Ablate, Variate, and Contemplate: Visual Analytics for Discovering Neural Architectures

CPSC 547

1

Machine Learning Background

- What is Machine Learning (ML)?
  - A machine learning model is an algorithm that predicts a target label from a set of predictor variables.
  - It learns the relationship between the features and target labels using training dataset.

  - Some technical terms:
    - Epoch
    - Loss
    - Training, validation and test dataset

MACHINE LEARNING
Machine learning begins to flourish

2

Neural Network (NN) Background

- How neural networks work?
  - Class of ML models inspired by message passing mechanisms in brain.
  - Two main components: Architecture and parameters for each architecture components
  - Architecture:
    - A computation graph mapping from input to output
    - The nodes of computation graphs are layers

Input Layer   Hidden Layer   Output Layer
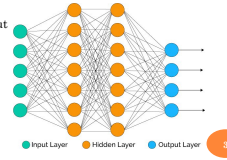
3

What is the problem?

- Configuration of layers and parameters are important in deep learning models.
- Small changes in parameter, huge difference in performance
- Training takes time and requires resources.
- The initial choice of NN architecture is a significant barrier for being successful.

"Designing neural networks is hard for humans. Even small networks can behave in ways that defy comprehension; large, multi-layer, nonlinear networks can be downright mystifying."

4

What are current Approach to solve this problem?

- Experiment with different configurations and architectures manually by using guidelines.

- Purely automated neural architecture search to generate and train the architectures.

- Using current visual analytical tools to make NN more interpretable and customizable.

5

Downsides of purely automatic Neural Search (ANAS)?

- Search thousands of architectures.
- Using very expensive resources for example:
  - Algorithms in reinforcement learning using 1800 GPU days
  - Evolutionary algorithms taking 3150 GPU days

- The best result might be too large for deploy if you do not have resources!

- Probably if we access this type of hardware either we have expertise for manually designing or have access to experts.

6

Downsides of Current Visual tools?

- They assume a good performant model architecture has been chosen!

- Use tools to fine tune it! How?
  - User can inspect how various components contribute to prediction.

  - Allow users to build and train toy models to check the effects of hyperparameters.

  - Debugging a network, which changes must be made for better performance, by analyzing activations, gradients, and failure cases.

7

What we really need?

- Initially sample small set of architectures, and then visualize it in the model space.

- Put human in the loop of neural architectures search.

- Human can do local, constraint, automated search for the models of interest and able to handcraft it easily.

- Provide a data scientist with an initial performant model to explore.

8

Their approach?

- Rapid Exploration of Model Architectures and Parameters (REMAP), a client/server tool for semi-automated NN search.

- Combination of global inspection(exploration) and local experimentation.

- Stop searching for architectures when model-builder found an acceptable model.

- Don't take much time, and not require huge resources, large category of end users!

9

What is their design study?

- Interview with four model builders
- Two type of questions:
  - 1) about practices in manually altering
  - 2) what visualization is good for non-experts for the human-in-the loop system for NN architecture search
- Interviews were held one-on-one using an online conferencing software and recorded audio.
- Establish a set of goals and tasks used in manual discovery of NN architectures by each participant.

10

What are their goals?

- G1: Find Baseline Model

  1) Start with a network you know is performant (either in literature review or pretrained neural network) as your baseline (priority on small model which train fast)

  2) Start fine-tune it by small changes like hyperparameters tuning/using different dropouts

11

What are their goals? (cont.)

- G2: Generate Ablation and Variation
Two tasks on performant network:

  Ablation studies: remove layers in a principled way and explore how this changes the performance of the network.

  Generate variations: generate variations of the architecture by switching out or re-parameterizing layers that were shown to be less useful by the ablations. Need to code for each version.

12

What are their goals? (cont.)

- G3: Explain/Understand Architectures
You might be able to glean a better understanding of how neural networks are constructed by viewing the generated architectures.

- G4: Human-supplied Constrained Search:
  - If there is sufficient time/resources/ clean data using Auto NA search is the best, there is no need for human.
  - If not, human can be controller by:
    - Defining constraints on search
    - Point an automated search to particular part

13

What are their tasks?

