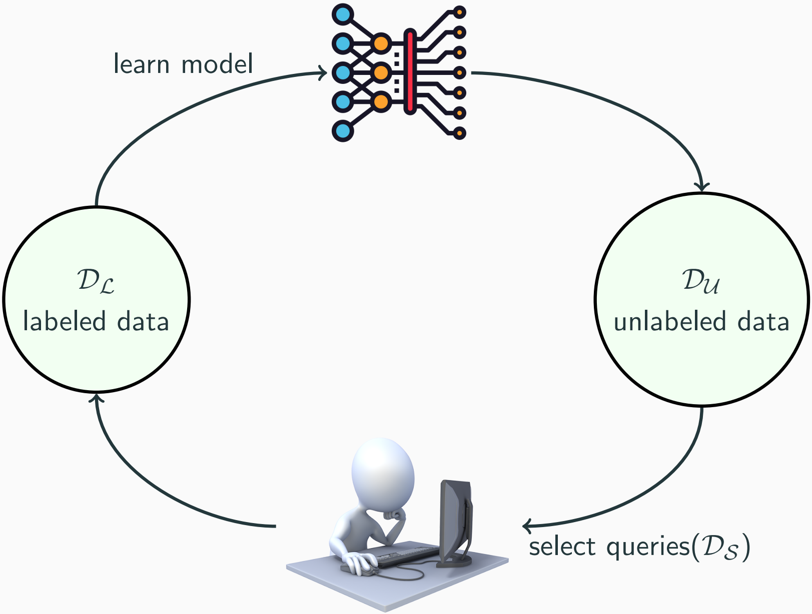
Investigating Stopping Criteria for Active Learning with Transformers

Yannick Dannies

Supervisor: Christopher Schröder

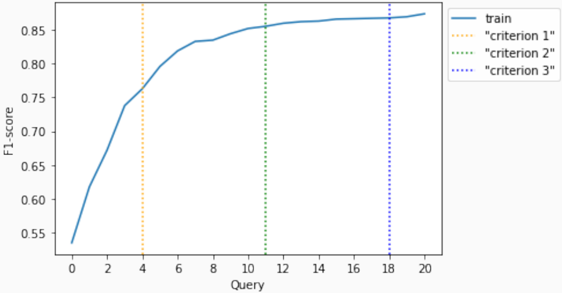
February 3, 2022

Active Learning Cycle



Stopping

- Stopping methods tell active learner when to halt
- Aggressive: stop sooner to reduce annotation cost
- Conservative: stop later to ensure higher performance



Schematic presentation of the aggressiveness in stopping

Motivation

- Previous stopping criteria not tested with transformers
  - Stabilizing Predictions (Bloodgood & Shanker, 2014)
  - Min-Error (Zhu & Hovy, 2007)
  - TotalConf (McDonald et al., 2020)
  - ...
- Instead tested with traditional models
  - SVMs, Logistic Regression, k-Nearest-Neighbors, ...

Motivation

- Transformers could behave different
  - Fine-tuning transformers is unstable (Mosbach et al.,2020)
  - Performance fluctuations may influence stopping
- Stopping similar to querying
  - Existing methods: often same principles as query strategies
  - Fitting query strategy can be used as stopping criterion

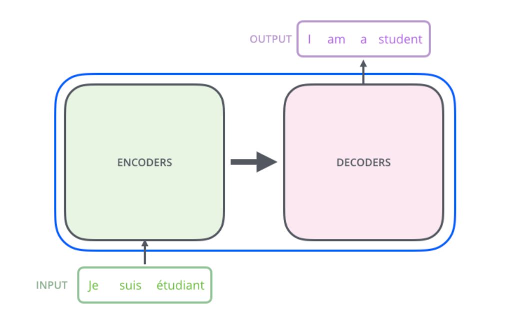
Research Questions

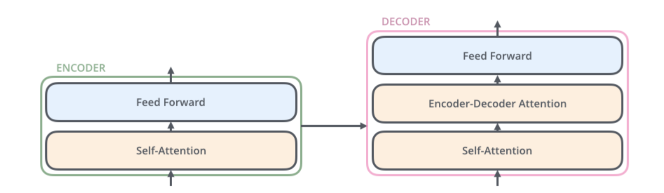
1. Are traditional stopping methods effective in combination with transformer models?
2. Is there a difference for binary and multi-class datasets?
3. How do existing hyperparameters influence stopping criteria?
4. Can *Discriminative Active Learning* (Gissin & Shalev-Shwartz, 2019) be used as a stopping method?

