Application of Adversarial Networks in search for four top quark production in CMS

Master thesis progress report
By Vichayanun Wachirapusitanand
Advisors: Dr Norraphat Srimanobhas
Prof Dr Freya Blekman
(Physics Department, Vrije Universiteit Brussel, Brussel, Belgium)
24 January 2019
Thesis duration

• Started from April 2018
  • Worked with International Institute for High Energy Physics (IIHE) in June 2018

• Expected to complete in April/May 2019

• Expected to present the findings in Siam Physics Congress, 2019, on 6 – 7 June 2019.
Thesis goal

• Using Adversarial Networks to help distinguish between four top quark processes and $t\bar{t}$ processes, while lessen the effects from uncertainty.
Large Hadron Collider (LHC)

- Proton-proton collider with centre-of-mass energy starting from 7 TeV, later upgraded to 14 TeV
- Initial design luminosity at $10^{34}$ cm$^{-2}$ s$^{-1}$
- Plans to upgrade collider energy and luminosity underway
Top quark properties

• Mass of $173.21 \pm 0.51$ (syst) $\pm 0.71$ (stat) GeV
• The heaviest quark to date, heavier than Higgs boson
• Occurs in SM processes
• Decays (mostly) into a W boson and a bottom quark
• Decays instantly before hadronisation occurs

Two interesting processes

Top-antitop ($t\bar{t}$) process (background)

Four top quark production (signal – 90 000x less frequent)

Chatrchyan S et al. Measurement of the $t\bar{t}$ production cross section and the top quark mass in the dilepton channel in pp collisions at $\sqrt{s} = 7$ TeV. Journal of High Energy Physics. 2011;2011(7).

Cross section

- An area in which a collision process occurs if a particle collides
- Proportional to probability in which a collision process occurs (larger cross section for a process = more collision events originating from that particular process)
- Usually measured in barn (1 b = $10^{-24}$ cm$^2$)
- Unique to each process, and can be calculated theoretically (Standard Model in this case).
Why four top quark production?

• Four top quark production can be used to calculate some other quantities in the Standard Model

• The cross section of this process can be altered by new theories, such as top quark compositeness, dark matter, or effective field theory.
Challenge

• For a search in four top quark production, the dominant background is ~ 90 000 times larger than the signal.
• Need to go through the immense data collection in LHC just to find a handful of events containing four top quark production.
Jets

• According to Quantum Chromodynamics, a single quark is not stable, and must couple with other particles, forming a shower of particles called particle jets.

• Can be used as a signature for a single quark in the detector.

• May be created from the process itself or radiation before or after the collision.
• Each top quark decays (almost always) into one bottom quark and one W boson.
• Each W boson decays into either two quarks or a lepton and a neutrino.

• Lepton + jets channel: Our analysis requires one top quark to decay into one bottom quark, a lepton, and a neutrino. Other top quarks must decay into three quarks.

• Expected number of jets
  • 4 in top-antitop production
  • 10 in top-antitop production
Adversarial networks

• Consists of two networks: discriminator and adversary
• Discriminator tries to classify an event…
• …while adversary tries to guess why the discriminator make that classification.
Dataset used in this analysis

- Simulated data: using simulated proton-proton collision events generated from top-antitop production (dominant background) as background events, and four top quark production as signal events.
  - Simulated with 13 TeV centre of mass energy
  - Simulated under the same virtual CMS detector
  - Simulated events containing a single electron or a single muon

- **Blinding policy**: Real data will not be used until we have made every possible checks whether the procedure is correct.
Training data choices

• Number of variables to be inputted into the network
  • 48 variables reflecting differences between top-antitop production and four top quark production, including kinematics of particles and overall topology of the event.
  • 46 variables – same set of variables with 2 of them removed (during the investigation, the removed variables are found to have negative contribution to the network performance)
Training data choices

- Number of jets in each event
  - All events in the available dataset
  - Events with 9 or more jets in the dataset, which has more events from four top quark production (signal) than in top-antitop production (background)

