Knowledge Distillation for Twitter Data

Mayank Musaddi | Intern ML Team

sprinklr

Brief about Distillation

Motivation

Huge models take a lot of time to infer output from the given input!

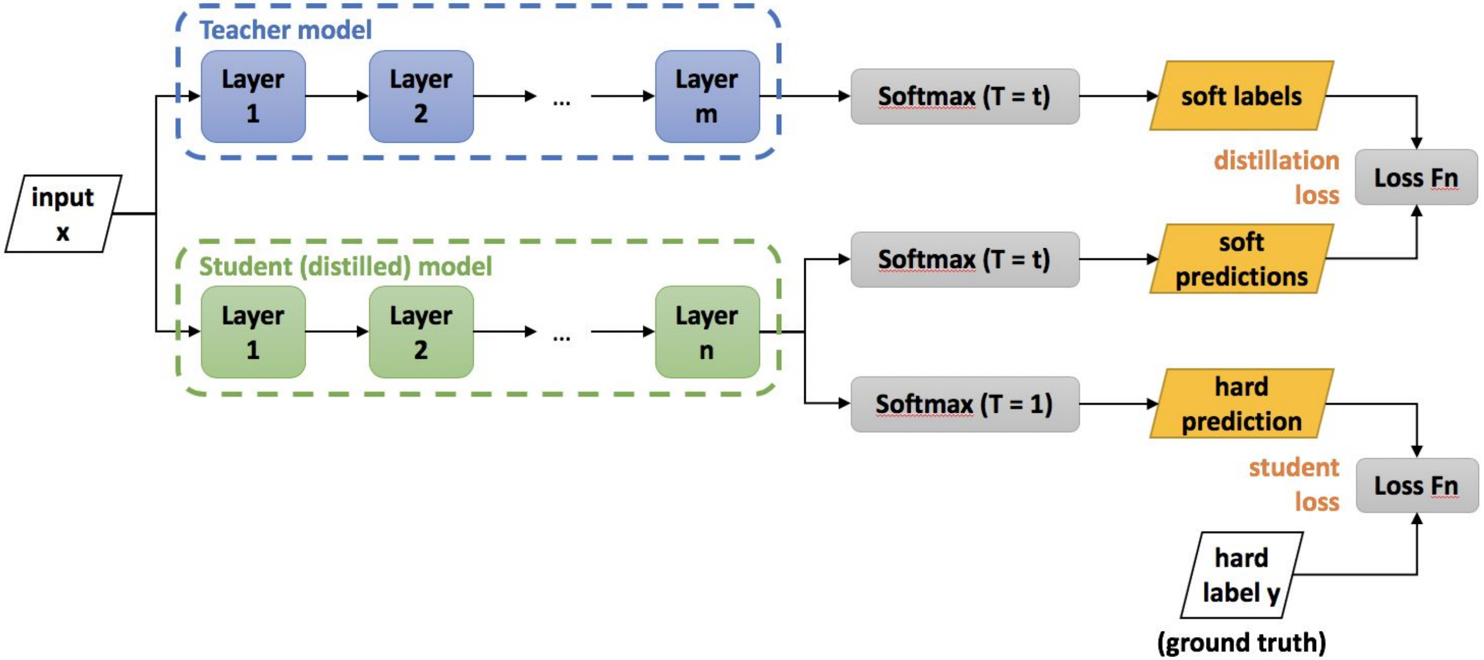
Due to large model size, there is a higher ram consumption and computation time

Idea

Train a smaller model with the help of the bigger model

Speeds up the inference time at the cost of some minimal accuracy loss

Distillation Pipeline



Teacher and Student has no architectural relationship

Distilling Roberta and XLM-Roberta

Task

Distilling the fine tuned knowledge of Twitter Specific Data of a larger teacher model to a smaller student model, for the Language Modelling Task.