Transformers (Vaswani et al.,2017)

- Current state-of-the-art NLP method
- Originally aimed at sequence to sequence tasks
- Advantages
  - Attention
  - Parallel processing of input sequence
  - Can handle long-range references/dependencies in sentences

Transformers





Categories of Stopping Methods

- Uncertainty-based criteria
  - Model probability for classes is confidence/uncertainty
  - Stop when it passes a treshold
- Prediction-based criteria
  - Stop when predictions do not change anymore
  - Stop when predictions only change minimally
- Metric-based criteria
  - Look how specific metric changes
  - Stop when it stays below/above threshold

TotalConf (McDonald et al., 2020)

- Measure overall confidence in classifying all unlabeled examples
- Effectiveness of active learner stops improving when said confidence stabilises
- Stop when confidence does not increase for *i* iterations

$$\mathit{TotalConf} = \frac{\sum_{d_{\mathcal{U}}} |\ell_1 - \ell_2|}{|\mathcal{D}_{\mathcal{U}}|}$$

$|\ell_1 - \ell_2|$: margin score between two most probable labels
$|\mathcal{D}_{\mathcal{U}}|$: number of samples in unlabeled pool | $d_{\mathcal{U}}$: unlabeled example

LeastConf (McDonald et al., 2020)

- Almost identical to TotalConf
- Measures and compares confidence for queried examples
- Stop, when confidence does not increase for *i* iterations

$$\mathit{LeastConf} = \frac{\sum_{d_{\mathcal{S}}} |\ell_1 - \ell_2|}{|\mathcal{D}_{\mathcal{S}}|}$$

$|\ell_1 - \ell_2|$: margin score between two most probable labels
$|\mathcal{D}_{\mathcal{S}}|$: number of queried samples | $d_{\mathcal{S}}$: queried example

Stabilizing Predictions (Bloodgood & Shanker, 2014)

- Figure out stopping point by only looking at the predictions
- Predictions stabilized $\leftrightarrow$ performance stabilized
- Stabilization is represented by agreement of predictions
- Stop when

$$\frac{1}{w}\sum_{t}^{t-w+1}\left|agreement_t - agreement_{t-1}\right| < \epsilon$$

$agreement_t$ - agreement at time step t | w - window size specified by user

Stabilizing Predictions

- General agreement (Artstein & Poesio, 2008):

$$\mathit{agreement} = \frac{A_o - A_e}{1 - A_e}$$

o - observed | e - expected

- Kappa (Cohen, 1960):

$$A_e = \sum_{c \in \{+1,-1\}} P(c|M_1) \cdot P(c|M_2)$$

$P(c|M)$ - probability of model M choosing class c for an example

Predicted Change of F Measure (Altschuler & Bloodgood, 2019)

- Wants to estimate performance change each learning iteration
- Performance within threshold for *i* iterations $\rightarrow$ stop learning
- F measure from contingency counts:

$$F(M_t) = \frac{2tp}{2tp + fp + fn}$$

t/f - true/false | p/n - positive/negative | M_t - model at time step t

Predicted Change of F Measure

		M_t		
		+	-	Total
M_{t-1}	+	a_+	b_+	$a_+ + b_+$
	-	c_+	d_+	$c_+ + d_+$
		$a_+ + c_+$	$b_+ + d_+$	n

contingency table for true positives

		M_t		
		+	-	Total
M_{t-1}	+	a_-	b_-	$a_- + b_-$
	-	c_-	d_-	$c_- + d_-$
		$a_- + c_-$	$b_- + d_-$	n

contingency table for true negatives

Predicted Change of F Measure

$$\Delta F = \frac{2(a_+ + c_+)}{2(a_+ + c_+) + b_+ + d_+ + a_- + c_-} - \frac{2(a_+ + b_+)}{2(a_+ + b_+) + c_+ + d_+ + a_- + b_-}$$

- Assumption: new model is correct
- $a_+ = a$, $a_- = 0$, $b_+ = 0$, $b_- = b$, $c_+ = c$, $c_- = 0$, $d_+ = 0$, $d_- = d$

Predicted Change of F Measure

- Change of F measure:

$$\Delta\hat{F} = \frac{2(a+c)}{2(a+c)} - \frac{2a}{2a+b+c} = 1 - \frac{2a}{2a+b+c}$$

- Can be used to predict next change
- Stop, when $\Delta\hat{F}$ lower than threshold ϵ for i iterations
    - User can specify threshold and iteration number

Discriminative Active Learning (Gissin & Shalev-Shwartz, 2019)