<table>
<thead>
<tr>
<th>Number of jets</th>
<th>Number of events in</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Signal</td>
<td>Background</td>
<td>Signal/total events ratio</td>
</tr>
<tr>
<td>7</td>
<td>23 972</td>
<td>150 535</td>
<td>0.14</td>
</tr>
<tr>
<td>8</td>
<td>51 874</td>
<td>266 969</td>
<td>0.16</td>
</tr>
<tr>
<td>9</td>
<td>44 288</td>
<td>110 178</td>
<td>0.29</td>
</tr>
<tr>
<td>10 or more</td>
<td>53 776</td>
<td>44 145</td>
<td>0.55</td>
</tr>
</tbody>
</table>
Example variable distributions from signal and background events with single electron and single muon.
Event weighting

• Using cross-section-based event weights to better represent the proportions of data in real collisions
  • larger cross section for a process = more collision events originating from that particular process

• Events weighted in a way that background (top-antitop production) has 831,760 events and signal (four top quark production) has 9.2 events (before event preselection)
  • Top-antitop production cross section: 831,760 fb
  • Four top quark production cross section: 9.2 fb
Official MVA discriminators used

Discriminator in 2015 analysis (BDT)
• Trained with boosted decision tree (BDT) via TMVA toolkit.
• Used in four top quark production analyses involving data collected in 2015.*

Discriminator in 2016 analysis (BDT1)
• Trained with boosted decision tree (BDT) via scikit-learn.
• Currently being used in four top quark production analyses involving data collected in 2016.

Real data dataset

- Using real data from proton-proton collision with 35.83 fb$^{-1}$ collected with CMS in 2016
  - Collected with 13 TeV centre of mass energy
  - Containing a single electron or a single muon
Neural network terminology

• **Loss**: A function used during neural network training to indicate how bad the neural network is performing.
  - Loss → 0 if the network is predicting events more and more correctly
  - Loss increases when the network gets worse at predicting events

- **Binary crossentropy**: One of the functions used to calculate loss of a neural network
  - \(- (y \log p + (1 - y) \log(1 - p))\) where \(y\) is truth value and \(p\) is predicted value
  - For individual events, returns 0 if predicted correctly, and grows much larger beyond 1 if predicted incorrectly
  - Can be used to punish the network heavily when it makes wrong predictions
Neural network structure

• One batch normalization layer (”applies a transformation that maintains the mean activation close to 0 and the activation standard deviation close to 1”)

• Number of hidden layers and neurons are permutated with hyperparameter search

• Using binary cross-entropy as its loss. 
  $$-(y \log p + (1 - y) \log(1 - p))$$
  where $y$ is truth value and $p$ is predicted value

• Trained using training part of data for 30 epochs
Neural network structure

• Pros
  • Batch normalization layer guarantees that all input variables are normalized.
  • Hyperparameter search helps us finding best hidden layer configurations.
  • Binary crossentropy loss function helps punishing the network heavily when wrong predictions are made.

• Cons
  • Takes a much longer time to find the optimal network configuration.
Receiver Operating Characteristic curve (ROC curve)

Signal acceptance (the ratio of signal events being accepted or considered as signal-like events)

More area under curve (AUC) means better signal acceptance under same background acceptance.

HT (sum of jets’ transverse momentum) has been widely used as a benchmark in comparing discriminators

Background acceptance (the ratio of background events being accepted or considered as signal-like events)

(Many variants of ROC curve may be used, such as signal efficiency vs background rejection, etc.)
Performance evaluation

• Calculate the area (AUC) under receiving operating characteristic (ROC) curve in
  • Low-jet multiplicity category (8 jets or lower)
  • 9 jets and 3 medium b-tagged jets (9J3M) category
  • 9 jets and 4 medium b-tagged jets (9J4M) category
  • 10 jets and 3 medium b-tagged jets (10J3M) category
  • 10 jets and 4 medium b-tagged jets (10J4M) category

• Calculated using simulated dataset from top-antitop production as background events and four top quark production as signal events on corresponding categories.
Hyperparameter search

• Two or three hidden layers with 50, 100, or 200 neurons each
  • Makes network training not time consuming
• Last layer is set to have tanh activation
• Followed by one final layer with sigmoid activation
  • $2 \times 2 + 2 \times 2 \times 2 = 36$ permutations in total
• Each permutation can take up to an hour of training with 30 epochs.
20 out of 144 possible neural network combinations, sorted by AUC from 10J4M category.

