For any distribution P on any concept space $\mathcal{H}$ and any realizable distribution $\mathcal{D}$ on a space of instances we have, for $0 < \delta \leq 1$, and $0 < \gamma \leq 1$, that with probability at least $1 - \delta$ over the choice of sample S of m instances, that with probability at least $1 - \gamma$ over the choice of h:

$$\ln\left(1 - \epsilon(h)\right) < \frac{\ln \frac{1}{P(C(S))} + \ln m + \ln \frac{1}{\delta} + \ln \frac{1}{\gamma}}{m-1}$$

where

- $C(S)$ is the set of hypotheses in $\mathcal{H}$ consistent with the sample S and $P(C(S)) = \sum_{h \in C(S)} P(h)$
- h is sampled from $Q(h) = \begin{cases} \frac{P(h)}{P(C(S))} & \text{if } h \in C(S) \\ 0 & \text{if } h \notin C(S) \end{cases}$

Proof: Essentially the same as that in (DA McAllister, 1999)

Following (G Valle-Perez et al., 2019), we make the following argument:

If SGD-trained neural networks:

- Reach 0 training error (realizability + trainability)
- Sample the 0-training-error region of parameter space close to uniformly within a bounded domain (unbiasedness in parameter space)

$\Rightarrow$

SGD-trained neural networks approximate the above model, with prior $P(h)$ determined by the parameter-function map upon uniform sampling of inputs.

The kernel in th
but is computat

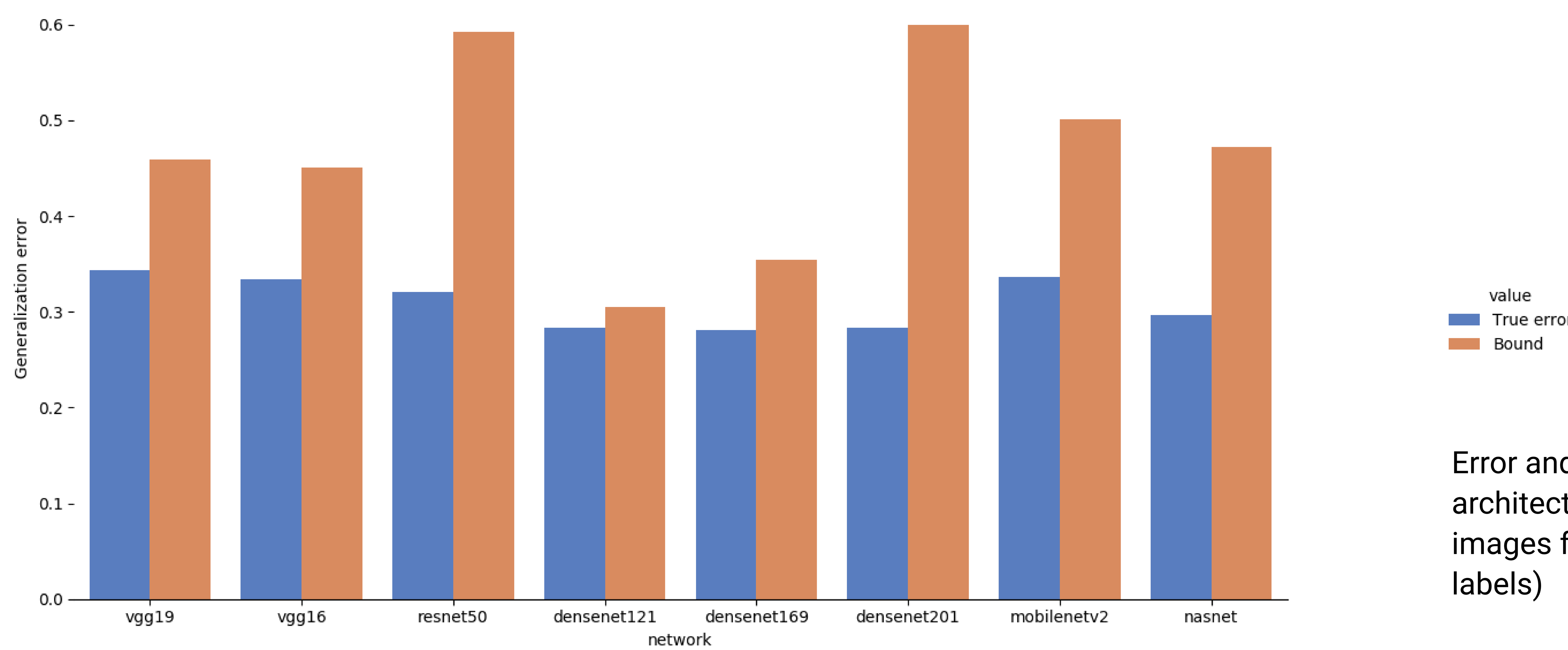
For other archit
al., 2019):