- Goal: labeled pool to represent true distribution of data
- Binary classifier: $\{u, l\}$ (u:unlabeled; l:labeled)
  - Chooses x instances
  - Trains again with chosen instances now as labeled
  - Repeats k times
- Check confidence that example is unlabeled
  - High value: informative example
- Low for all unlabeled examples $\rightarrow$ labeled set close to representing true distribution

Discriminative Stopping Criteria

```
active_learning(data, actual_stopping_criterion):
    ...
    # train "supervised" classifier on
    # unlabeled/labeled pools
    clf = train_binary_classifier(data)
    predictions = clf.predict_stop_set()
    pred_probabilities = clf.predict_probs_stop_set()
    stop = actual_stopping_criterion.stop(
                predictions,
                pred_probabilities
                )
    if stop is True:
        stop_active_learning()
    ...
```

Experiment Setup

- Classifier:  (Devlin et al.,2018)
- Datasets:
  - Binary: IMDb, SST-2
  - Multi-label: AG-News, DBpedia
- Active Learning:
  - Initialization set size: 25 examples
  - Query size: 25 examples
  - Repetitions: 5
  - Query Strategies: Prediction Entropy, Contrastive Active Learning
  - Stopping Criteria: Stabilizing Predictions, TotalConf, LeastConf, Predicted Change of F measure, Discriminative Criteria

Experiment Costs

Traditional Criterion:

$$\begin{aligned} cost &= |Query\ \ Strategies| \cdot |Datasets| \\ &\cdot |Queries| \cdot |Repetitions| \\ &= 2 \cdot 4 \cdot 20 \cdot 5 \\ &= 800 \text{ x training transformer model} \end{aligned}$$

Discriminative Criterion:

$$\begin{aligned} cost &= |Query\ \ Strategies| \cdot |Datasets| \\ &\cdot |Queries| \cdot |Repetitions| \cdot 2 \\ &= 2 \cdot 4 \cdot 20 \cdot 5 \cdot 2 \\ &= 1600 \text{ x training transformer model} \end{aligned}$$

Datasets - AG-News

- Train set size: 120000 | Test set size: 7600

Text	Label
Wall St. Bears Claw Back Into the Black (Reuters) Reuters - Short-sellers. Wall Street's dwindling band of ultra-cynics, are seeing green again.	business
Dolphins Too Have Born Socialites (Reuters) Reuters - Some people are born to be the life and soul of the party – and so it seems are some dolphins	science/tech
Soon after, a financial planner stopped by his desk to drop off brochures about insurance benefits available through his employer...	world
Dreaming done, NBA stars awaken to harsh Olympic reality (AFP) - National Basketball Association players trying to win ...	sports

AG-News examples

Datasets - IMDb

- Train set size: 75000 | Test set size: 25000

Text	Label
Brilliant over-acting by Lesley Ann Warren. Best dramatic hobo lady I have ever seen, and love scenes in clothes warehouse are second to none...	pos
I liked the film. Some of the action scenes were very interesting, tense and well done. I especially liked the opening scene which had a semi truck in it...	pos
I saw this at the premiere in Melbourne. It is shallow, two-dimensional, unaffecting and, hard to believe given the subject matter, boring...	neg
This is one of the dumbest films, I've ever seen. It rips off nearly ever type of thriller and manages to make a mess of them all	neg

IMDb examples

Query Strategies

Prediction Entropy (Roy & McCallum, 2008)

- Selects highest entropy examples to reduce overall entropy

$$-\sum_{j=1}^{c} P(y_i = j|x_i) log P(y_i = j|x_i)$$

Contrastive Active Learning (Margatina et al.,2021)

- Selects instances that are highly different from their close neighbors (Kullback-Leibler)

$$\frac{1}{m}\sum_{j=1}^{m} KL(P(y_j|x_j^{knn})||P(y_i|x_i))$$

where x_j^{knn} are the m nearest neighbors of instance x_i

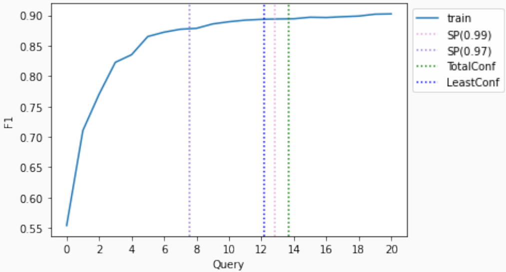
Prediction Entropy

		SP(0.99)	**SP**(0.97)	**DF**(0.05)	**DF**(0.04)	**TotalConf**	**LeastConf**	**Final-F1**
AG-News	F1	0.894	0.876	-	-	0.896	0.891	0.902
	Query	12.8	7.6	-	-	13.7	12.2	20
IMDb	F1	0.896	0.885	0.875	0.884	0.894	0.862	0.903
	Query	13.5	9.2	8.0	8.8	11.2	10.0	20

