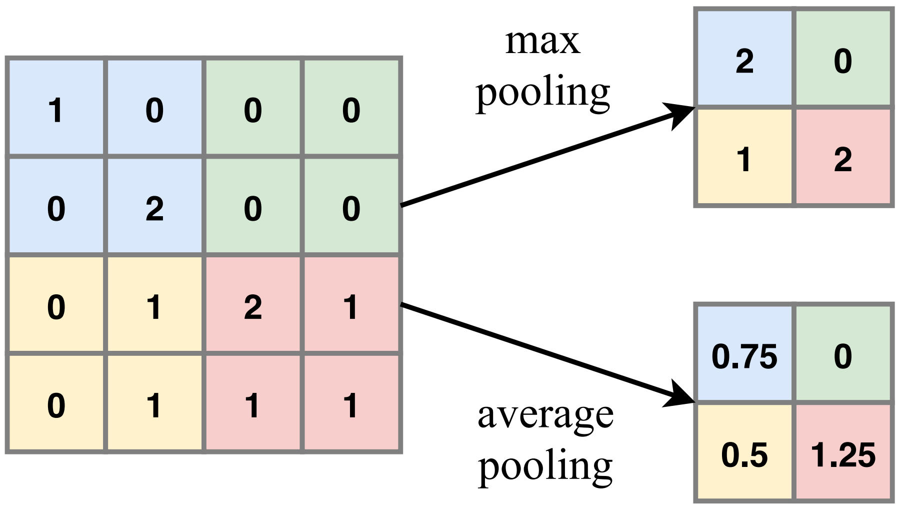

Figure 2.6.: CKM Fitter constraints on α from combined $\mathrm{B} \to \pi\pi$, $\mathrm{B} \to \rho\pi$ and $\mathrm{B} \to \rho\rho$ decay analyses.

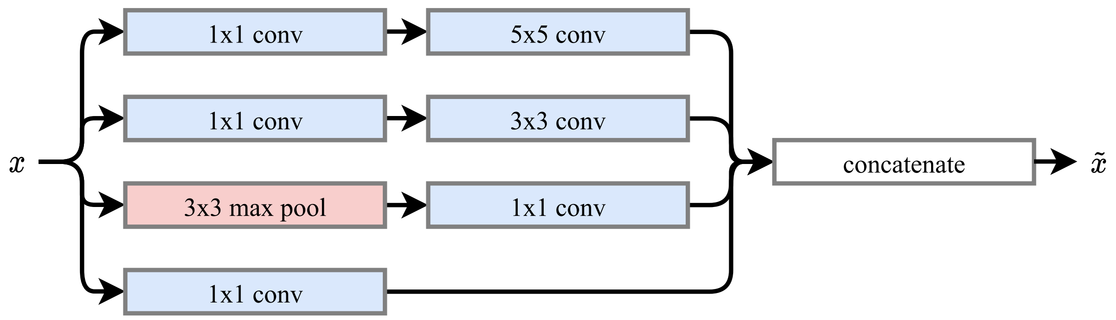

Figure 2.7.: CKM Fitter constraints on α from combined $B \to \pi\pi$, $B \to \rho\pi$ and $B \to \rho\rho$ decay analyses.