Choice of Student Model :

- Raw Model half the size of Teacher
  - Needs to be first distilled on the same data the teacher was trained on, to capture semantic understanding.
  - Provides flexibility in making changes of parameters and model type.
- Pre Trained small model with same hidden state size
  - Already has captured the semantic information when trained on the same dataset which the teacher was trained on
  - Only need to distill on the fine tuned knowledge

Distilling Roberta for English Tweets

Teacher: Roberta for English Tweets (as currently under training by Sanjay)

Student: DistilRoBERTa

- 6 layers, 768 dimension and 12 heads, totalizing 82M parameters (compared to 125M parameters for RoBERTa-base)
- On average DistilRoBERTa is twice as fast as Roberta-base.

Accuracy comparison:

Task	MNLI	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE
RoBERTa-base:	87.6	91.9	92.8	94.8	63.6	91.2	90.2	78.7
DistilRoBERTta-base:	84.0	89.4	90.8	92.5	59.3	88.3	86.6	67.9

Preprocessing Corpus

The raw tweets were preprocessed and the following were pruned:

- Repetitive Tweets
- URLs and links
- User IDs
- HTML
- Punctuations
- Multiple Spaces
- Non Alphanumeric

Training Specifics

Weighed Losses

Distillation Loss (5): KL Divergence Loss between Teacher and Student Outputs

Language Modelling Loss (2): Cross Entropy Loss for MLM Task

Mean Squared Error Loss (0): Mean Squared error between output logits of Teacher and Student

Cosine Loss (1): Cosine Embedding Loss between the hidden states of Teacher and Student

Hyperparameters

Temperature : 2

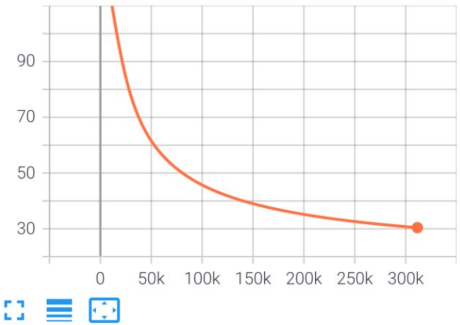
Batch Size : 5 -> 64

Learning Rate : 5e-4 with Warm up proportion of 0.05

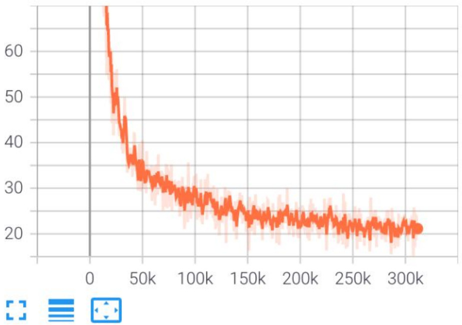
Optimizer : Adamw (ϵ 1e-6)

Loss Evaluation (Batch Size 5)

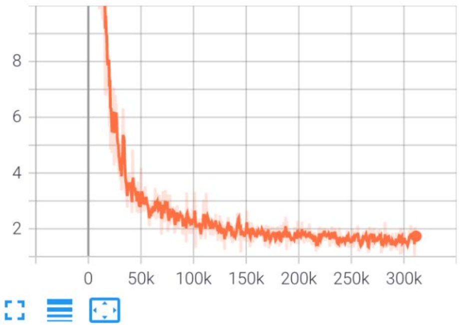
losses/cum_avg_loss_epoch
tag: losses/cum_avg_loss_epoch



