Evaluation

CSE 598 Introduction to Deep Learning

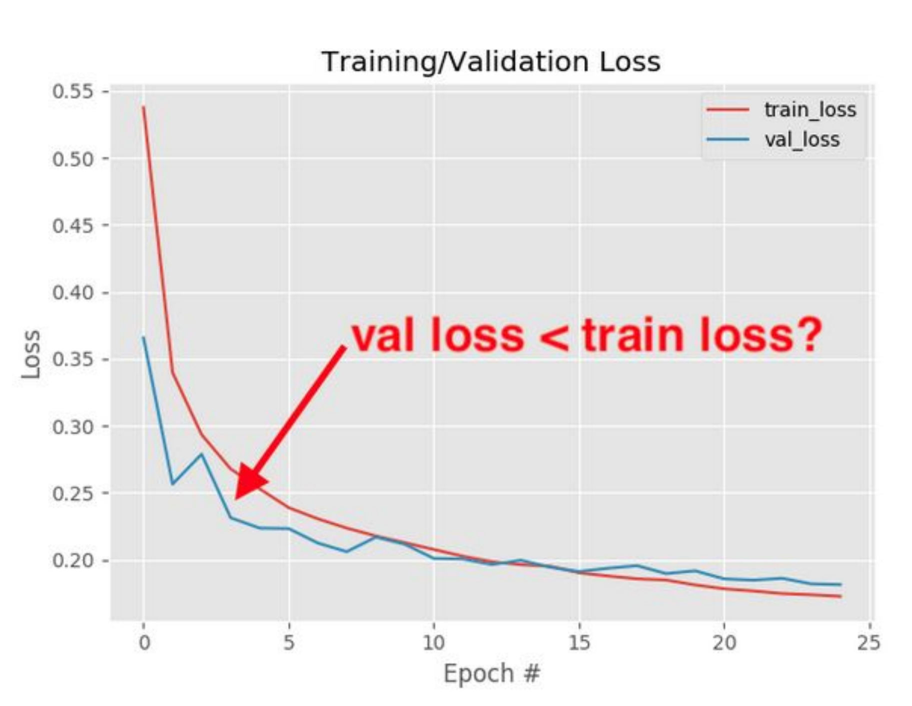
Evaluation

Remember the recipe:

1. Until you get tired of trying different stuff ...
   a. Define an architecture
      i. For example:
         1. add dropout, add more hidden layers, make the hidden layers larger, etc.
         2. make the hidden state of an LSTM larger or smaller, stack several LSTMs, etc.
         3. add more convolutions and pooling layers, change filter sizes, etc.
   b. Train with the training and validation datasets
      i. Usually for however many epochs you observe a decrease in the loss function
      ii. Fit parameters with the training dataset, "evaluate" with the validation set after each epoch
      iii. Early Stopping: have some patience (== stop after a small number of epochs without improvement)
2. Select the best model based on the results with the validation dataset
3. AFTER you are done declaring the winner, evaluate with the test dataset

Evaluation

- Generally, the validation loss should go down quickly (after each epoch)



Evaluation

- Losses are great internally, to fit parameters and the like
  - we know we need to backpropagate a float to update weights
- But we need metrics that tell us how good we are at solving a problem
  - more intuitive metrics

- If your model classifies instances into bins:
  - Accuracy: How many predictions are correct?
  - Precision, Recall and F-measure

Accuracy

- How often does your model get the label right?
  - Very simple:
    - count the number of predictions (== number of instances in the test dataset)
    - count the number of correct predictions (== number of instances in the test dataset that the model gets right
    - divide
- Example (5 instances):
  - GOLD 1: shirt 2: shirt 3: sandal 4: boot 5: dress
  - PREDICTED 1: shirt 2: shirt 3: shirt 4: boot 5: shirt

  Accuracy: 3 / 5 = 0.6 (or 60%)

Accuracy

- Accuracy is not a good choice if the dataset is not balanced
  - Balanced == frequency of all labels is roughly the same
  - Example: Build a classifier to predict whether a red Ferrari will drive by **699 S Mill Ave, Tempe, AZ 85281** within one second
    - Say that one red Ferrari drives by each month (the ground truth)
      - rich people get lost (and attend ASU too)
    - There are 30 x 24 x 60 x 60 = 2,592,000 seconds in a month
    - Assuming it takes a red Ferrari one second to drive the SCAI building, a classifier that always predicts NO will get 2,591,999 / 2,592,000 = 1.0 Accuracy
      - despite it never predicts the YES label
      - This classifier is useless — it is the same than the majority label