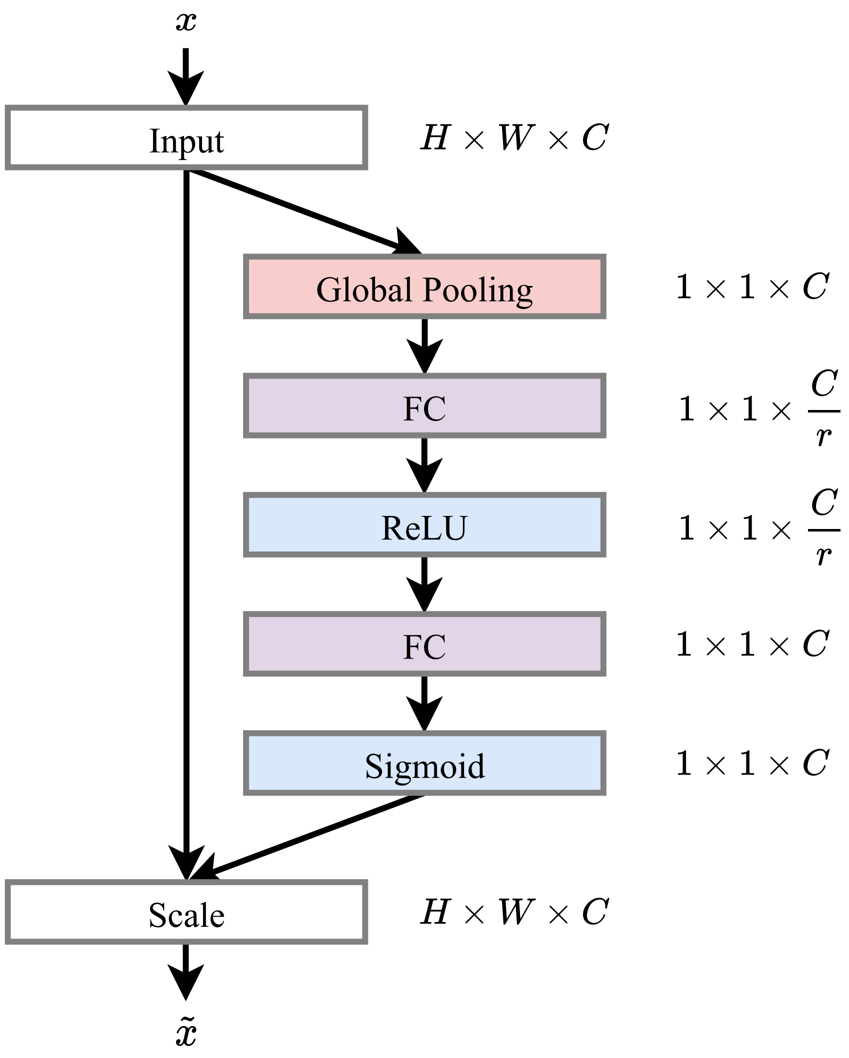

Figure 2.8.: CKM Fitter constraints on α from combined $B \to \pi\pi$, $B \to \rho\pi$ and $B \to \rho\rho$ decay analyses.

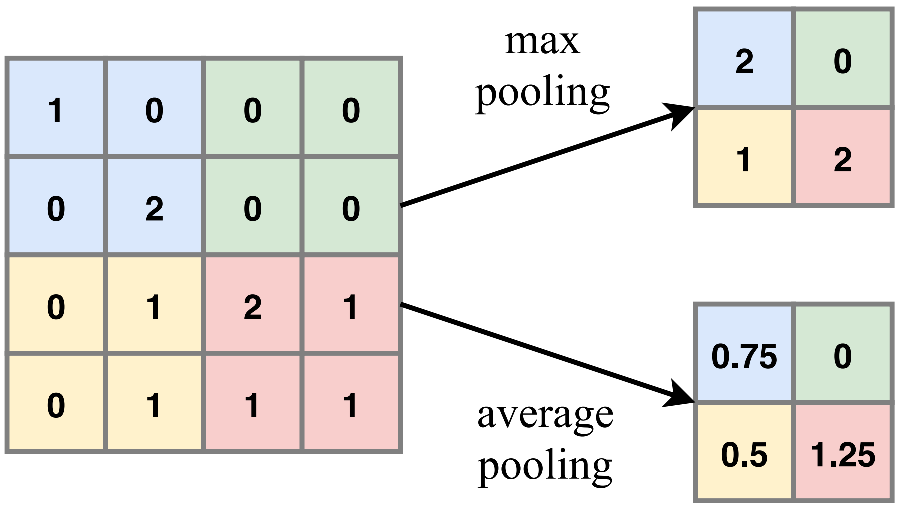

Figure 2.9.: CKM Fitter constraints on α from combined $\mathrm{B} \to \pi\pi$, $\mathrm{B} \to \rho\pi$ and $\mathrm{B} \to \rho\rho$ decay analyses.

machine learning model for event or particle classification, or combined in some other way for energy estimation.