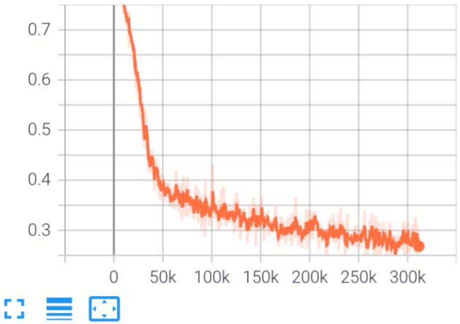
losses/loss
tag: losses/loss



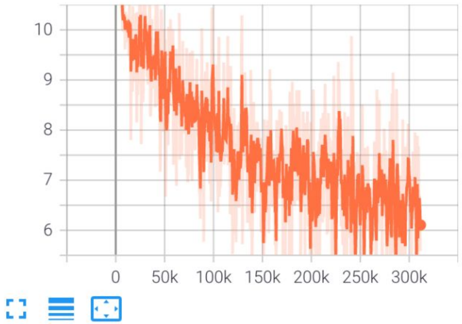
losses/loss_ce
tag: losses/loss_ce



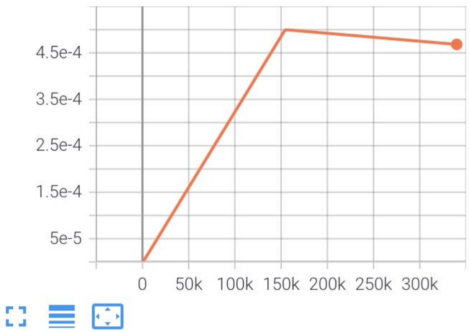
losses/loss_cos
tag: losses/loss_cos



losses/loss_mlm
tag: losses/loss_mlm

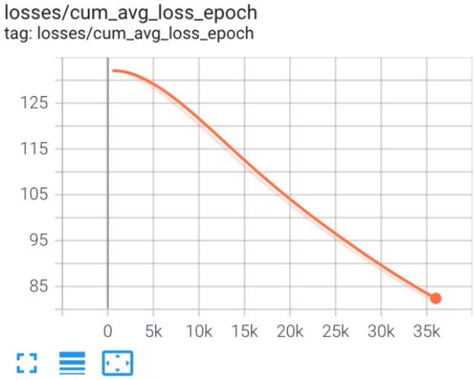


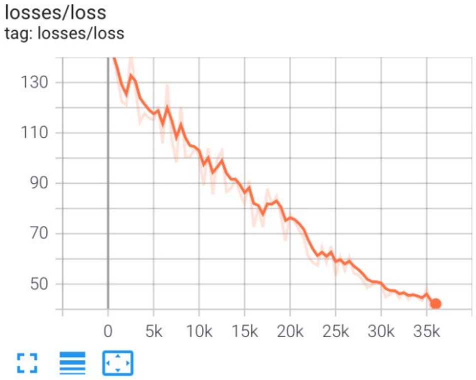
learning_rate/lr
tag: learning_rate/lr

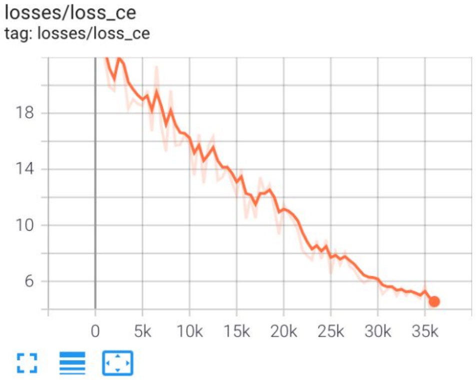


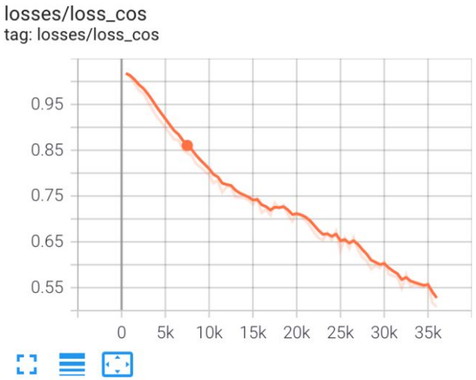
8491

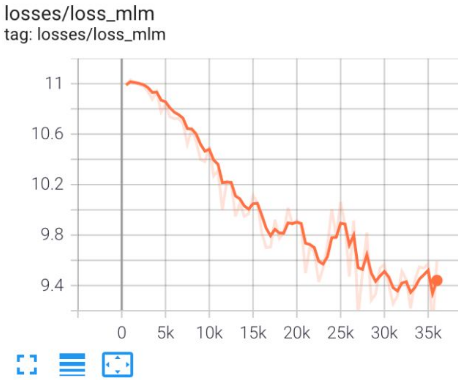
Loss Evaluation (Batch Size 64)