Accuracy

- Many real classifications are not balanced:
  - Will it rain tomorrow in Dallas?
    - "Always NO" gets good accuracy
-

		Predicted/Classified	
		Negative	Positive
Actual	Negative	998	0
	Positive	1	1

Accuracy

- Many real classifications are not balanced:
  - Will it rain tomorrow in Dallas?
    - "Always NO" gets good accuracy

```
from sklearn.metrics import confusion_matrix, ConfusionMatrixDisplay
cm = confusion_matrix(ground_truth, predictions)
```

	precision	recall	f1-score	support
T-shirt/top	0.86	0.83	0.85	1000
Trouser	1.00	0.97	0.99	1000
Pullover	0.83	0.88	0.85	1000
Dress	0.85	0.95	0.90	1000
Coat	0.84	0.83	0.84	1000
Sandal	0.94	0.99	0.96	1000
Shirt	0.78	0.68	0.73	1000
Sneaker	0.90	0.97	0.94	1000
Bag	0.99	0.97	0.98	1000
Ankle boot	1.00	0.88	0.93	1000
accuracy			0.90	10000
macro avg	0.90	0.90	0.90	10000
weighted avg	0.90	0.90	0.90	10000

Precision, Recall and F-score

		Predicted	
		Negative	**Positive**
Actual	**Negative**	True Negative	False Positive
	Positive	False Negative	True Positive

Precision, Recall and F-Measure

- Assuming we care about the POSITIVE class:
  - Precision: Out of all POSITIVE predictions, how many are correct?

$$\textbf{Precision} = \frac{\text{True Positive}}{\text{Actual Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Positive}}$$

		Predicted	
		Negative	**Positive**
Actual	**Negative**	True Negative	False Positive
	Positive	False Negative	True Positive

numerator

denominator

Precision, Recall and F-score

- Assuming we care about the POSITIVE class:
  - Recall: How many of the actual (true) POSITIVE instances did the model get right?

$$\textbf{Recall} = \frac{\text{True Positive}}{\text{Predicted Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Negative}}$$

		Predicted	
		Negative	**Positive**
Actual	**Negative**	True Negative	False Positive
	Positive	False Negative	True Positive

numerator

denominator

Precision, Recall and F-score

- Examples:
  - PRED: + + + + + + - -
  - GOLD: + + + + - - - -
  - Precision: 4/6 = 0.66 Recall: 4/4 = 1.0

  - PRED: + + + + + + + +
  - GOLD: + + + + - - - -
  - Precision: 4/8 = 0.50 Recall: 4/4 = 1.0

Precision, Recall and F-score

- Do you care about precision or recall?
  - Obviously you want P = 1.0 and R = 1.0
    - but you won't get that for any *difficult* problem. Be suspicious if you do. Don't celebrate.

Precision, Recall and F-score

- Do you care about precision or recall?
  - Obviously you want P = 1.0 and R = 1.0
    - but you won't get that for any *difficult* problem. Be suspicious if you do.
  - Oh well, it depends
    - Consider binary classification, and we define POSITIVE labels as

Arrest person X because he committed a crime	P	R
Student will drop the class (and the instructor wants to avoid that)	P	R
Patient has a bad disease	P	R
Patient is healthy and should be discharged from the hospital	P	R
Car accident in the next second (so apply brakes immediately)	P	R
It is safe to pass the slow car ahead (change lanes, accelerate, etc.)	P	R

Precision, Recall and F-score

- Do you care about precision or recall?
  - Arrest person X because he committed a crime P R
    - A false positive means that...
    - A false negative means that ...
    - high P and low recall VS. low R and high recall

Precision, Recall and F-score

- Do you care about precision or recall?
  - Student will drop the class (and the instructor wants to avoid that)   P   R
    - A false positive means that...
    - A false negative means that ...
    - high P and low recall VS. low R and high recall

Precision, Recall and F-score

- Do you care about precision or recall?
    - Patient has a bad disease P R
        - A false positive means that...
        - A false negative means that ...
        - high P and low recall VS. low R and high recall

Precision, Recall and F-score

- Do you care about precision or recall?
  - Patient is healthy and should be discharged from the hospital P R
    - A false positive means that...
    - A false negative means that ...
    - high P and low recall VS. low R and high recall

Precision, Recall and F-score

- Do you care about precision or recall?
    - Car accident in the next second (so apply brakes immediately)   P   R
        - A false positive means that...
        - A false negative means that ...
        - high P and low recall VS. low R and high recall

Precision, Recall and F-score

- Do you care about precision or recall?
  - It is safe to pass the slow car ahead (change lanes, accelerate, etc.)   P   R
    - A false positive means that...
    - A false negative means that ...
    - high P and low recall VS. low R and high recall

