Bootstrapping - noisy labels

January 27, 2016

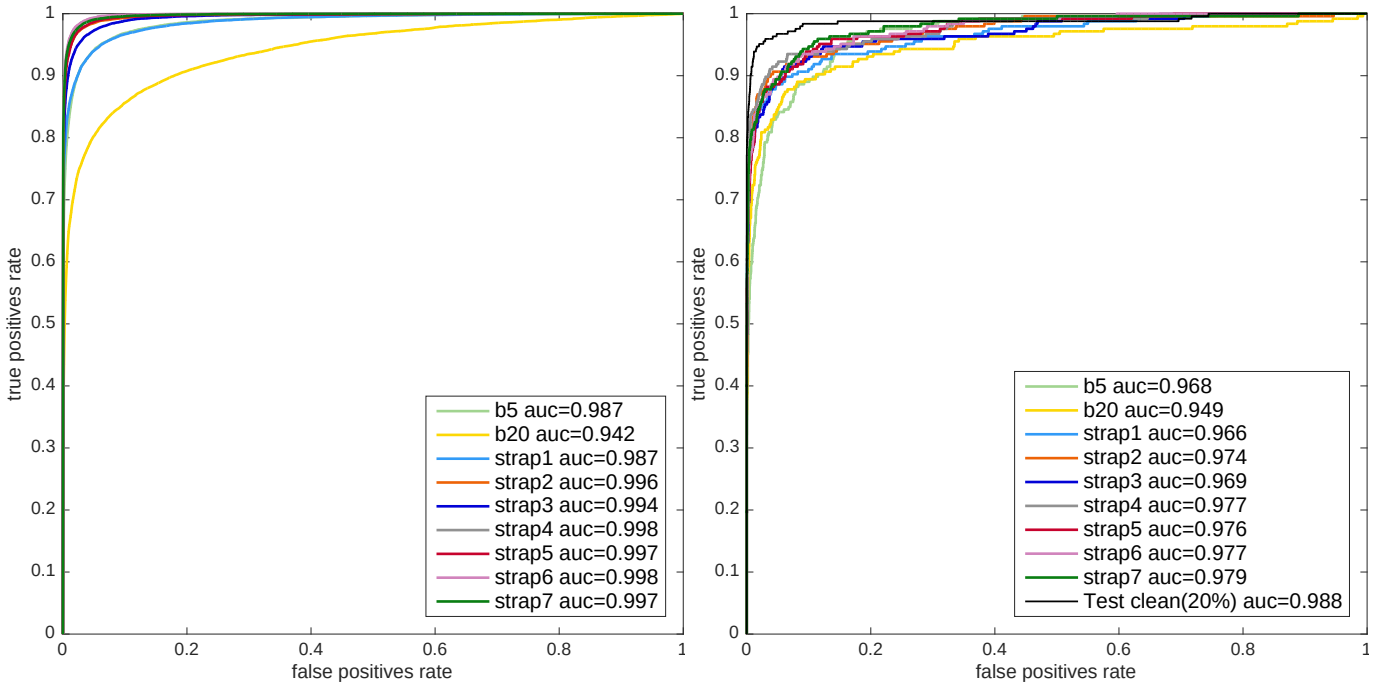
Bootstraping the baseline 20%[1] (th=30)

Update rules : relabelling the training set

(1) new decision score $= \frac{1}{2}\Big($ CNN score + previous score $\Big)$, where CNN score $= 100\ \hat{\mathbb{P}}(\text{label} = 1)$
Update training set labels with a threshhold at $th^{(i+1)} = \frac{th^{(i)}+50}{2}$

(2: mean) new decision score $= mean($CNN score , all previous scores $)$
Update training set labels with a threshhold at $th = 50$

Performance (1)