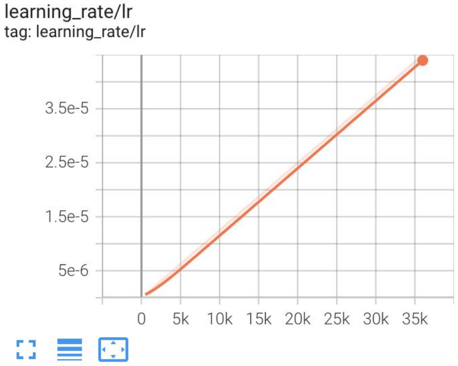












8492

Embeddings Cosine Similarity Evaluation

#	Base Words	Distilled Model	Roberta Model
		Distilled on Twitter Data	Base Model
01	ma	my (0.99755), anna (0.9975), Yeah (0.99744), f (0.99741), ay (0.99739), ta (0.99722), Yo (0.99722), na (0.99719), it (0.99716), em (0.99713)	son (0.99207), ji (0.99207), na (0.9918), min (0.99179), dad (0.99166), la (0.99166), tu (0.99159), aka (0.99155), ki (0.99147), ha (0.99146)
02	me	youe (0.99284), cancel (0.99273), i'll (0.99272), meals (0.99227), my (0.99225), maintain (0.99221), toxic (0.9922), somehow (0.99216), purse (0.99206), ma (0.9919)	us (0.97861), te (0.97637), go (0.97631), he (0.97577), ma (0.97554), am (0.97503), my (0.97456), boy (0.974), you (0.97391), boys (0.97386)
03	lmao	Lmao (0.99693), Lmaoooo (0.99401), alll (0.98697), gyallllll (0.9868), a (0.98592), smh (0.9859), coffin (0.98575), xo (0.98547), lute (0.98539), soo (0.98526)	Lmao (0.94968), lady (0.94861), mila (0.94662), loved (0.94649), logo (0.94577), doooooooooo (0.94445), lover (0.9439), soo (0.94344), lakh (0.94296), xo (0.94294)

Embeddings Cosine Similarity Evaluation

#	Base Words	Distilled Model	Roberta Model
		Distilled on Twitter Data	Base Model
04	gonna	gotta (0.99166), gave (0.99162), Gona (0.99083), dna (0.99064), goth (0.99027), wanna (0.99007), omg (0.98998), lady (0.98998), Goolge (0.98993), Gaga (0.98972)	gotta (0.97185), gods (0.9683), suddenly (0.96369), nonsense (0.96005), gents (0.9598), gives (0.9593), naturally (0.95926), goth (0.95868), grip (0.95857), guts (0.95831)
05	as	aslong (0.99745), real (0.99647), ur (0.99642), en (0.99641), Is (0.99628), is (0.9962), ir (0.99612), And (0.99602), Out (0.99601), asleep (0.996)	ta (0.98462), te (0.98453), sit (0.98437), alpha (0.98364), winner (0.98352), there (0.98342), hire (0.98333), they (0.98313), being (0.98312), want (0.98309)
06	Yeah	Okay (0.99899), Yo (0.99889), Looks (0.99862), lib (0.99859), wait (0.9985), bet (0.99825), AU (0.99823), bc (0.99818), Smith (0.99818), ships (0.99816)	Yes (0.98318), yeah (0.97725), Apparently (0.97724), Maybe (0.97654), whatever (0.97449), Wow (0.97445), Exactly (0.97437), matched (0.97436), SHIP (0.97435), locked (0.97434)

