Dice Loss for Data-imbalanced NLP Tasks

Li et al. 2019

Minbyul Jeong

Data Mining & Information Systems Lab.
Department of Computer Science and Engineering,
College of Informatics, Korea University

1 Introduction

Motivation: Many NLP tasks (e.g., MRC, NER, POS etc..) have data imbalance issue

Task	# neg	# pos	ratio
CoNLL03 NER	170K	34K	4.98
OntoNotes5.0 NER	1.96M	239K	8.18
SQuAD 1.1 (Rajpurkar et al., 2016)	10.3M	175K	55.9
SQuAD 2.0 (Rajpurkar et al., 2018)	15.4M	188K	82.0
QUOREF (Dasigi et al., 2019)	6.52M	38.6K	169

Table 1: Number of positive and negative examples and their ratios for different data-imbalanced NLP tasks.

S1: the staff are great. (positive)
S2: the location is great but the staff are surly and unhelpful .. (hard-negative)
S3: the staff are surly and unhelpful. (easy-negative)

(1) Negative Examples outnumber Positive Examples
→ Easy-Negative Examples overwhelms the training procedure

(2) Cross Entropy criteria is accuracy-oriented objective
→ each training instance contributes equally to the objective function
(F1 score concerns about Positive examples)

Goal: Suggesting modified loss function of Dice coefficient + Focal Loss with comparable results

1 Introduction

(1) Overwhelming effect of easy-negative examples

→ model is **not sufficiently learned to distinguish** positive examples & hard-negative examples.

→ Using Focal Loss to deemphasize confident examples during training

(2) Training – Test discrepancy :

(accuracy based training – F1 score based evaluation)

→ **Each training instance contributes equal to objective function** however, F1 score concerns about positive examples

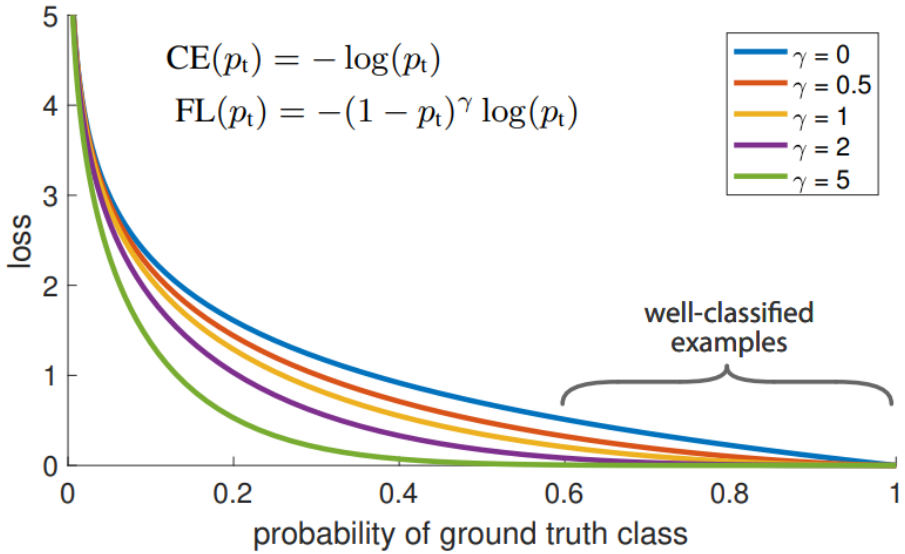
→ Dice Loss or Tversky Index to replace Cross Entropy loss function

2 Methods

$$\mathrm{CE} = -\frac{1}{N}\sum_{i}\sum_{j\in\{0,1\}} y_{i,j}\log p_{i,j}$$

$$\text{Weighted CE} = -\frac{1}{N}\sum_{i}\alpha_i\sum_{j\in\{0,1\}} y_{i,j}\log p_{i,j} \qquad \alpha_i = \lg\left(\frac{n-n_t}{n_t}+K\right)$$

$$\mathrm{FL}(p_t) = -(1-p_t)^{\gamma}\log(p_t).$$



2 Methods

Sørensen–Dice coefficient (DSC)

$$\mathrm{DSC}(A,B) = \frac{2|A \cap B|}{|A| + |B|}$$

Gauge the similarity of two sets A and B (Symmetric)