Precision, Recall and F-score

- F-score: a combination of Precision and Recall
  - not just an average, allows to assign different weights to Precision and Recall

$$F_\beta = (1 + \boxed{\beta^2}) \cdot \frac{\text{precision} \cdot \text{recall}}{(\boxed{\beta^2} \cdot \text{precision}) + \text{recall}}.$$

- If beta = 1, same importance to Precision and Recall:

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

Precision, Recall and F-score

- If beta = 1, same importance to Precision and Recall:

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

- F1 score is the harmonic mean
  - Think of it as a mean that is biased towards the smaller of P and R
    - the larger the difference, the larger the bias

P	R	F1
0.50	0.50	0.50
1.00	0.50	0.67
0.50	1.00	0.67
0.40	0.99	0.57
0.60	0.90	0.72
0.60	1.00	0.75

Precision, Recall and F-score

- Usually we calculate P, R and F1-score for each label
    - one label against all other labels

```
from sklearn.metrics import confusion_matrix, ConfusionMatrixDisplay
cm = confusion_matrix(ground_truth, predictions)
```

	precision	recall	f1-score	support
T-shirt/top	0.86	0.83	0.85	1000
Trouser	1.00	0.97	0.99	1000
Pullover	0.83	0.88	0.85	1000
Dress	0.85	0.95	0.90	1000
Coat	0.84	0.83	0.84	1000
Sandal	0.94	0.99	0.96	1000
Shirt	0.78	0.68	0.73	1000
Sneaker	0.90	0.97	0.94	1000
Bag	0.99	0.97	0.98	1000
Ankle boot	1.00	0.88	0.93	1000
accuracy			0.90	10000
macro avg	0.90	0.90	0.90	10000
weighted avg	0.90	0.90	0.90	10000

Precision, Recall and F-score

- Averages ...
  - micro: calculate metrics globally: total true positives, false negatives and false positives
    - [takes into account the support for each label]
  - macro: calculate metrics for each label, then calculate average
    - [all labels are equally important regardless of support]
  - weighted: calculate metrics for each label, then calculate weighted average based on the support
    - [takes into account the support for each label, different than micro!]
  - perhaps you only care about one (or some) labels (e.g., POSITIVE)

Precision, Recall and F-score

- Averages ...
  - micro: calculate metrics globally: total true positives, false negatives and false positives
  - macro: calculate metrics for each label, then calculate average
  - weighted: calculate metrics for each label, then calculate weighted average based on the support
  - perhaps you only care about one (or some) labels (e.g., POSITIVE)

```
>>> y_pred = [1, 1, 0]
>>> y_true = [1, 1, 1]
>>> print(classification_report(y_true, y_pred, labels=[1, 2, 3]))
              precision    recall  f1-score   support

           1       1.00      0.67      0.80         3
           2       0.00      0.00      0.00         0
           3       0.00      0.00      0.00         0

   micro avg       1.00      0.67      0.80         3
   macro avg       0.33      0.22      0.27         3
weighted avg       1.00      0.67      0.80         3
```

Precision, Recall and F-score

- Defining true positives, false positive and false negatives is sometimes not straightforward

After	-	(AM-TMP*	*
a	-	*	*
bad	-	*	*
start	-	*)	*
,	-	*	*
Treasury	-	(A1*	(A1*
bonds	-	*)	*)
were	-	*	*
buoyed	buoy	(V*)	*
by	-	(A0*	*
a	-	*	*
late	-	*	*
burst	-	*	*
of	-	*	*
buying	-	*)	*
to	-	(AM-ADV*	*
end	end	*	(V*)
modestly	-	*	(AM-MNR*
higher	-	*)	*)
.	-	*	*

Precision, Recall and F-score

```
After       -     (AM-TMP*   *
a           -     *          *
bad         -     *          *
start       -     *)         *
,           -     *          *
Treasury    -     (A1*       (A1*
bonds       -     *)         *)
were        -     *          *
buoyed      buoy  (V*)       *
by          -     (A0*       *
a           -     *          *
late        -     *          *
burst       -     *          *
of          -     *          *
buying      -     *)         *
to          -     (AM-ADV*   *
end         end   *          (V*)
modestly    -     *          (AM-MNR*
higher      -     *)         *)
.           -     *          *
```

- Defining true positives, false positive and false negatives is sometimes not straightforward
    - *After a bad start, Treasury bonds were buoyed by a late burst of buying to end modestly higher.*
    - boued
        - M-TMP, when: After a bad start
        - A1, what: Treasury bonds
        - A0, who: by a late burst of buying
        - M-ADV: to end modestly higher
    - end
        - A1, what: Treasury bonds
        - M-MNR, how: modestly higher