Embeddings Cosine Similarity Evaluation

#	Base Words	Distilled Model	Roberta Model
		Distilled on Twitter Data	Base Model
01	Congress	Open (0.99867), Republican (0.99865), Law (0.99864), USA (0.99863), Women (0.99861), R (0.99861), Vote (0.9986), John (0.99859), Know (0.99857), pass (0.99857)	House (0.98114), Senate (0.97894), terms (0.97667), SHIP (0.97656), Michelle (0.9757), Commission (0.97562), Government (0.97559), Jackson (0.97555), AI (0.97553), standing (0.97547)
02	Clinton	NBC (0.99912), President (0.99905), Enough (0.99903), Pres (0.99903), Christ (0.99902), Press (0.99898), criminal (0.99898), Further (0.99897), CBC (0.99897), Canada (0.99893)	Obama (0.97663), Bush (0.97557), Putin (0.97494), Williams (0.97369), Johnson (0.9716), Carter (0.9711), Brown (0.96883), Brian (0.96836), Moore (0.96835), Arsenal (0.9682)
03	Trump	Out (0.99821), New (0.99774), Vote (0.99773), Great (0.99772), GOP (0.99762), Russia (0.99761), And (0.99754), USA (0.99754), Stop (0.99753), Congress (0.99747)	realDonaldTrump (0.97479), Donald (0.97165), President (0.97092), Putin (0.97081), Obama (0.9705), Trump2020 (0.96923), Carter (0.9686), America (0.96804), Johnson (0.96708), Williams (0.96697)
04	Vote	Republican (0.9992), Women (0.99918), Final (0.99917), Right (0.99916), Law (0.99914), political (0.99908), anything (0.99908), pass (0.99907), 32 (0.99907), Further (0.99907)	Watch (0.96664), vote (0.96374), Listen (0.96323), Yes (0.96182), Comment (0.96127), Move (0.96089), Judge (0.95884), Join (0.9581), Yeah (0.95776), Ryan (0.95773)

Embeddings Cosine Similarity Evaluation

#	Base Words	Distilled Model	Roberta Model	Roberta Model
		Distilled on Twitter Data	Base Model	Fine tuned on Twitter
05	M	Rich (0.9992), Read (0.99917), Car (0.99917), Black (0.99914), Tip (0.99912), 45 (0.99912), Very (0.99911), Far (0.99909), dam (0.99909), 93 (0.99908)	P (0.98796), N (0.98671), S (0.98565), F (0.98558), J (0.98526), R (0.98519), G (0.98438), E (0.9843), H (0.98338), MF (0.98184)	MF (0.9475), MBS (0.9472), G (0.94712), May (0.94401), PM (0.94275), Im (0.94224), Me (0.94167), A (0.9412), mm (0.94088), K (0.9405)
06	f	of (0.99896), cause (0.99889), friend (0.99889), no (0.99887), holy (0.99887), ugh (0.99887), yeah (0.99887), rap (0.99886), mm (0.99884), she (0.99883)	s (0.98585), c (0.98458), g (0.98305), b (0.98274), t (0.98132), o (0.98022), a (0.98007), p (0.97972), d (0.97971), n (0.97758)	gf (0.95717), frail (0.95127), foul (0.94696), if (0.94502), fade (0.94492), tf (0.94399), fren (0.94389), farming (0.94073), af (0.93814), foster (0.93648)
07	af	f**k (0.99625), ma (0.99607), murder (0.99603), real (0.99602), Yeah (0.99578), but (0.9957), f (0.99567), die (0.99565), it (0.99557), off (0.99555)	if (0.99241), ash (0.99198), ma (0.99114), rn (0.99109), bra (0.99107), oh (0.99102), wow (0.99097), ol (0.99096), mom (0.99094), bet (0.99088)	Olave (0.95029), ban (0.94921), ash (0.9477), met (0.94744), an (0.94659), if (0.94647), air (0.94571), fly (0.94557), oil (0.94549), owl (0.94548)

Similarity Scores

Comparison of Teacher and Student Model

Measured Percentage Similarity of Words in Top 10 Nearest Embeddings for a set of words

For DistilRoBERTa distilled on 60k steps:

Percentage Similarity with RoBERTa base : 16.04%

RoBERTa fine tuned : 17.45%

For DistilRoBERTa distilled on 200k steps:

Percentage Similarity with RoBERTa base : 14.68%

RoBERTa fine tuned : 22.10%

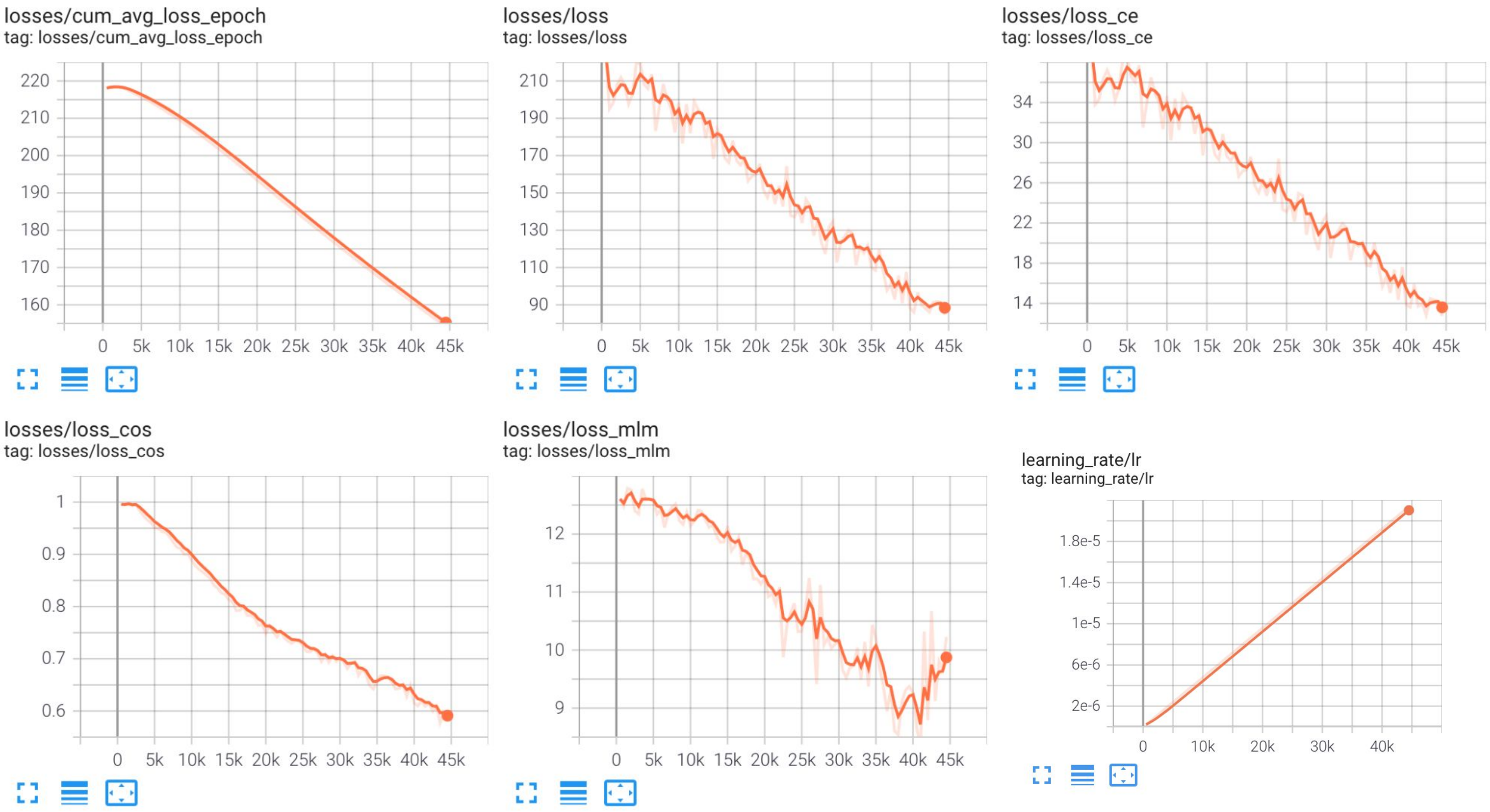
Distilling XLM-RoBERTa for Latin Tweets

Teacher: XLM-RoBERTa for Latin Tweets (as currently under training by Yash)

Student: XLM-RoBERTa with 6 layers

- 6 layers, 768 dimension and 6 heads
- Same as teacher model with 6 layers and 6 heads removed

Loss Evaluation (Batch Size 5)



8496

Task Specific Distillation

Distilling knowledge of existing Flair NER Model to DistilRoberta

Teacher: Flair NER -> BiLSTM Model

Trained on Identifying Brand, Location, Object and People from a sentence.