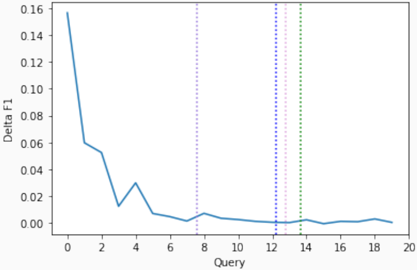
SP - Stabilizing Predictions | DF - Predicted Change of F Measure

Plot AG-News

Performance:

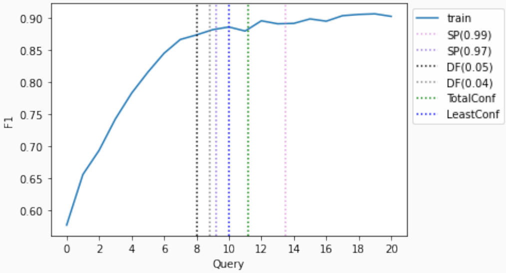


F1 Change:

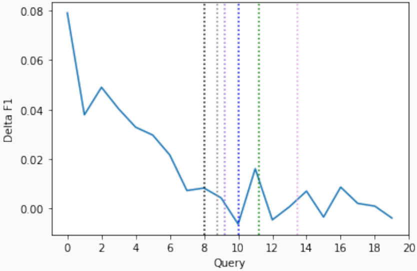


Plot IMDb

Performance:



F1 Change:



Discussion

- Criteria seem to work in general
- No visible influence from instability of transformers
- AG-News stabilizes quite early, most tested criteria stop relatively late
- Opposite for IMDb, does not really stabilize, most criteria stop early

Outlook

- Compare with traditional machine learning model
- Test discriminative stopping methods
- Test on all datasets, query strategies
- If time: create more discriminative approaches

References

- Cohen, Jacob. "A coefficient of agreement for nominal scales." Educational and psychological measurement 20, no. 1 (1960): 37-46.
- Bloodgood, Michael, and K. Vijay-Shanker. "A method for stopping active learning based on stabilizing predictions and the need for user-adjustable stopping." arXiv preprint arXiv:1409.5165 (2014).
- Altschuler, Michael, and Michael Bloodgood. "Stopping active learning based on predicted change of f measure for text classification." In 2019 IEEE 13th International Conference on Semantic Computing (ICSC), pp. 47-54. IEEE, 2019.
- McDonald, Graham, Craig Macdonald, and Iadh Ounis. "Active Learning Stopping Strategies for Technology-Assisted Sensitivity Review." In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2053-2056. 2020.
- Gissin, Daniel, and Shai Shalev-Shwartz. "Discriminative active learning." arXiv preprint arXiv:1907.06347 (2019).
- Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." In Advances in neural information processing systems, pp. 5998-6008. 2017.
- Artstein, Ron, and Massimo Poesio. "Inter-coder agreement for computational linguistics." Computational Linguistics 34, no. 4 (2008): 555-596.

References

- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).
- Margatina, Katerina, Giorgos Vernikos, Loïc Barrault, and Nikolaos Aletras. "Active learning by acquiring contrastive examples." arXiv preprint arXiv:2109.03764 (2021).
- Roy, Nicholas, and Andrew McCallum. "Toward optimal active learning through monte carlo estimation of error reduction." ICML, Williamstown 2 (2001): 441-448.
- Mosbach, Marius, Maksym Andriushchenko, and Dietrich Klakow. "On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines." arXiv preprint arXiv:2006.04884 (2020).

Image Sources

- Folie 2:
    - Deep Learning Model: flaticon.com
    - Oracle: pngwing.com
- Folie 8:
    - Illustrations: jalammar.github.io/illustrated-transformer
- Folie 20:
    - Bert: medium.com/analytics-vidhya/a-gentle-introduction-to-implementing-bert-using-hugging-face-35eb480cff3

Predicted Change of F Measure

		M_t +	M_t -	Total
Truth	+	$a_+ + c_+$	$b_+ + d_+$	n_+
	-	$a_- + c_-$	$b_- + d_-$	n_-
		$a + c$	$b + d$	n

contingency table for model vs ground truth

		M_{t-1} +	M_{t-1} -	Total
Truth	+	$a_+ + b_+$	$c_+ + d_+$	n_+
	-	$a_- + b_-$	$c_- + d_-$	n_-
		$a + b$	$c + d$	n

contingency table for previous model vs ground truth