144 combinations include
- 3×3 + 3×3×3 network architectures
- 2 different variable sets
- 2 different training dataset

Green = better than the best official MVA performance in the same channel.
Yellow = performance in between the two official MVAs (BDT and BDT1).
Red = worse than the worst official MVA performance in the same channel.

<table>
<thead>
<tr>
<th>Category with best signal/background ratio</th>
<th>8-J</th>
<th>9J3M</th>
<th>9J4M</th>
<th>10J3M</th>
<th>10J4M</th>
</tr>
</thead>
<tbody>
<tr>
<td>FourTopsHypermodelAlt/9up_vars_with_extras_200_200_100_.hdf5</td>
<td>0.6676</td>
<td>0.61825</td>
<td>0.62649</td>
<td>0.66664</td>
<td>0.66723</td>
</tr>
<tr>
<td>FourTopsHypermodelAlt/9up_vars_remove_less_AUC_200_100_.hdf5</td>
<td>0.66141</td>
<td>0.60823</td>
<td>0.60847</td>
<td>0.66406</td>
<td>0.66453</td>
</tr>
<tr>
<td>FourTopsHypermodelAlt/9up_vars_with_extras_200_100_50_.hdf5</td>
<td>0.68391</td>
<td>0.60688</td>
<td>0.61138</td>
<td>0.65822</td>
<td>0.66332</td>
</tr>
<tr>
<td>FourTopsHypermodelAlt/9up_vars_remove_less_AUC_100_200_200_.hdf5</td>
<td>0.65179</td>
<td>0.59955</td>
<td>0.61729</td>
<td>0.65536</td>
<td>0.66324</td>
</tr>
<tr>
<td>FourTopsHypermodelAlt/9up_vars_remove_less_AUC_200_200_200_.hdf5</td>
<td>0.66613</td>
<td>0.59917</td>
<td>0.605</td>
<td>0.65705</td>
<td>0.6596</td>
</tr>
<tr>
<td>FourTopsHypermodelAlt/9up_vars_with_extras_200_200_50_.hdf5</td>
<td>0.67838</td>
<td>0.61945</td>
<td>0.6183</td>
<td>0.65883</td>
<td>0.65823</td>
</tr>
<tr>
<td>FourTopsHypermodelAlt/vars_with_extras_200_50_200_.hdf5</td>
<td>0.67654</td>
<td>0.61355</td>
<td>0.61618</td>
<td>0.65904</td>
<td>0.65798</td>
</tr>
<tr>
<td>FourTopsHypermodelAlt/9up_vars_with_extras_100_200_200_.hdf5</td>
<td>0.66749</td>
<td>0.56915</td>
<td>0.59742</td>
<td>0.63355</td>
<td>0.6749</td>
</tr>
<tr>
<td>FourTopsHypermodelAlt/9up_vars_with_extras_200_100_100_.hdf5</td>