A : set contains of all positive examples which are predicted

B : set contains of all positive examples which are golden

$$\mathrm{DSC} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FN} + \mathrm{FP}} = F1 \rightarrow \frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FN}+\mathrm{FP}} = \frac{2\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}}{\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}+\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}} = \frac{2\mathrm{Pre}\times\mathrm{Rec}}{\mathrm{Pre}+\mathrm{Rec}}$$

$$\mathrm{DSC}(x_i) = \frac{p_{i1} \cdot y_{i1}}{p_{i1} + y_{i1}} \rightarrow \mathrm{DSC}(x_i) = \frac{p_{i1} \cdot y_{i1} + \gamma}{p_{i1} + y_{i1} + \gamma} \rightarrow \mathrm{DL} = \frac{1}{N}\sum_i \left[1 - \frac{2p_{i1}y_{i1} + \gamma}{p_{i1}^2 + y_{i1}^2 + \gamma}\right]$$

Negative examples contributes to training

Faster convergence

2 Methods

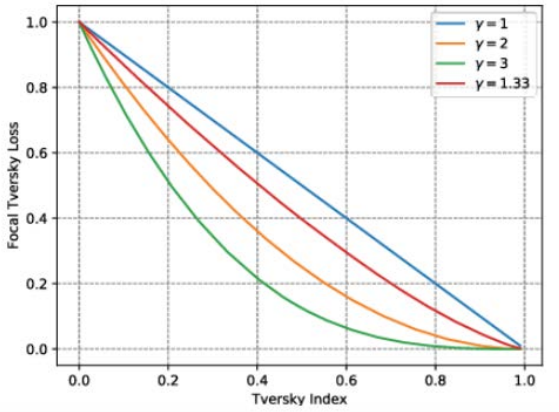
Tversky Index (TI)

$$\mathrm{TI} = \frac{|A \cap B|}{|A \cap B| + \alpha|A \backslash B| + \beta|B \backslash A|}$$

Gauge the similarity of two sets A and B (Asymmetric)
Flexibility in controlling tradeoff between FN and FP

$$\mathrm{TL} = \frac{1}{N}\sum_i \left[1 - \frac{p_{i1}y_{i1}}{p_{i1}y_{i1} + \alpha\, p_{i1}y_{i0} + \beta\, p_{i0}y_{i1}}\right]$$

Focal Tversky Loss (FTL)



$$FTL_c = \sum_c (1 - TI_c)^{1/\gamma}$$

2 Methods

Sørensen–Dice coefficient (DSC)

$$\mathrm{DSC}(A,B) = \frac{2|A \cap B|}{|A| + |B|}$$

Gauge the similarity of two sets A and B (Symmetric)

A : set contains of all positive examples which are predicted

B : set contains of all positive examples which are golden

$$\mathrm{DSC} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FN} + \mathrm{FP}} = F1 \rightarrow \frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FN}+\mathrm{FP}} = \frac{2\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}}{\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}+\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}} = \frac{2\mathrm{Pre}\times\mathrm{Rec}}{\mathrm{Pre}+\mathrm{Rec}}$$

$$\mathrm{DSC}(x_i) = \frac{p_{i1} \cdot y_{i1}}{p_{i1} + y_{i1}}$$

$$+ \qquad \rightarrow \quad \mathrm{DSC}(x_i) = \frac{(1-p_{i1})p_{i1} \cdot y_{i1}}{(1-p_{i1})p_{i1} + y_{i1}}$$

$$\mathrm{FL}(p_t) = -(1-p_t)^{\gamma}\log(p_t).$$

3 Evaluation

(1) POS tagging dataset

	CTB5			CTB6			UD1.4		
Model	P	R	F	P	R	F	P	R	F
Joint-POS(Sig)(Shao et al., 2017)	93.68	94.47	94.07	-	-	90.81	89.28	89.54	89.41
Joint-POS(Ens)(Shao et al., 2017)	93.95	94.81	94.38	-	-	-	89.67	89.86	89.75
Lattice-LSTM(Zhang and Yang, 2018)	94.77	95.51	95.14	92.00	90.86	91.43	90.47	89.70	90.09
BERT-Tagger(Devlin et al., 2018)	95.86	96.26	**96.06**	94.91	94.63	**94.77**	95.42	94.17	**94.79**
BERT+WeightCE	96.45	96.41	96.43(+0.37)	95.34	96.22	95.78(+1.01)	96.09	97.08	96.58(+1.79)
BERT+FL	96.11	97.42	96.76(+0.70)	95.80	95.08	95.44(+0.67)	96.33	95.85	96.81(+2.02)
BERT+DL	96.77	98.87	97.81(+1.75)	94.08	96.12	95.09(+0.32)	96.10	97.79	96.94(+2.15)
BERT+DSC	97.10	98.75	**97.92(+1.86)**	96.29	96.85	**96.57(+1.80)**	96.24	97.73	**96.98(+2.19)**

Table 3: Experimental results for POS datasets. WeightCE denotes weighted cross-entropy, FL denotes focal loss, DL denotes dice loss and DSC denotes adjusted dice coefficient.