- Starting from baseline models takes time/ hundreds of million parameters and cannot easily experimented

  - task1) Quickly search for baseline architectures through a visual overview of models

- Ablation and Variation actions/ human should provide simple constraint on architecture

  - task2) Generate local, constrained searches in the neighborhood of baseline models

- Support visual comparisons to help user have strategy for generating variations and ablation and explore in space model

  - task3) Visually compare subsets of models to understand small, local differences in architecture

14

Visual Model Selection challenges?

First challenge:

- The parameter space for NN is potentially infinite (we can always add layers!)

- To interpret model space:
  - Two additional projections based on two type of model interpretability identified in Lipton's work [1].
    - Structural
    - Post-hoc
  - 2-D Projections are generated from distance metrics using scikit-learn's implementation of Multidimensional Scaling.

15

What is structural interpretability ?

- How the Components of a model function.

- A distance metric based on structural interpretability would place models with similar computational components, or layers, close to each other in the projection.

- How they implement?
  - They used OTMANN distance, an Optimal Transport-based distance metric.

16

What is post-hoc interpretability?

- Understanding a model based on its predictions.

- A distance metric based on post-hoc interpretability would place models close together in the projection if they have similar predictions on a held-out test set.

- How they implement?
  - They used the edit distance between the two architectures' predictions on the test set.

Visual Model Selection challenges? (cont.)

Second challenge:

- Finding visual encoding and embedding techniques for visualizing NN that enables comparison of networks

- While conveying shape and computation of networks.

Their visual encoding?

- Sequential Neural Architecture Chips (SNACs)

- A space-efficient, adaptable encoding for feed-forward neural networks

- It explicitly uses properties of NN such as the sequence of layers, in its visual encoding

---

SNACs

- Easy visual comparisons across several architectures via juxtaposition in a tabular format.
- Layer type is redundantly encoded with both color and symbol.
- Activation layers have glyphs for three possible activation functions:
  - hyperbolic tangent ($tanh$), rectified linear unit (ReLU), and sigmoid
- Dropout layers feature a dotted border to signify that some activations are being dropped.

Developing initial set of architectures of REMAP?

- A starting set of models is initially sampled from the space in a preprocessing stage, but how?
  1. A small portion of random schema based on ANAS

  1. Using Markov chains dictates the potential transition probabilities from layer to layer:
     - Starting from an initial state, the first layer is sampled, then its hyperparameters are sampled from a grid. Then, its succeeding layer is sampled based on what valid transitions are available.

  2. Transition probabilities and layer hyperparameters were chosen based on similar schemes in the ANAS literature, as well as conventional rules of thumb.
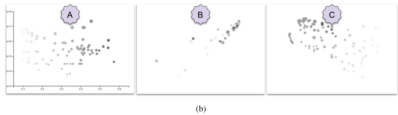

How the whole user interface look like?
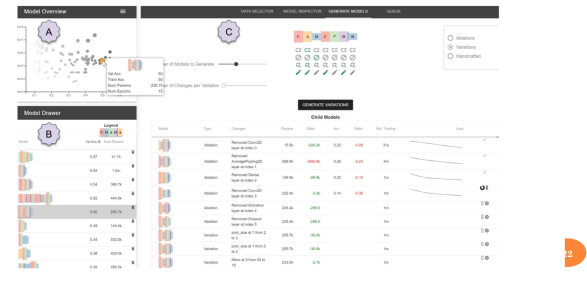
The interface components

- The Model Overview
  - Represented by a scatter plot
  - Three types
  - Find the baseline model here from the pretrained models.
    - Circle represents trained neural net
    - The darkness of the circle encodes the model accuracy
    - The radius of the circle encodes the log of the number of parameters

(b)

---

The interface components (cont.)

- The Model Drawer
  - Retaining a subset of interesting models during analysis
  - Drag model of interest here and compare them

The interface components (cont.)

- The Data Selection Panel
- If users are particularly interested in performance on certain classes in the data, select a data class
- By selecting individual classes from the validation data, users can update the darkness of circles in the model overview to see how all models perform on a given class.
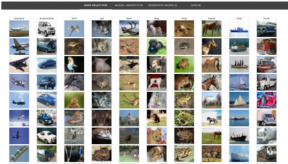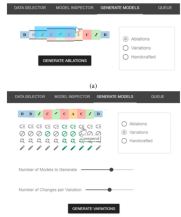

The interface components (cont.)

