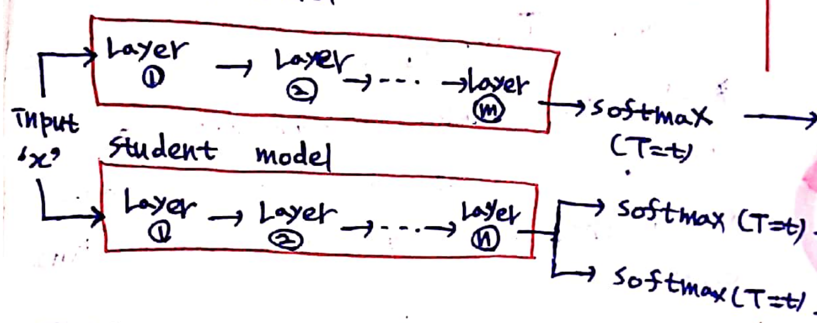
knowledge distillation ①

↳ 미리 잘 학습된 큰 네트워크 (teacher network)의 지식을 실제로 사용하고자 하는 작은 네트워크에게 전달하는 것.

「distilling the knowledge in a neural network」

★ knowledge distillation = 작은 네트워크도 큰 네트워크와 비슷한 성능을 낼 수 있도록 학습과정에서 큰 네트워크의 지식을 작은 네트워크에게 전달하여 작은 네트워크의 성능을 높이겠다는 목적.

teacher model



Total Loss $= (1-\lambda)\cdot L_{CE}(\sigma(Z_s), \hat{y}) + \overset{㉠}{2\lambda \cdot T^2 \cdot L_{CE}}\left(\sigma\left(\frac{Z_t}{T}\right), \sigma\left(\frac{Z_s}{T}\right)\right)$

L_{CE} = cross entropy loss.

Z_s = output logits of student network

Z_t = 〃 of teacher network

$\hat{y}$ = ground truth(one-hot)

λ = balancing parameter

↳ λ가 크면 ㉠: 오른쪽항 loss를 더 중요하게 보고 학습하겠다는 의미.

T = temperature : softmax 함수가 입력값이 큰 것은 아주 크게, 작은 것은 아주 작게 만드는 성질을 완화해줍니다.

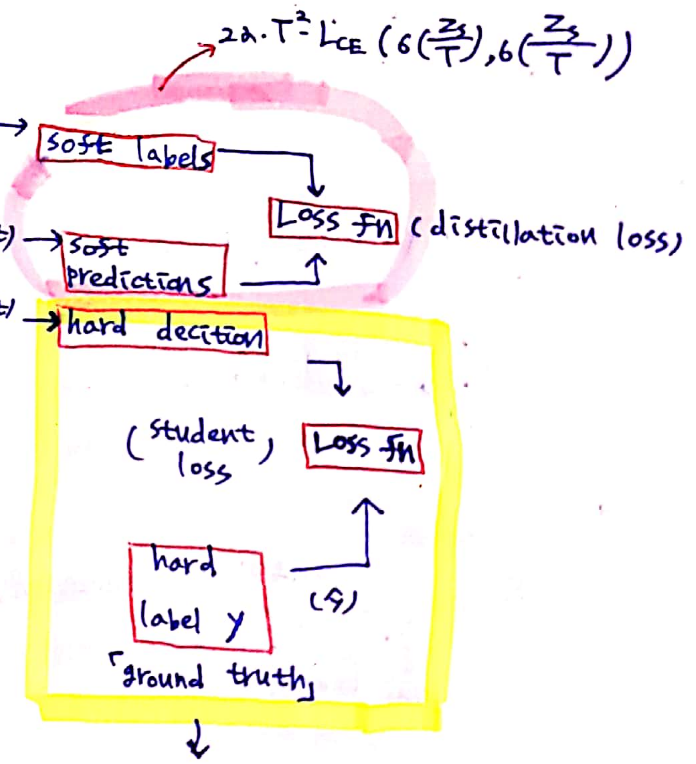
ex)

$$\begin{pmatrix} \text{bear} \\ \text{cat} \\ \text{dog} \end{pmatrix} = \overset{\text{hard label}}{\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}} \Rightarrow \overset{\text{soft label}}{\begin{pmatrix} 0.05 \\ 0.75 \\ 0.2 \end{pmatrix}}$$

정보손실 발생.