3 Evaluation

(2) NER dataset

English CoNLL 2003			
Model	P	R	F
ELMo(Peters et al., 2018)	-	-	92.22
CVT(Clark et al., 2018)	-	-	92.6
BERT-Tagger(Devlin et al., 2018)	-	-	92.8
BERT-MRC(Li et al., 2019)	92.33	94.61	**93.04**
BERT-MRC+WeightCE	93.32	92.78	93.05(+0.01)
BERT-MRC+FL	93.13	93.09	93.11(+0.06)
BERT-MRC+DL	93.22	93.12	93.17(+0.12)
BERT-MRC+DSC	**93.41**	**93.25**	**93.33(+0.29)**

English OntoNotes 5.0			
Model	P	R	F
CVT (Clark et al., 2018)	-	-	88.8
BERT-Tagger (Devlin et al., 2018)	90.01	88.35	89.16
BERT-MRC(Li et al., 2019)	92.98	89.95	**91.11**
BERT-MRC+WeightCE	89.99	92.92	91.43(+0.32)
BERT-MRC+FL	90.13	92.34	91.22(+0.11)
BERT-MRC+DL	91.70	92.06	91.88(+0.77)
BERT-MRC+DSC	**91.59**	**92.56**	**92.07(+0.96)**

Chinese MSRA			
Model	P	R	F
Lattice-LSTM (Zhang and Yang, 2018)	93.57	92.79	93.18
BERT-Tagger (Devlin et al., 2018)	94.97	94.62	94.80
Glyce-BERT (Wu et al., 2019)	95.57	95.51	95.54
BERT-MRC(Li et al., 2019)	96.18	95.12	**95.75**
BERT-MRC+WeightCE	96.08	94.79	95.43(-0.32)
BERT-MRC+FL	95.45	95.89	95.67(-0.08)
BERT-MRC+DL	96.20	96.68	96.44(+0.69)
BERT-MRC+DSC	**96.67**	**96.77**	**96.72(+0.97)**

Chinese OntoNotes 4.0			
Model	P	R	F
Lattice-LSTM (Zhang and Yang, 2018)	76.35	71.56	73.88
BERT-Tagger (Devlin et al., 2018)	78.01	80.35	79.16
Glyce-BERT (Wu et al., 2019)	81.87	81.40	80.62
BERT-MRC(Li et al., 2019)	82.98	81.25	**82.11**
BERT-MRC+WeightCE	83.45	83.87	83.66(+1.55)
BERT-MRC+FL	83.63	82.97	83.30(+1.19)
BERT-MRC+DL	83.97	84.05	84.01(+1.90)
BERT-MRC+DSC	**84.22**	**84.72**	**84.47(+2.36)**

3 Evaluation

(2) NER dataset

English CoNLL 2003			
Model	P	R	F
ELMo(Peters et al., 2018)	-	-	92.22
CVT(Clark et al., 2018)	-	-	92.6
BERT-Tagger(Devlin et al., 2018)	-	-	92.8
BERT-MRC(Li et al., 2019)	92.33	94.61	**93.04**
BERT-MRC+WeightCE	93.32	92.78	93.05(+0.01)
BERT-MRC+FL	93.13	93.09	93.11(+0.06)
BERT-MRC+DL	93.22	93.12	93.17(+0.12)
BERT-MRC+DSC	**93.41**	**93.25**	**93.33(+0.29)**

English OntoNotes 5.0			
Model	P	R	F
CVT (Clark et al., 2018)	-	-	88.8
BERT-Tagger (Devlin et al., 2018)	90.01	88.35	89.16
BERT-MRC(Li et al., 2019)	92.98	89.95	**91.11**
BERT-MRC+WeightCE	89.99	92.92	91.43(+0.32)
BERT-MRC+FL	90.13	92.34	91.22(+0.11)
BERT-MRC+DL	91.70	92.06	91.88(+0.77)
BERT-MRC+DSC	**91.59**	**92.56**	**92.07(+0.96)**

Chinese MSRA			
Model	P	R	F
Lattice-LSTM (Zhang and Yang, 2018)	93.57	92.79	93.18
BERT-Tagger (Devlin et al., 2018)	94.97	94.62	94.80
Glyce-BERT (Wu et al., 2019)	95.57	95.51	95.54
BERT-MRC(Li et al., 2019)	96.18	95.12	**95.75**
BERT-MRC+WeightCE	96.08	94.79	95.43(-0.32)
BERT-MRC+FL	95.45	95.89	95.67(-0.08)
BERT-MRC+DL	96.20	96.68	96.44(+0.69)
BERT-MRC+DSC	**96.67**	**96.77**	**96.72(+0.97)**