<td>0.65926</td>
<td>0.6077</td>
<td>0.59782</td>
<td>0.65721</td>
<td>0.65739</td>
</tr>
<tr>
<td>FourTopsHypermodelAlt/vars_remove_less_AUC_100_200_50_.hdf5</td>
<td>0.66908</td>
<td>0.61602</td>
<td>0.607</td>
<td>0.65599</td>
<td>0.65651</td>
</tr>
<tr>
<td>FourTopsHypermodelAlt/9up_vars_with_extras_200_100_100_.hdf5</td>
<td>0.62773</td>
<td>0.5709</td>
<td>0.5813</td>
<td>0.64196</td>
<td>0.65597</td>
</tr>
<tr>
<td>FourTopsHypermodelAlt/9up_vars_remove_less_AUC_200_200_100_.hdf5</td>
<td>0.67344</td>
<td>0.61516</td>
<td>0.61294</td>
<td>0.6582</td>
<td>0.65377</td>
</tr>
<tr>
<td>FourTopsHypermodelAlt/9up_vars_with_extras_200_50_100_.hdf5</td>
<td>0.69619</td>
<td>0.60747</td>
<td>0.6145</td>
<td>0.65506</td>
<td>0.6553</td>
</tr>
<tr>
<td>FourTopsHypermodelAlt/9up_vars_with_extras_200_50_50_.hdf5</td>
<td>0.64131</td>
<td>0.57333</td>
<td>0.57943</td>
<td>0.63763</td>
<td>0.65026</td>
</tr>
<tr>
<td>FourTopsHypermodelAlt/9up_vars_with_extras_50_50_50_.hdf5</td>
<td>0.63286</td>
<td>0.59137</td>
<td>0.5957</td>
<td>0.64947</td>
<td>0.65005</td>
</tr>
<tr>
<td>FourTopsHypermodelAlt/9up_vars_remove_less_AUC_200_100_100_.hdf5</td>
<td>0.66794</td>
<td>0.60769</td>
<td>0.60271</td>
<td>0.65634</td>
<td>0.64889</td>
</tr>
<tr>
<td>FourTopsHypermodelAlt/9up_vars_with_extras_50_50_50_.hdf5</td>
<td>0.66235</td>
<td>0.60154</td>
<td>0.60424</td>
<td>0.64714</td>
<td>0.64816</td>
</tr>
<tr>
<td>FourTopsHypermodelAlt/9up_vars_with_extras_200_50_50_.hdf5</td>
<td>0.67362</td>
<td>0.60978</td>
<td>0.5907</td>
<td>0.65947</td>
<td>0.64715</td>
</tr>
<tr>
<td>FourTopsHypermodelAlt/9up_vars_with_extras_200_50_50_.hdf5</td>
<td>0.65272</td>
<td>0.60242</td>
<td>0.58806</td>
<td>0.65114</td>
<td>0.64662</td>
</tr>
<tr>
<td>FourTopsHypermodelAlt/vars_with_extras_200_100_200_.hdf5</td>
<td>0.6915</td>
<td>0.60806</td>
<td>0.62259</td>
<td>0.64042</td>
<td>0.64632</td>
</tr>
</tbody>
</table>