While this has worked leveraging the enormous amount of work in machine learning especially deep learning surely would prove valuable. The key thing that has been done is to move away from human engineered features to machine learning models that discover the underlying features that work best in clasifiying of regressing a particualar task.

Water chernkov detectors are especially paired to this task as the output from our detectors is essentially an 'image' of the event and so classification models that work well on images shoud work well on seperating our types of events.

Firstly, the principle issue with matching water cherenkov detectors and deep learning is representing the cylindrical detector output of either a 2d flat map that a typically conv network can use as input of use a more complicated graph network approach as some other people have tried. Typically people have just ignored the endcaps, but this is not optimal as these contain nearing half the detected light in a standard CHIPS detector. Other appraoches include the x+ x- approach. But as hited at by tomothy report, for a primarilly event classification task, removing any distritions due to the detector shape is the most important things. Therefore, the approach of veiwing the "image" from the interaction vertex point preserves the cherenkov ring strucutre as best possible.

The hough space is used to find this vertex and direction from which the images are produces, so it also made sense to include this along with time as a seperate channel

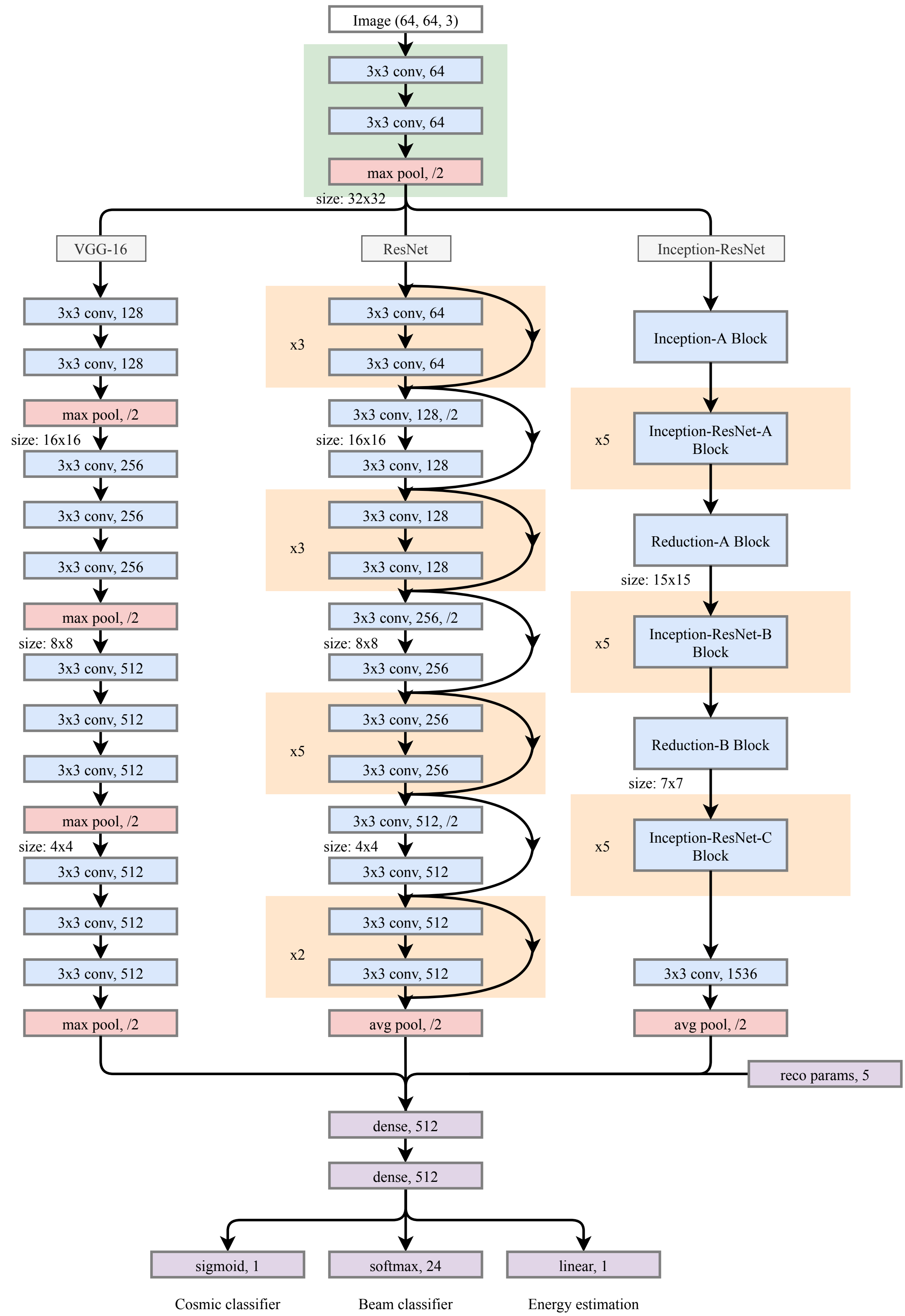

Figure 2.10.: CKM Fitter constraints on α from combined $B \to \pi\pi$, $B \to \rho\pi$ and $B \to \rho\rho$ decay analyses.

in the output. You can see fro the vertext position is best. Using a more unformly distributes sample of events leads to a greater abiity to distinguish the types which is important for subsequant energy reconstruction.

Many possible models have been developed for convolutional neural networks over th years.