Precision, Recall and F-score

After	-	(AM-TMP*	*
a	-	*	*
bad	-	*	*
start	-	*)	*
,	-	*	*
Treasury	-	(A1*	(A1*
bonds	-	*)	*)
were	-	*	*
buoyed	buoy	(V*)	*
by	-	(A0*	*
a	-	*	*
late	-	*	*
burst	-	*	*
of	-	*	*
buying	-	*)	*
to	-	(AM-ADV*	*
end	end	*	(V*)
modestly	-	*	(AM-MNR*
higher	-	*)	*)
.	-	*	*

- Defining true positives, false positive and false negatives is sometimes not straightforward
  - *After a bad start, Treasury bonds were buoyed by a late burst of buying to end modestly higher.*
  - boued
    - M-TMP, when: After a bad start — *After a bad start*
    - A1, what: Treasury bonds — *bonds*
    - A0, who: by a late burst of buying — *a burst* OR *buying*
    - M-ADV: to end modestly higher — *to end modestly higher*
  - end
    - A1, what: Treasury bonds — *bond*
    - M-MNR, how: modestly higher — *higher*

Precision, Recall and F-score

- Defining true positives, false positive and false negatives is sometimes not straightforward
- Named Entities
    - Arizona State University *University*
        - Arizona (?)
    - University of Arizona
    - ASU *Repeats*
    - Arizona State *Repeats*
    - Phoenix
    - Phoenix metropolitan area (?)
    - ...

Arizona State University

From Wikipedia, the free encyclopedia

Not to be confused with University of Arizona.

Arizona State University (**ASU** or **Arizona State**) is a public research university[8] in the Phoenix metropolitan area.[9] Founded in 1885 by the 13th Arizona Territorial Legislature, ASU is one of the largest public universities by enrollment in the U.S.[10]

Not all problems are classification problems

- But we kind of need a metric that can be calculated automatically
- How good is the caption for a picture?
  - Even if you have a “gold” caption ….
    - *A person drinking water*
    - *An individual with a hat is drinking from a glass*
- How good is a translation?
- How good is a summary?
- How good is a dialogue system?
  - Interesting paper to read:
    - Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, Joelle Pineau. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. EMNLP 2016
      - https://www.aclweb.org/anthology/D16-1230/, https://vimeo.com/239251122

Not all problems are classification problems

- Machine translation quality will ideally look at
  - [translating from a source to a target language]
  - Does the translation mean the same than the source sentence? [adequacy]
  - How good (== readable) is the translation? [fluency]

- BLEU is commonly used metric to evaluate machine translation
  - It has issues, but still widely used
  - At a very high level, look at n-gram overlap between
    - reference translation (ground truth)
    - system translation (generated by some system)

Not all problems are classification problems

- n-gram: sequence of *n words*

Take this sentence: *Once you stop learning, you start dying*

unigram	bigram	trigram
Once	Once you	Once you stop
you	you stop	you stop learning
stop	stop learning	stop learning, you
learning	learning you	learning, you start
you	you start	you start dying
start	start dying	
dying		

https://towardsdatascience.com/bleu-bilingual-evaluation-understudy-2b4eab9bcfd1

Not all problems are classification problems

$$\text{Precision} = \frac{\text{No. of candidate translation words occuring in any reference translation}}{\text{Total no. of words in the candidate translation}}$$

Candidate 1: the the the the the the the.

Candidate 2: the cat is mat the on

Reference: The cat is on the mat.

Ribeiro, Marco Tulio, Tongshuang Wu, Carlos Guestrin and Sameer Singh. Beyond Accuracy: Behavioral Testing of NLP models with CheckList. ACL (2020). https://www.aclweb.org/anthology/2020.acl-main.442/

- What is an adversarial example?
- What are your two favorite failures from Table 1? Can you justify the failures?
- What are your two favorite failures from Table 2? Can you justify the failures?
- What are your two favorite failures from Table 3? Can you justify the failures?
- What do you think of the tools big companies make available in the cloud?

Dodge, Jesse, Suchin Gururangan, D. Card, Roy Schwartz and Noah A. Smith. Show Your Work: Improved Reporting of Experimental Results. EMNLP (2019). https://www.aclweb.org/anthology/D19-1224/

- What is the standard way to decide whether a model is better than another model?
- Is it important to ensure research results are reproducible?
- How do they dene computational budget?
- What do you think of their conclusions? Write around 3-5 sentences for each of the three points.