(a) Training set

(b) Test set

[1] Balanced dataset with 20% of positive samples

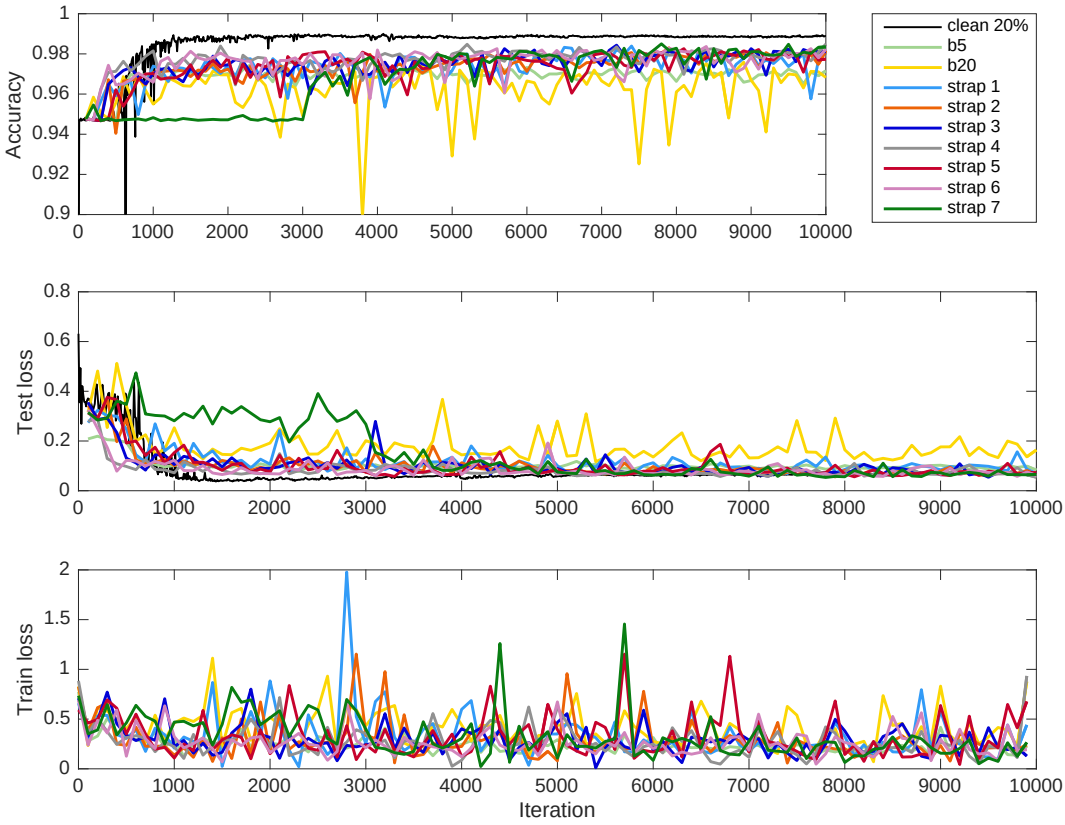
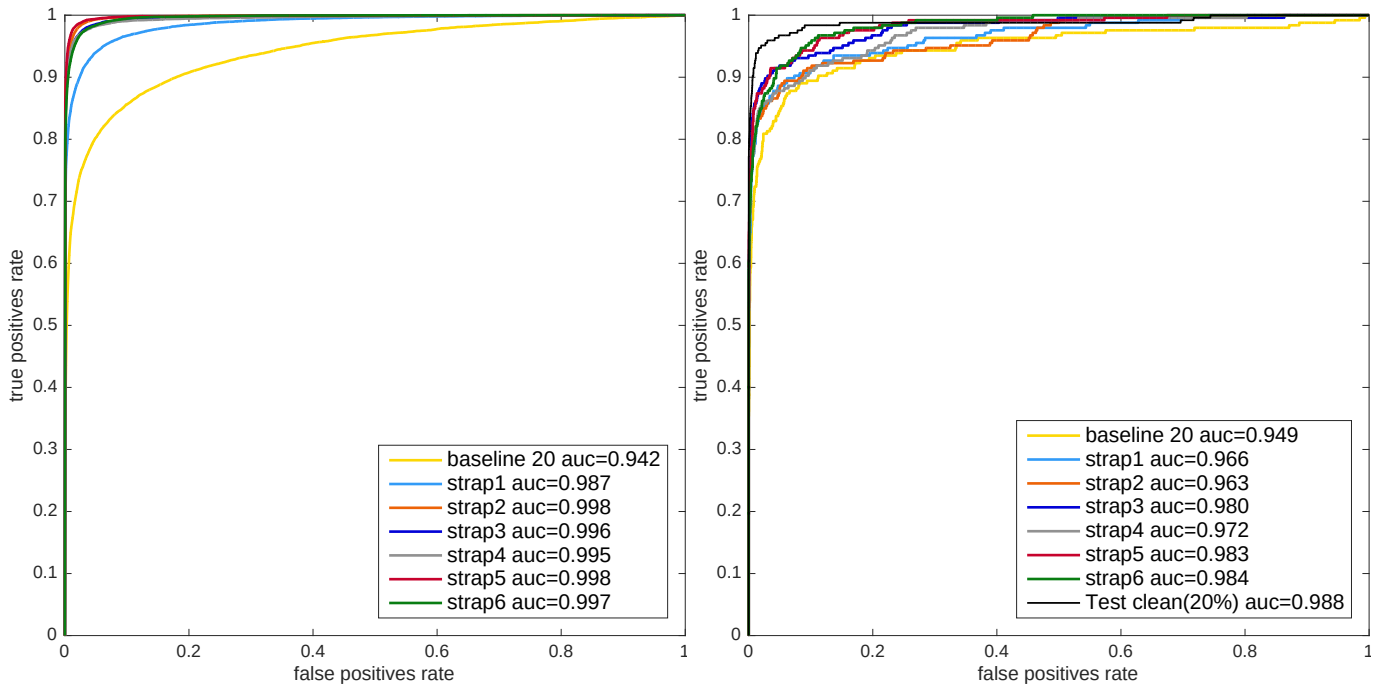