Firstly, the principle issue with matching water cherenkov detectors and deep learning is representing the cylindrical detector output of either a 2d flat map that a typically conv network can use as input of use a more complicated graph network approach as some other people have tried. Typically people have just ignored the endcaps, but this is not optimal as these contain nearing half the detected light in a standard CHIPS detector. Other appraoches include the x+ x- approach. But as hited at by tomothy report, for a primarilly event classification task, removing any distritions due to the detector shape is the most important things. Therefore, the approach of veiwing the "image" from the interaction vertex point preserves the cherenkov ring strucutre as best possible.

2.4. Intro, General theory and previous work

- Standard water cherenkov analysis is via a likelihood hood fit to the ring assuming some event topology hypothesis. This is used in super-k with fitqun and what has been previously implemented for CHIPS in the WCSimAnalysis package. - Previous work in HEP has applied deep learning to a variety of problems... - A MPL with a single hidden layer can be shown to apprximate any function arbitrarily accurately. Give REF for this. - Conv is a set of 'filters' that when applied via scanning across an input image result in a feature map - Pooling used as each layer requires less complexity and it is less important about the location and that the feature exists.

2.5. A CVN for Chips

- Approaches in the past for event classification using CNNs for water cherenkov detectors have taken a few Approaches to generating the input image representation. - Projecting onto a 2d surface "outside" the detector REF: Cern summer report in Ref. [19]

2.6. Things to talk about

- Convolutional neural network theory and history - Recent uses in experiments - CHIPS implementation tellig the story - Explainability (ablation, clustering, layer weights etc...)