BDT (2015) | 0.6701 | 0.6102 | 0.6019 | 0.6426 | 0.6400 |
BDT1 (2016) | 0.7244 | 0.6444 | 0.6375 | 0.6597 | 0.6272 |
Best neural network chosen contains neural layers with 200, 200, and 100 neurons, with last layer having tanh activation.

This network has a slightly worse performance in low-jet multiplicities, but has a higher performance in signal-rich categories (10J3M and 10J4M), crucial to the discovery. Also, it gives the output range close to [0, 1], making it more plausible to physicists.
ROC curve for inclusive category (left) and 10J4M signal-rich category (right)
Keras output distribution in the range of [0, 1], inclusive, for single electron channel (left) and single muon channel (right). Using cross-section-based weights result in extremely huge number of events having discriminator value close to 0.
For visualization, we can apply log10 to the output. The distribution still gives the same ROC curve, therefore there are no changes in network performance.

Keras output distribution, inclusive, for signal and background dataset.
Distribution of negative log10 value of network output for single electron

By using uneven bins in a histogram, we can guarantee that every NN output histograms from every category has no bins with zero background events, which can be used in limit-setting tool.

Furthermore, to make sure that no events are left behind due to the range of $[-\infty, 0]$, the distribution is flipped using negative value. Events with an extremely low NN output will fall in the overflow bin.
Discriminator conclusion

• Using neural network containing layers with 200, 200, and 100 neurons, where last layer having tanh activation.
  • Slightly higher performance in signal-rich categories, with higher probability to contain signal events.
  • Gives a better sounding output range close to [0, 1].

• Using log10 value of the network output to better represent the output distribution.

• There are still uncertainties in the distribution.
Using NN output in analysis

- The output distribution can be used to calculate the cross section limit for the four-tops process, using asymptotic $\text{CL}_S$ method.
  - Calculate the upper limit of the strength of signal process $\mu$ from
    \[ E[n_i] = \mu s_i + b_i \]
    where $E[n_i], s_i, b_i$ is the number of events in each bin in data histogram, signal process histogram, and background process histogram
- Requires histograms of the output distribution from certain background processes.
- All histograms must not contain any bins with zero background events.

<table>
<thead>
<tr>
<th></th>
<th>Upper limit of four top quark production cross section (as multiples of 9.2 fb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expected</td>
<td>$16.4^{+9.8}_{-5.7}$</td>
</tr>
</tbody>
</table>

Expected limit obtained are from traditional neural networks only.

**Better result with adversarial neural networks?**

Adversarial Networks

• Two sets of networks
  • Discriminator: Minimizes the loss $L_f$ occurred due to incorrect predictions on whether the event is signal or background
  • Adversarial: Maximizes the loss $L_r$ occurred due to incorrect predictions on whether the signal/background comes from particular variable distribution or samples

• Total loss defined by $L_f - \lambda L_r$, where $\lambda$ is a controllable hyperparameter
Proposal: Adversarial networks

- Train the adversarial network with a higher level objective, can be upper limit from $\text{CL}_S$ method.
- Train with training events with adjusted uncertainty, and let the adversarial network guess whether an input event is from unadjusted sample or adjusted sample.
- We can use different configurations and compare the upper limits.
Proposal: Adversarial networks

- HEP example in arXiv:1611.01046
  - Trains the adversarial network with approximate median significance (AMS) as a higher objective. (AMS is dependent of discriminator output.)
  - Trains the adversarial network to determine whether an input event contains pileup (more than one collision event - $Z = 1$) or not ($Z = 0$).
  - Adversarial networks are trained with different parameter configurations in order to compare AMS values with each other.
Naïve attempt on adversarial networks

• Two neural networks
  • Discriminator: two hidden layers with 100 neurons each. Discriminates the event into signal (four top quark production) and background (top-antitop production)
  • Adversary: two hidden layers with 10 neurons each. Determines whether the event comes from samples with Jet Energy Scale (JES) modifications, the most important uncertainty, or not.

• Using binary cross-entropy as loss functions for both networks.
• Using the same set of 13 input variables used in BDT discriminator in 2016 analysis.
Naïve attempt on adversarial networks

• Individual outputs for networks:
  • Discriminator: one output node with output [0, 1] representing signal or background
  • Adversary: three output nodes with softmax activation representing whether the event comes from central samples (no JES modification), JESup samples (with upper JES modification), or JESdown samples (with lower JES modification)
Naïve attempt on adversarial networks

Adversarial training does not change the ROC curve within events with 9 jets category.
Naïve attempt on adversarial networks

Adversarial training also does not change the ROC curve within events with 10+ jets category.

→ Adversarial training fails?
Naïve attempt on adversarial networks

Flaws

• Discriminator and adversary network not deep enough
• Not using a larger set of variables
• Discriminator network not using the same configuration

• Lots of room for improvement for adversarial networks!
To-do list

Done
• Half of the adversarial network studied.
• Currently using CMS datasets for neural network training.
• Using neural network discriminator to calculate the upper limits of four top quark production

To do
• Design adversarial part of the network and evaluate its performance
• Calculate the new expected upper limits of four top quark production

Overall progress: Can be completed by April/May 2019