where h_{θ_i} is the
samples $\theta_i \sim$
the network

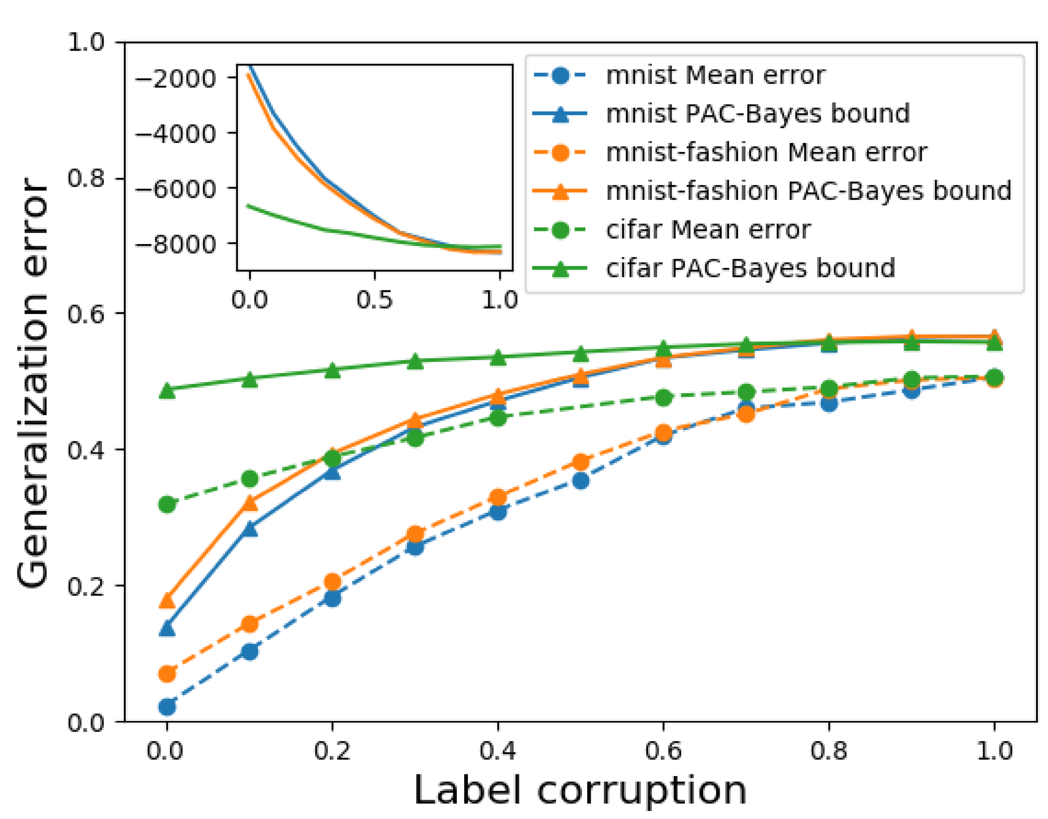
We take M to b
We use this em

Experiments: Tighter bounds on deep learning architectures

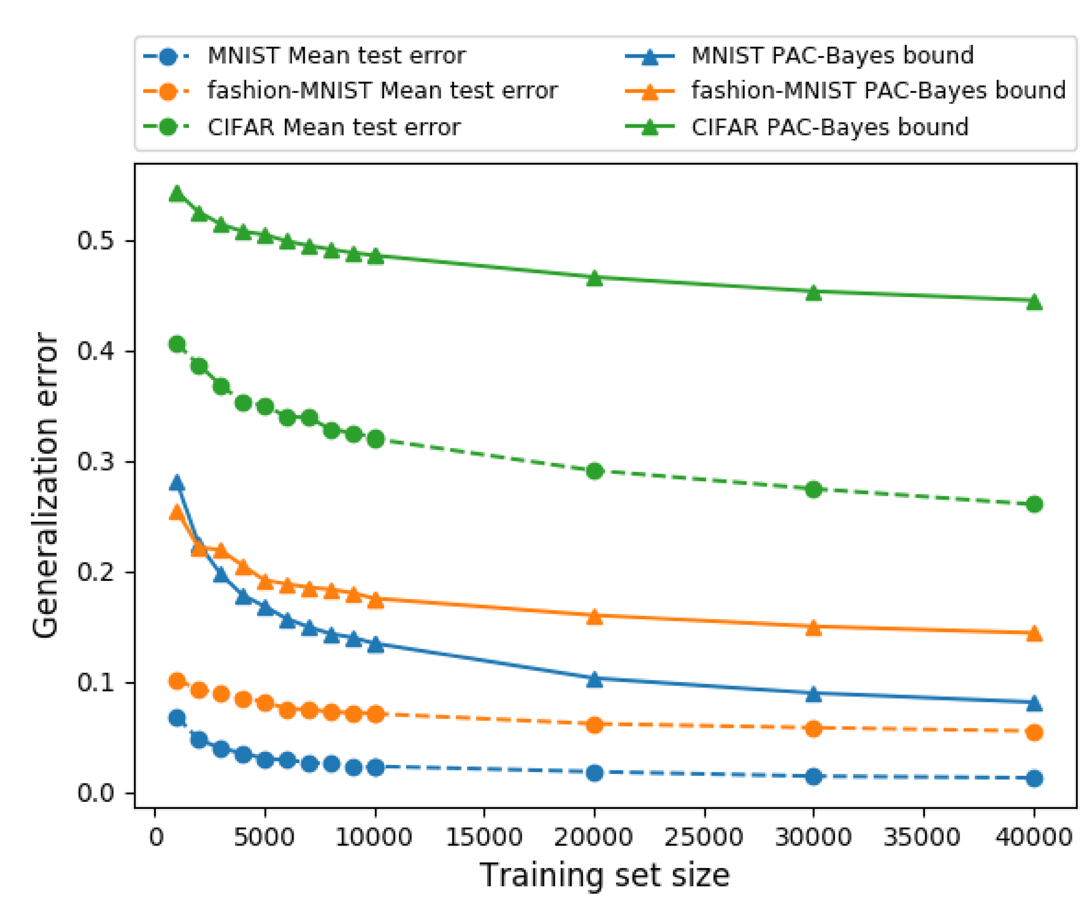
We trained a range of neural network architectures on several standard datasets, and **compared the test error with the PAC-Bayes bound** calculated from the training data.