2.7. Diagrams

- Example event hit maps, showing the different possible channels for different types of event. - True neutrino energy distributions for different categories for energy estimation. - Network architecture plots. - Neutrino energy and estimated neutrino energy distributions on same plots. - reco-true/true neutrino energy distributions for the different event types, QEL, DIS, RES etc... - Plot comparing lepton energy reconstruction between old reco and new estimation, just for nice CC events. - Individual block diagrams for the different models I try + a squeeze-exitation diagram - Arbitrary vs 8-bit precision for all the channels - Example activation maps for a few events (explain) - Number of events in each category within the training sample - Training loss and accuracy vs epoch/iteration - Classifier output plots - Table of the final number of expected events and efficiency and purity of the signal at the chosen cut value (nuel and numu) - t-SNE and PCA plots, nicely coloured - Example events from the t-SNE plot, showing the different types - Confusion matrices - Parrallel coordinates plots for tuning the hyperparameters - Time taken comparison with old reconstruction (just inference time for all stages) - Final energy distribution of selected events given a number of years running

2.8. References

Cern summer report in Ref. [19] CHIPS cosmic rate in Ref. [20] Nova first CVN paper in Ref. [21] - CNN's have been widely applied in various computer vision tasks to solve image recongnition and analysis problems. - The core problem in HEP is the correct categorisation of particle interactions - This is usually done by reconstructing high-level componenets suh as clusters, tracks, showers, jets and rings. and then summarising these objects energies, directions and shapes, these wuantities are then fed into k-nearest neighbours, BDTs or MLPs to seperate sign from bkg. - Prone to failure, mistakes in the reconstruction, and limitation to what has been implemented by humans. - computer vision moved away from specifically constructed features to sing ML CNNs to discover the features. - Manu HEP problems including water cherenkov detectors essentially result in an 'image' of an event, which are well suited to these tools. - MLP are widely used in HEP, Nova context enriched CVN paper in Ref. [22] Nova energy recontruction CVN in Ref. [23] Watchmal/Triumf Cherenkov variational autoencoders in Ref. [24] Daya bay paper in Ref. [25] SHiP GAN simulation paper in Ref. [26] New ideas with x+ x- mapping in Ref. [27] DUNE TDR in Ref. [28] Initial CNN visualisation paper in Ref. [29] Original t-SNE paper in Ref. [30] Original 'dropout' paper in Ref. [31] VGG paper in Ref. [32] Improved resnet paper in Ref. [33] Inception-resnet paper in Ref. [34] Squeeze-and-excitation networks paper in Ref. [35] MobileNetV2 paper in Ref. [36] Multi-task learning how to weight paper in Ref. [37] Grad-CAM paper in Ref. [38] Amazing machine learning for physicists thing in Ref. [39] - Deep neural networks have emerged as one of the most powerful supervised learning techniques. - They truly caught the attention of the wider ML communityr in 2012 when A. Krizhevsky, I, Sutskever and G. Hinton used a GPU to train AlexNet, lowering the error rate on the image classification task ImageNet by 12- Such was the rapid pace of advance afterwards that the ResNet model acheived a 3.57% error just three years later. - Many high level libraries have now been formed, predominently led by Tensorflow (from google) and pyTorch (from facebook) making it easier to quickly code and implemenetd DNNs. - Neural networks are neural-inspired nonlinear models for supervised learning. Constructed from the basic building blocks of a "neuron". -

2.9. Reference notes

Appendix A.

Pointless extras

"Le savant n'étudie pas la nature parce que cela est utile;
il l'étudie parce qu'il y prend plaisir,
et il y prend plaisir parce qu'elle est belle."
— Henri Poincaré, 1854–1912

Appendixes (or should that be "appendices"?) make you look really clever, 'cos it's like you had more clever stuff to say than could be fitted into the main bit of your thesis. Yeah. So everyone should have at least three of them...

A.1. Like, duh

Padding? What do you mean?

A.2. $y = \alpha x^2$

See, maths in titles automatically goes bold where it should (and check the table of contents: it *isn't* bold there!) Check the source: nothing needs to be specified to make this work. Thanks to Donald Arsenau for the teeny hack that makes this work.

Colophon

This thesis was made in LaTeX 2_ε using the "hepthesis" class [40].