- The Model Inspection Panel
- See more granular information about a highlighted model.
- By Confusion Matrix/Training curve

The interface components (cont.)

- The Generate Models tab
- currently selected, allows for users to create new models via ablations, variations, or handcrafted templates.

- Each child model is embedded into the model overview, and can be moved to the model drawer to become a model baseline.

The interface components (cont.)

- The Generate Models tab
- Users can view the current training progress of models
- can view the history of all training across all models in the Queue tab.
- Can reorder/delete

Global inspection and local experimentation

Global inspection

User first explore an overview of a set of pre-trained small models

- Visual Overview of set of models leads user identify interesting cluster of architecture

Local experimentation

Then user guide to discovery of new models via operations on existing models

- Semi-automated search through model space
- Run ablation (effects of removing) and Variation experiments (replacing/adding layers)
- Hand craft new models using a simple graphical interface

An Ablation Study

- Ablations create a set of models, one for each layer with that layer removed.

- the network is retrained with each feature of interested turned off, one at a time.

- The goal of ablations is to determine the effect/importance of each feature of a network.

- This might then drive certain features to be pruned, or for those features to be duplicated.

- Train those models for the same number of epochs as the parent model, and display to the user the change in validation accuracy.

Variation

- Several new models' generation by random atomic changes of an existing model

- By default, the variation command will randomly remove, add, replace, prepend, or reparametrize layers.

- The Variations feature runs constrained searches in the neighborhood of a selected model.
  - Users can constrain the random generation of variations by specifying a subset of types of variations for a given layer, as well as the number of variations allowed per model.

- This might then drive certain features to be pruned, or for those features to be duplicated.

How they evaluate ReMap

- Using expert feedbacks

- Case study

How they evaluate ReMap by Using expert feedbacks?

- Same participants.
- two-hour online interview.
- audio and screen sharing are recorded, show demo first
- Two tasks of unconstraint and constrain search are given to them:
  - on discovering a performant neural network architecture for image classification.
  - on the CIFAR-10 dataset, a collection of 50,000 training images and 10,000 testing images each labeled as one of ten mutually exclusive classes using app features.
- Task1) find NN has highest accuracy on first 10,000 images.
- Task2) find NN can deploy on mobile app (up to 100,000 parameters) used to only classify two labels of cats and birds.

33

Hand-craft the models

- User can handcraft the model to whatever he knows, train them to have his trade-offs.
- Was added based on feedback from a validation study with model builders
- Remove, add, or modify any layer in the model by clicking on a layer or connections between layer

34

How they evaluate ReMap by Using case study?

- discover CNN for classification of sketches
- Quick Draw dataset contains millions of sketches of 50 classes
- To solve each problem, perform three tasks

35

REMAP Generalizability

It

- It is generalizable as long as we have two components of:
  - a set of projections of models
  - a local sampling method to generate methods

- All projections are are generalizable to any machine learning model .

36

REMAP Scalability

- Remove the size cap of REMAP
  - Train more larger models applicable for industry.

- Visual encoding not support skip connections which has additional linkage between layers.

- The scope is limited to network architectures that are linked lists:
  - because they are simpler to understand
  - A common architecture that are more performant than non-neural network models for image classification problems

37

Discussion) REMAP Advantages

- User can trade-off between the size of model, the performance of individual classes, and the overall performance of the resulting model.

- User can constrain on number of parameters, using his domain knowledge and deployment scenario.

- Global and local inspection of networks (Model Selection)

- Allowing user-directed exploration of the model space :
  - Provide starting point for user to find models that match their understanding of the data, the importance of particular classes, or particular number of parameters

- Manually construct/modify architecture via a simple drag-and-drop interface

38

Discussion) REMAP disadvantages

- Only consider non-expert user with limited source of architecture, the baseline models should be small and trainable on more typical hardware. Not state of the art!

- Constrained on generated baseline model, cannot have fine-grained control over the model building process at first stage.

- Better for education/ or playing with data and NNs.

- More audience but less useful results in real applications.

- We can encode the number of parameters by each layer as well

39

Any Questions?

40