Figure 1: Accuracy & losses - b5: baseline with original dataset (5% of positive samples), b20: baseline with balanced dataset, strap i: ith iteration of bootstrapping

Performance (2: mean)



(a) Training set

(b) Test set

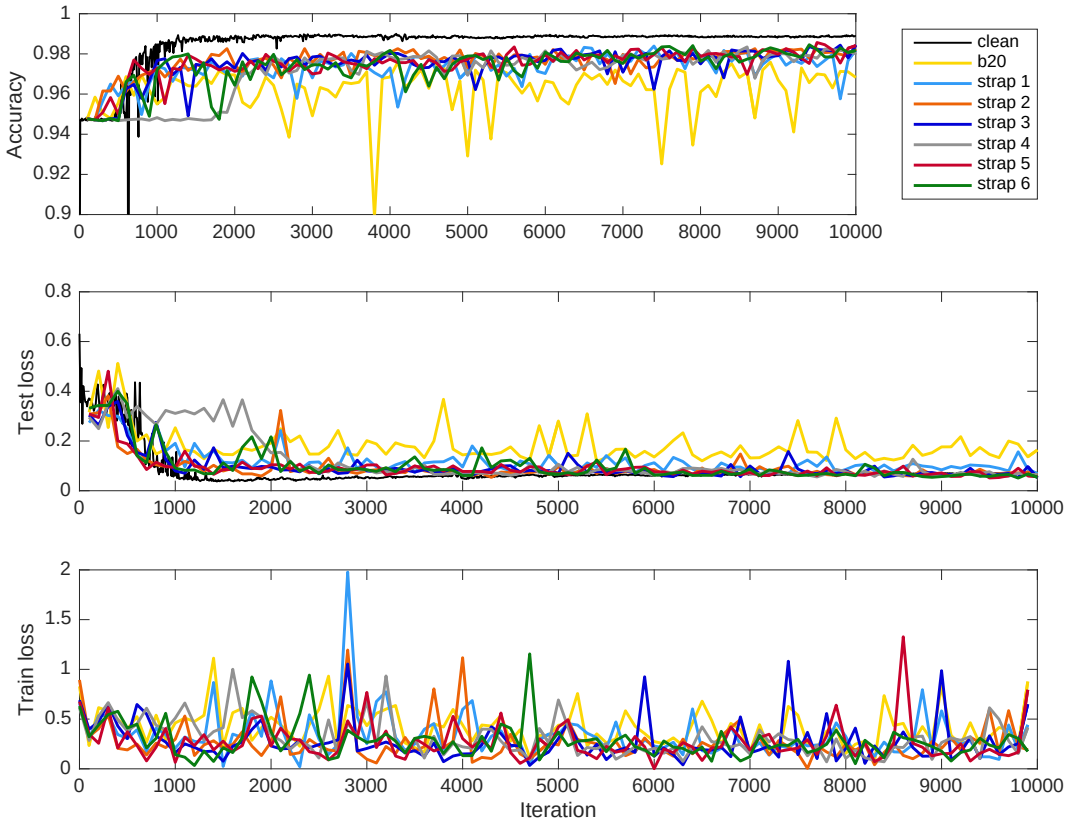


Figure 2: Accuracy & losses

Labels update

	baseline 20	strap 1	strap 2	strap3	strap 4	strap 5	strap 6
strap 1	1039//3325						
strap 2	963//4659	368//1778					
strap 3	971//4811	441//1995	408//552				
strap 4	902//4960	426//2198	450//812	344//562			
strap 5	923//5014	445//2250	448//843	330//581	319//352		
strap 6	1055//4857	640//2156	768//874	708//670	520//264	594//305	
strap 7	950//5119	552//2435	660//1133	607//936	405//516	496//574	127//494