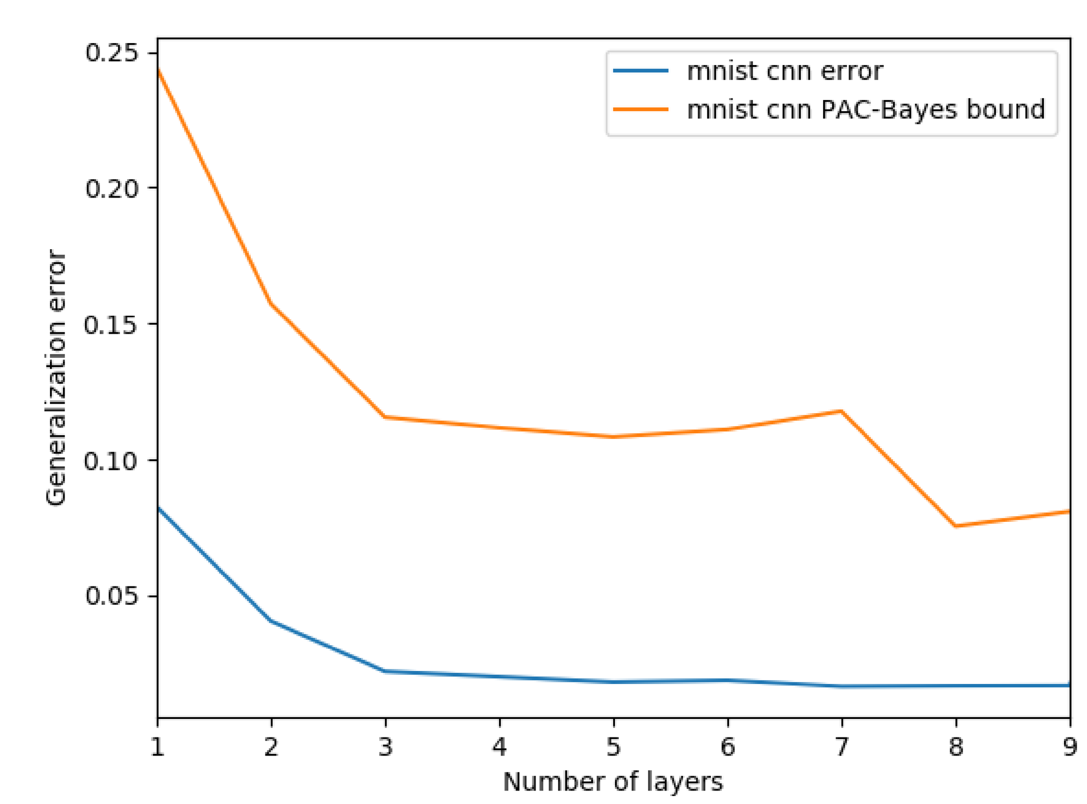
Error and bound, for different architectures trained on a sample of 10k images from CIFAR10 (with binarized labels)



Error and bound, versus label corruption (fraction of target labels which are randomized). Inset shows value of $P(C(S))$. Network is a 4-layer CNN without pooling.



Error and bound, versus training set size m, for a 4-layer CNN without pooling



Error and bound, versus number of layers for a CNN with max pooling, trained on a sample of 10k MNIST images

Limitations

- The bound
nonasymp
- The bound
choice se
networks.
hyperpara
- How well
is also an
- The calcu
which typi
MCMC). I
is probabl
- It isn't cle
available.
- As discus
versus tes
above que

Refs:

Valle-Perez et al., 2
simple functions. P

J Lee et al., 2017. D

A Garriga-Alonso e
2019

R Novak et al., 201
Published in ICLR 2

AGG Mathews et al

G Yang, 2019. Scali
Gradient Independe

M Kääriäinen et al.,