Chinese OntoNotes 4.0			
Model	P	R	F
Lattice-LSTM (Zhang and Yang, 2018)	76.35	71.56	73.88
BERT-Tagger (Devlin et al., 2018)	78.01	80.35	79.16
Glyce-BERT (Wu et al., 2019)	81.87	81.40	80.62
BERT-MRC(Li et al., 2019)	82.98	81.25	**82.11**
BERT-MRC+WeightCE	83.45	83.87	83.66(+1.55)
BERT-MRC+FL	83.63	82.97	83.30(+1.19)
BERT-MRC+DL	83.97	84.05	84.01(+1.90)
BERT-MRC+DSC	**84.22**	**84.72**	**84.47(+2.36)**

3 Evaluation

(3) Machine Reading Comprehension dataset

	SQuAD v1.1		SQuAD v2.0		QuoRef	
	EM	F1	EM	F1	EM	F1
Model						
QANet (Yu et al., 2018b)	73.6	82.7	-	-	34.41	38.26
BERT (Devlin et al., 2018)	84.1	90.9	78.7	81.9	58.44	64.95
BERT+FL	84.67(+0.57)	91.25(+0.35)	78.92(+0.22)	82.20(+0.30)	60.78(+2.34)	66.19(+1.24)
BERT+DL	84.83(+0.73)	91.86(+0.96)	78.99(+0.29)	82.88(+0.98)	62.03(+3.59)	66.88(+1.93)
BERT+DSC	**85.34(+1.24)**	**91.97(+1.07)**	**79.02(+0.32)**	**82.95(+1.05)**	**62.44(+4.00)**	**67.52(+2.57)**
XLNet (Yang et al., 2019)	88.95	94.52	86.12	88.79	64.52	71.49
XLNet+FL	88.90(-0.05)	94.55(+0.03)	87.04(+0.92)	89.32(+0.53)	65.19(+0.67)	72.34(+0.85)
XLNet+DL	89.13(+0.18)	95.36(+0.84)	87.22(+1.10)	89.44(+0.65)	65.77(+1.25)	72.85(+1.36)
XLNet+DSC	**89.79(+0.84)**	**95.77(+1.25)**	**87.65(+1.53)**	**89.51(+0.72)**	**65.98(+1.46)**	**72.90(+1.41)**

Table 5: Experimental results for MRC task. FL denotes focal loss, DL denotes dice loss and DSC denotes adjusted dice coefficient.

3 Evaluation

(4) Paraphrase Identification dataset

Identifying whether two sentences have the same meaning or not

Model	MRPC F1	QQP F1
BERT (Devlin et al., 2018)	88.0	91.3
BERT+FL	88.43(+0.43)	91.86(+0.56)
BERT+DL	88.71(+0.71)	91.92(+0.62)
BERT+DSC	**88.92(+0.92)**	**91.57(+0.27)**
XLNet (Yang et al., 2019)	89.2	91.8
XLNet+FL	89.25(+0.05)	91.19(-0.61)
XLNet+DL	89.33(+0.13)	91.37(-0.43)
XLNet+DSC	**89.78(+0.58)**	**92.53(+0.73)**

Table 6: Experimental results for PI task. FL denotes focal loss, DL denotes dice loss and DSC denotes adjusted dice coefficient.

4 Ablation Studies

Difference between Accuracy-oriented tasks & hyper-parameters of alpha

	SST-2	SST-5
Model	Acc	Acc
BERT+CE	94.9	55.57
BERT+DL	94.37	54.63
BERT+DSC	94.84	55.19

Table 7: The effect of dice loss on sentiment classification. BERT+CE refers to fine-tune BERT model and set cross-entropy as the training objective.

α	**Chinese Onto4.0**	**English QuoRef**
$\alpha = 0.1$	80.13	63.23
$\alpha = 0.2$	81.17	63.45
$\alpha = 0.3$	84.22	65.88
$\alpha = 0.4$	84.52	**68.44**
$\alpha = 0.5$	84.47	67.52
$\alpha = 0.6$	**84.67**	66.35
$\alpha = 0.7$	81.81	65.09
$\alpha = 0.8$	80.97	64.13
$\alpha = 0.9$	80.21	64.84

Table 8: The effect of α in Tversky Index.