Table 1: $|0 \to 1|//|1 \to 0|$

Labels update - mean

	baseline 20	strap 1	strap 2	strap3	strap4	strap 5
strap 1	1039//3325					
strap 2	556//4983	78//2219				
strap 3	554//4857	89//2106	392//268			
strap 4	493//5016	63//2300	336//432	199//419		
strap 5	617//4933	122//2152	407//296	268//281	402//195	
strap 6	908//4827	434//2067	896//388	736//352	889//285	628//231

We drop the initial score when computing the new score for the 5^{th} bootstrap iteartion and consider only the previous score for the 6^{th} iteration

Confusion matrices

q00= $\mathbb{P}(\text{true label} = 0|\text{predicted label} = 0)$
q11= $\mathbb{P}(\text{true label} = 1|\text{predicted label} = 1)$
Evaluated on a subset of the training set (606 negatives / 214 positives):

	q00	q11
baseline 5%	.9788	.7065
baseline 20%	.9811	.7065
strap 1	.9879	.8375
strap 2	.9837	.8973
strap 3	.9881	.8993
strap 4	.9867	.9366
strap 5	.9867	.9433
strap 6	.9881	.9241
strap 7	.9868	.9708

Mean	q00	q11
strap 2	.9794	.9209
strap 3	.9823	.9155
strap 4	.9809	.9281
strap 5	.9795	.9275
strap 6	.9866	.9110

False positives & false negatives in the training set

	baseline	baseline 20%	strap 1	strap 2	strap 3	strap 4	strap 5	strap 6	strap 7
size	381942	79135	..	..	..	..	..	..	..
(1)%	4.14%	20%	17.11%	15.33%	15.28%	14.87%	14.83%	15.2%	14.73%
FP	1461	1590	513	729	704	604	1176	264	403
FN	3919	4036	2198	670	1052	596	456	941	789

Mean:

	baseline	baseline 20%	strap 1	strap 2	strap 3	strap 4	strap 5	strap 6
size	381942	79135	..	..				
(1)%	4.14%	20%	17.11%	14.41%	14.56%	14.28%	14.55%	15.05%
FP	1461	1590	513	492	540	244	839	361
FN	3919	4036	2198	618	783	1179	274	1169

(Mis)classified samples from the training set

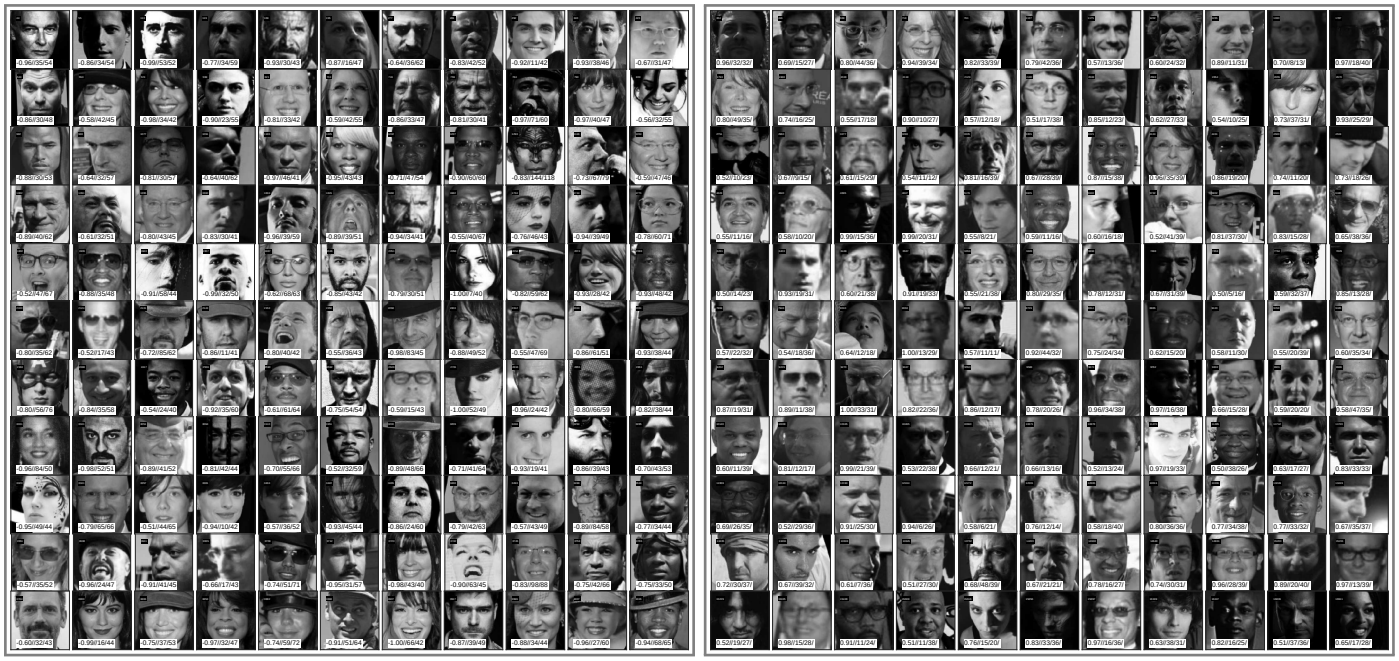
The baseline:

(a) False negatives - 85% mislabelled

(b) False positives - 46% mislabelled

Figure 3: Annotation : CNN score / Initial classifier score

1st iteration:



(a) False negatives - 78% mislabelled

(b) False positives - 42% mislabelled

Figure 4: Strap 1: annotation = CNN score//scores history

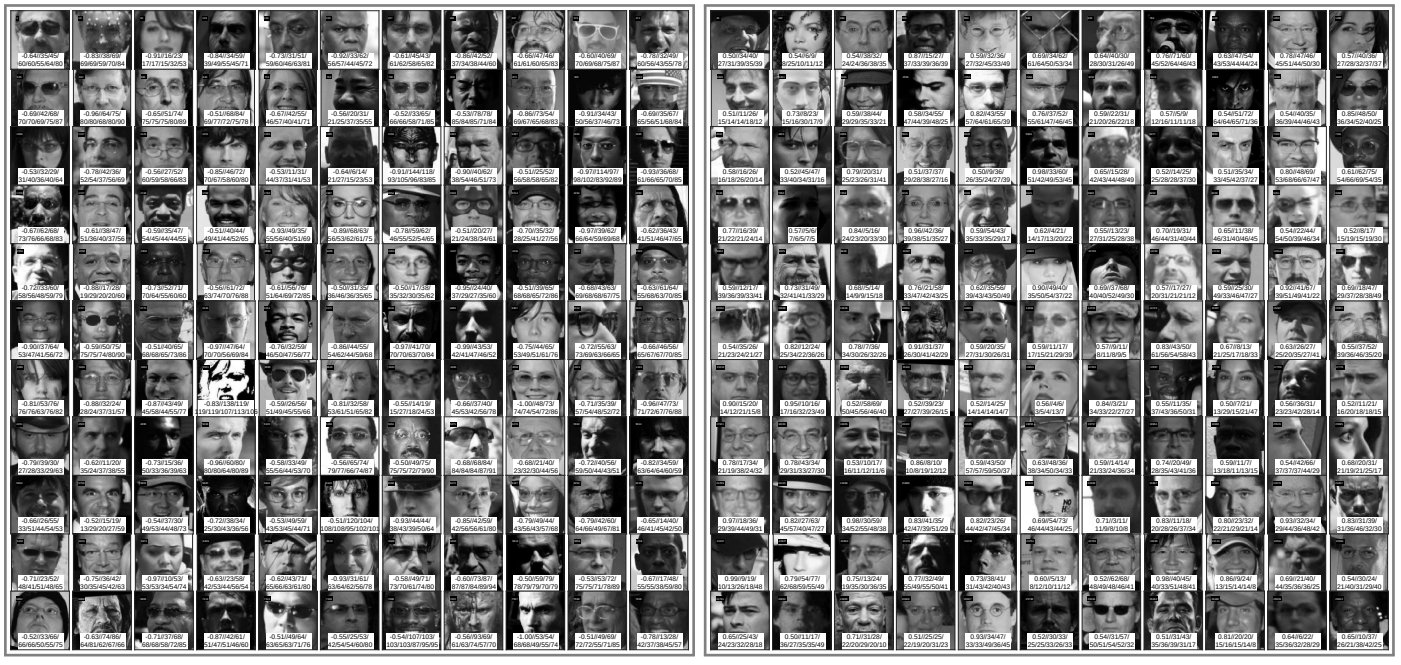
7th iteration:

(a) False negatives - 39% mislabelled

(b) False positives - 55% mislabelled

Figure 5: Strap 7: annotation = CNN score//scores history

6th -mean iteration:



(a) False negatives - 40% mislabelled

(b) False positives - 43% mislabelled

Figure 6: Strap 5 -mean: annotation = CNN score//scores history