Training and Validation

UW ECE 657A - Core Topic

Mark Crowley

- Data types, sources, nature, scales and distributions
- Data representations, transformation and normalization
- Experimental Methodology: Training-Test-Validation, Albation Studies

Lecture Outline

1. Dividing the Data
    - Error Estimation and Training Methods
    - Validation
2. Ablation Studies
    - Definition
    - Uses
    - An Example

Lecture Outline

1. Dividing the Data
    - Error Estimation and Training Methods
    - Validation

2. Ablation Studies
    - Definition
    - Uses
    - An Example

Training and Error Estimation

- Resubstitution
- Holdout Method
- Leave-one-out
- K-fold Cross Validation
- Bootstrap

Error Estimation and Training Methods

Resubstitution Method: Uses all the available data for training and then the same data is used for testing

- Optimistically biased estimate, especially when the ratio of sample size to dimensionality is small

Holdout Method: Uses half the data for training and the other half is used for testing; training and test sets are independent

- Pessimistically biased estimate
- Different partitionings may give different results

Leave-one-out Method: A classifier is designed using (n- 1) samples and evaluated on the remaining one sample; this is repeated n times with different training sets of size (n-1)

- Estimate is unbiased but it has a large variance
- Large computational requirement because n different classifiers have to be designed

Error Estimation Methods

K-fold cross validation (Rotation): divide the available samples into K disjoint subsets for training and the remaining subset for test. Can repeat the process with different partitioning.

- Estimate has lower bias than the holdout method and is cheaper to implement than leave-one-out method.

Bootstrap Method: Generate many bootstrap sets of size n by sampling with replacement.

- Bootstrap estimates can have lower variance than the leave-one-out method.
- Computationally more demanding.
- Useful in small sample size situations.

Validation

- Sometimes we need to tune certain parameters of the classifier such as *k* in the K Nearest Neighbours classifier or parameters that control the architecture of a neural network.
- This can be done using a **validation set**, leaving the test set to test the optimal choice
- So, the data is divided into 3 subsets: **Training, Validation** and **Testing**

Lecture Outline

1. Dividing the Data
   - Error Estimation and Training Methods
   - Validation

2. Ablation Studies
   - Definition
   - Uses
   - An Example

Ablation Studies

Once you have a trained model that gives you some kind of response,

- State-of-the-Art (SotA) accuracy, precision, performance!
- Unprecedented solution to a previously believed unsolvable problem?
- The first practical solution to this problem ever presented *(at least to our knowledge)*.

What what if someone asks you?

- Why does it work?
- Which part is most important?
- Where is the computational bottleneck (ie. what is slowing it down?)
- followup...is the slow part *that* essential?

How do you know figure out **why** it is working?

Definitions

Definition

Ablation - the removal, especially of organs, abnormal growths, or harmful substances, from the body by mechanical means, as by surgery. – Dictionary.com

Definition from (Fawcett and Hoos, 2013):

> *Our use of the term ablation follows that of (Aghaeepour and Hoos, 2013) and loosely echoes its meaning in medicine, where it refers to the surgical removal of organs, organ parts or tissues. We ablate (i.e., remove) changes in the settings of algorithm parameters to better understand the contribution of those changes to observed differences in algorithm performance.*

Motivation for Ablation Studies

As one person puts it (see this twitter thread by @fchollet (`https://threader.app/thread/1012721582148550662`)

- How do you determine 'causality' between which parts of your system are responsible for the performance?
- Advice: "Spend at least ~10% of your experimentation time on an honest effort to disprove your thesis."

Automated Parameter Tuning

These tools and other algorithm configuration tools help to set the many complex parameters needed to achieve optimal, or at least maximal, performance. But they spit out the parameters without any explanation.

So in (Fawcett and Hoos, 2013 and 2016) they propose ways to:

> *"help these algorithm developers answer questions about the high-quality configurations produced by these tools, specifically about which parameter changes contribute most to improved performance."*

What Ablation Is and Isn't

It Is...

- a good way to determine what parts of your model are *useful*, which are *necessary* and which *may be unnecessary*
- an approach to help you *understand* and *explain* you model to others by showing how each part contributes to your state-of-the-art performance
- essentially a method for improving your model selection/design process

What Ablation Is and Isn't

It Is...

- a good way to determine what parts of your model are *useful*, which are *necessary* and which *may be unnecessary*
- an approach to help you *understand* and *explain* you model to others by showing how each part contributes to your state-of-the-art performance
- essentially a method for improving your model selection/design process
- **Therefore...** : all ablation analysis should be done on
    1. The Test Dataset?
    2. The Training Dataset?
    3. A Validation Training Dataset?

What Ablation Is and Isn't

It Is...

- a good way to determine what parts of your model are *useful*, which are *necessary* and which *may be unnecessary*
- an approach to help you *understand* and *explain* you model to others by showing how each part contributes to your state-of-the-art performance
- essentially a method for improving your model selection/design process
- **Therefore...** : all ablation analysis should be done on
  1. The Test Dataset?
  2. The Training Dataset?
  3. A Validation Training Dataset?
- **Answer:** 3. If the goal is to use ablation to improve the model design, then such analysis must happen on a held out validation dataset, not the final testing dataset.

What Ablation Is and Isn't

It Is Not...

- a *regularization method*
- a way to improve your testing numbers (accuracy, recall, confidence) higher
- a way to fill in the space of your paper with more experiments and graphs
- it *will* do this, but that is not the purpose.
  - Advice: If you cannot fill a 6-9 page paper with your own background, theories, data, methodology and results then adding two pages of ablation studies will not save you.

What Ablation Is and Isn't

It Is Not...