Student: DistilRoBERTa base -> 6 Layer Transformer Model

Experiments

- Simple Finetuning of DistilRoBERTa on the Training Corpus
- Distilling the Flair NER model with Token Classification and Distillation Loss

Results 1:

Simple Finetuning of DistilRoBERTa on Training Corpus

Epochs: 30

Batch Size: 16

Learning Rate: 5e-05

Weight Decay: 0.01

Accuray: 96%

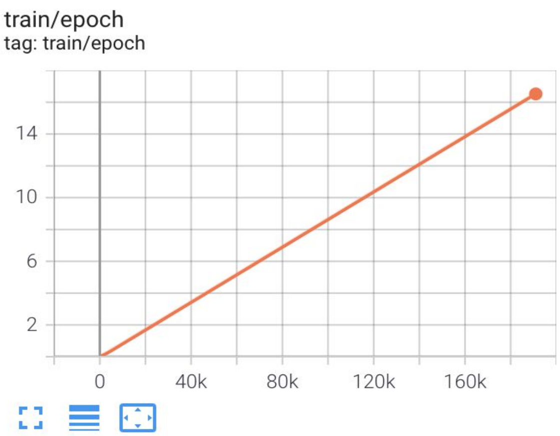
F1 Score: 0.76

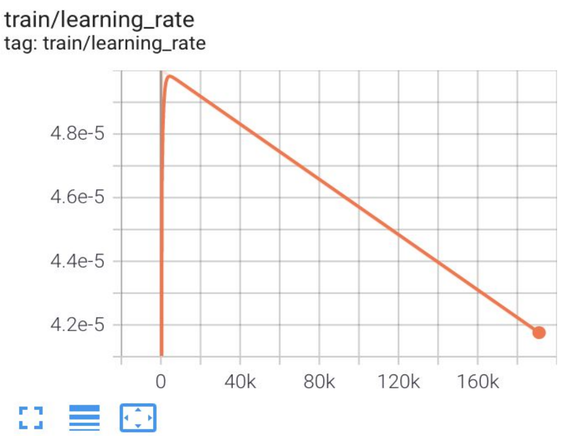
Precision: 0.763

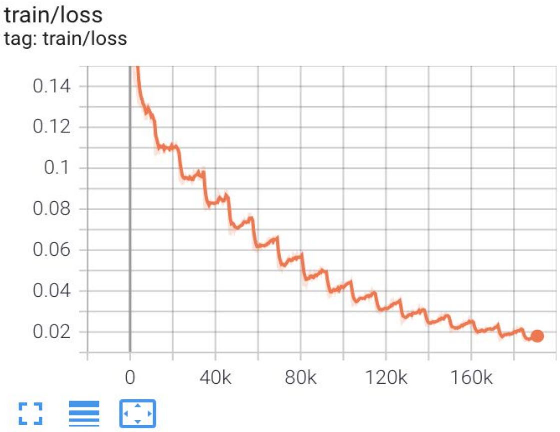
Recall: 0.756

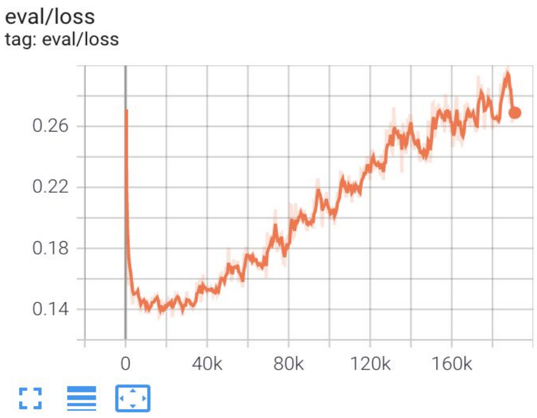
The inference time was still higher than the Flair Model:

- The student model still has high number of parameters:
- Sentences of Fixed Size as input, hence padded sentences as input
- Predicting 300 * 12 (Words * Classes) logits for forward pass of one sentence


train/epoch
tag: train/epoch
14
10
6
2
0
40k
80k
120k
160k


train/learning_rate
tag: train/learning_rate
4.8e-5
4.6e-5
4.4e-5
4.2e-5
0
40k
80k
120k
160k


train/loss
tag: train/loss
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
40k
80k
120k
160k


eval/loss
tag: eval/loss
0.26
0.22
0.18
0.14
0
40k
80k
120k
160k

8461

Results 2:

Distilling the Flair NER model with Token Classification and Distillation Loss

Epochs: 30

Batch Size: 16

Learning Rate: 5e-05

Weight Decay: 0.01

Temperature: 2

Accuray: 93.12%

F1 Score: 0.56

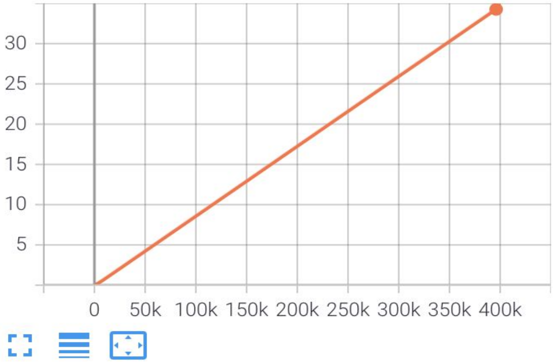
Precision: 0.51

Recall: 0.63

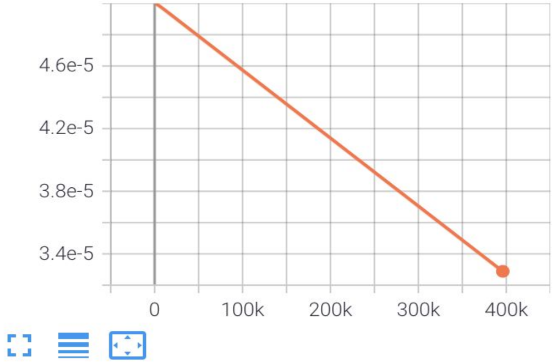
The results were analysed the fall in accuracy were attributed due to the following reasons:

- Class Imbalance
- High Padding and introduction of pad class
- Different Embedding Types

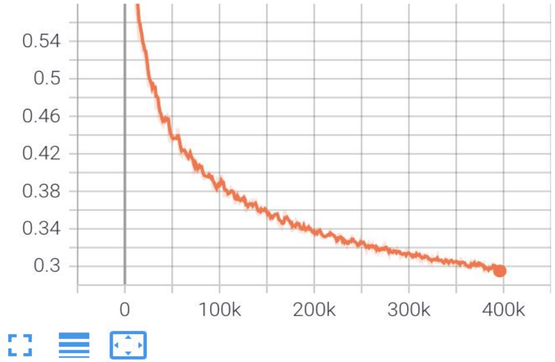
train/epoch
tag: train/epoch



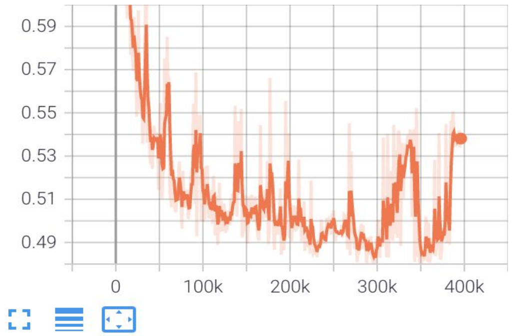
train/learning_rate
tag: train/learning_rate



train/loss
tag: train/loss



eval/loss
tag: eval/loss



8460

Conclusion

Range Of Distillation Experiments

Distillation of Language Models

Task Specific Distillation

Distillation of LSTM to Transformers

Pluggable Distillation Pipeline

An easily pluggable distillation pipeline for transformer library for the language modelling task

Distilled Language Model for Twitter Specific Data

Covering three distilled models, RoBERTa for English Tweets, XLM RoBERTa for Hindi Tweets and XLM RoBERTa for Latin Tweets

Thank you

mayank.musaddi@sprinklr.com