Bibliography

[1] J. Chadwick, Verh. Phys. Gesell. **1**6, 383 (1914).

[2] W. Pauli, Letter to the physical society of Tübingen (1930).

[3] E. Fermi, Z. Phys. **8**8, 161 (1934).

[4] H. Bethe and R. Peierls, Nature **1**33, 532 (1934).

[5] C. Cowan Jr, Detection of the free neutrino: a confirmation , 103 (1956).

[6] G. Danby *et al.*, Physical Review Letters **9**, 36 (1962).

[7] M. L. Perl *et al.*, Physical Review Letters **3**5, 1489 (1975).

[8] T. S. Electroweak *et al.*, Physics Reports **4**27, 257 (2006).

[9] K. Kodama *et al.*, Physics Letters B **5**04, 218–224 (2001).

[10] R. Davis Jr, D. S. Harmer, and K. C. Hoffman, Physical Review Letters **2**0, 1205 (1968).

[11] J. N. Bahcall, N. A. Bahcall, and G. Shaviv, Physical Review Letters **2**0, 1209 (1968).

[12] K. S. Hirata *et al.*, Physical Review Letters **6**3, 16 (1989).

[13] A. Abazov *et al.*, Physical review letters **6**7, 3332 (1991).

[14] P. Anselmann *et al.*, Physics Letters B **3**27, 377 (1994).

[15] J. N. Bahcall and R. K. Ulrich, Reviews of Modern Physics **6**0, 297 (1988).

[16] Q. Ahmad *et al.*, Physical Review Letters **8**9, 011302 (2002).

[17] K. Hirata *et al.*, Phys. Lett. **2**05, 416 (1988).

[18] R. Becker-Szendy *et al.*, Physical Review D **4**6, 3720 (1992).

[19] T. Theodore, (2016).

[20] S. V. Cao, H. Junting, L. Karol, and N. Federico, (2013).

[21] A. Aurisano *et al.*, Journal of Instrumentation **1**1, P09001–P09001 (2016).

[22] F. Psihas *et al.*, Physical Review D **1**00 (2019).

[23] P. Baldi, J. Bian, L. Hertel, and L. Li, Physical Review D **9**9 (2019).

[24] A. Abhishek, W. Fedorko, P. de Perio, N. Prouse, and J. Z. Ding, Variational autoencoders for generative modelling of water cherenkov detectors, 2019, 1911.02369.

[25] E. Racah *et al.*, 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA) (2016).

[26] C. Ahdida *et al.*, Journal of Instrumentation **1**4, P11028–P11028 (2019).

[27] L. Berns, **1**468, 012165 (2020).

[28] B. Abi *et al.*, arXiv preprint arXiv:2002.03005 (2020).

[29] M. D. Zeiler and R. Fergus, (2013), 1311.2901.

[30] L. v. d. Maaten and G. Hinton, Journal of machine learning research **9**, 2579 (2008).

[31] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, arXiv preprint arXiv:1207.0580 (2012).

[32] K. Simonyan and A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, 1409.1556.

[33] K. He, X. Zhang, S. Ren, and J. Sun, Identity mappings in deep residual networks, 2016, 1603.05027.

[34] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, 2016, 1602.07261.

[35] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, Squeeze-and-excitation networks, 2017, 1709.01507.

[36] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, Mobilenetv2: Inverted residuals and linear bottlenecks, 2018, 1801.04381.

[37] A. Kendall, Y. Gal, and R. Cipolla, Multi-task learning using uncertainty to weigh

losses for scene geometry and semantics, 2017, 1705.07115.

[38] R. R. Selvaraju *et al.*, International Journal of Computer Vision **1**28, 336–359 (2019).

[39] P. Mehta *et al.*, Physics Reports **8**10, 1–124 (2019).

[40] A. Buckley, The hepthesis LaTeX class.

List of figures

List of tables