- a *regularization method* why not?
- a way to improve your testing numbers (accuracy, recall, confidence) higher
- a way to fill in the space of your paper with more experiments and graphs
- it *will* do this, but that is not the purpose.
  - Advice: If you cannot fill a 6-9 page paper with your own background, theories, data, methodology and results then adding two pages of ablation studies will not save you.

An Example

A nice example explained here: https://stats.stackexchange.com/questions/380040/what-is-an-ablation-study-and-is-there-a-systematic-way-to-perform-it

They use

- (Uijlings, 2012) : “Selective Search for Object Recognition.”
- SIFT feature extraction
- SVM supervised training
- loop and strengthen hypotheses.

They then feed the output into a 5CNN+2FC network:

- (Girshick, 2014) : “Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation.”

Selective Search Architecture

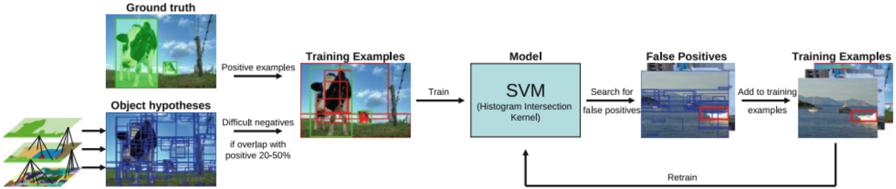


Fig. 3 The training procedure of our object recognition pipeline. As positive learning examples we use the ground truth. As negatives we use examples that have a 20–50 % overlap with the positive examples. We iteratively add hard negatives using a retraining phase

Figure: The original Selective Search architecutre from [Uijlings, 2013]

R-CNN Architecture

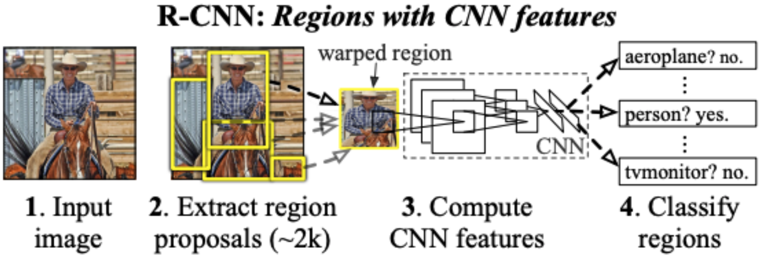
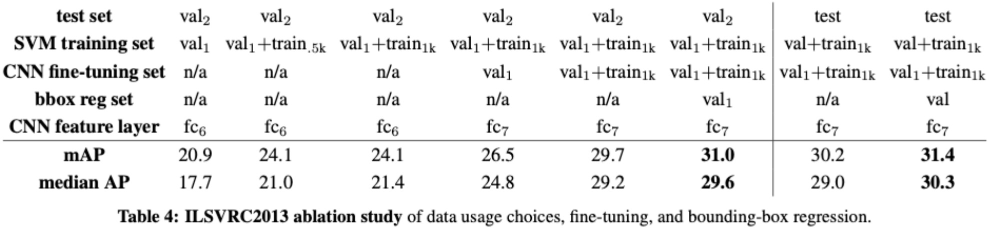


Figure: The modified R-CNN architecutre which uses SS, and performs an ablation study to justify their design. [Girshick, 2014]

Ablation Results



test set	val$_2$	val$_2$	val$_2$	val$_2$	val$_2$	val$_2$	test	test
SVM training set	val$_1$	val$_1$+train$_{.5k}$	val$_1$+train$_{1k}$	val$_1$+train$_{1k}$	val$_1$+train$_{1k}$	val$_1$+train$_{1k}$	val+train$_{1k}$	val+train$_{1k}$
CNN fine-tuning set	n/a	n/a	n/a	val$_1$	val$_1$+train$_{1k}$	val$_1$+train$_{1k}$	val$_1$+train$_{1k}$	val$_1$+train$_{1k}$
bbox reg set	n/a	n/a	n/a	n/a	n/a	val$_1$	n/a	val
CNN feature layer	fc$_6$	fc$_6$	fc$_6$	fc$_7$	fc$_7$	fc$_7$	fc$_7$	fc$_7$
mAP	20.9	24.1	24.1	26.5	29.7	**31.0**	30.2	**31.4**
median AP	17.7	21.0	21.4	24.8	29.2	**29.6**	29.0	**30.3**

Table 4: ILSVRC2013 ablation study of data usage choices, fine-tuning, and bounding-box regression.

Figure: Ablation Table Results from (Girshick, 2014)

What they Found

- Investigate which parts are needed and which aren't
- They found that the SIFT features were not as critical if there was a *high-capacity CNN* to localize objects
- They also found that the CNN could be *pre-trained* on a large, unrelated dataset of images
- Then it can be fine-tuned for the specific problem. This worked better than specialized computer vision methods, such as SIFT.

Conclusion

They would only have found this through ablation experiments.

References

- (Fawcett and Hoos, 2013) Chris Fawcett and Holger H. Hoos. Analysing differences between algorithm configurations through ablation. Proceedings of the 10th Metaheuristics International Conference (MIC 2013), pp. 123-132, 2013.
- (Girschick, 2014) Girshick, Ross, Jeff Donahue, Trevor Darrell, and Jitendra Malik. “Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation.” ArXiv:1311.2524 , October 22, 2014. http://arxiv.org/abs/1311.2524.
- (Newell, 1975) Allen Newell. A Tutorial on Speech Understanding Systems. In Speech Recognition: Invited Papers Presented at the 1974 IEEE Symposium. New York: Academic. p. 43.
- (Uijlings, 2013) Uijlings, J. R. R., K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders. “Selective Search for Object Recognition.” International Journal of Computer Vision 104, no. 2 (September 2013): 154–71. https://doi.org/10.1007/s11263-013-0620-5.