↳ 이런 정보의 손실 없이 teacher network의 분류결과를 student network의 분류결과와 비교시켜, student network가 teacher network를 모방하도록 학습시킵니다.

teacher, student network의 분류결과의 차이를 Loss에 포함.



$(1-\lambda)\cdot L_{CE}(\sigma(Z_s), \hat{y})$

「ground truth와 student의 분류 결과와의 차이를 cross entropy loss로 계산」

★ knowledge distillation = 미리 학습시킨 teacher network의 출력을 내가 실제로 사용하고자 하는 작은 모델인 student network가 모방하여 학습함으로써 상대적으로 적은 parameter를 가지고 있더라도 모델의 성능을 높이는 방법론.

① Sparse node activation loss

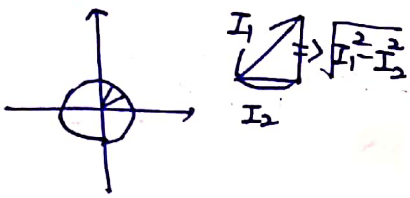
= check, variable node에 Lp norm을 더함으로써 얻어진다. (is obtained by adding an Lp norm on the check, variable node.

• norm = 벡터공간에서 벡터의 크기를 재기 위해서 사용. $\|x\|_p = \left(\sum_{i=1}^{N} |x_i|^p\right)^{\frac{1}{p}}$ ②

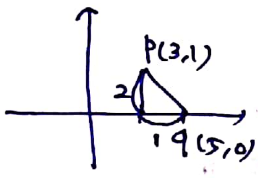
k-nearest neighbors: distance metric

1 ↳ L2 (euclidean) distance

$$d_2(I_1, I_2) = \sqrt{\sum_p (I_1^p - I_2^p)^2}$$



$d_1(p,q) = \|p-q\|$, $p = (3,1)$, $q = (5,0)$



$d_1(p,q) = 3$

L_1 regularization

$$\text{cost} = \frac{1}{n} \cdot \sum_{i=1}^{n} \{ \boxed{L(y_i, \hat{y}_i)} + \frac{\lambda}{2} |w| \}$$

기존의 cost function.

실제 ↳ 예측

$$L_{kd} = \sum_{t=1}^{T_{student}} L_d(t)$$

② Knowledge distillation loss

↳ imposes an additional constraint on the check node.

「we suggest using the knowledge distillation method by incorporating a teacher decoder into the training procedure.」 → At test time, decoding is performed using the student neural decoder.

Architecture

↳ teacher decoder = min-sum

student decoder = neural min-sum

teacher > student : BER

「$T_{teacher} > T_{student}$」

* since the teacher network has more layers, / imitating the teacher's node activations, will result in lower decoding error.

<B. The knowledge distillation Loss term>

we propose a new loss function to guide the training of the neural decoder. (student) /

$\mu_{c,v}^{teacher}(t)$, $\mu_{c,v}^{student}(t)$ = teacher, student check node messages at iteration 't'.

* the student network messages $\mu_{c,v}^{student}(t)$ at iteration 't' tried to mimic the teacher network messages $\mu_{c,v}^{teacher}(t+t_0)$ at iteration $t+t_0$.

$$L_{kd}(t) = \sum_{v=1}^{n} \sum_{c' \in N(v)} \| \sigma(\mu_{c,v}^{teacher}(t+t_0)) - \sigma(\mu_{c,v}^{student}(t)) \|_p^p$$

p = norm order

c) The sparse node activation Loss term
sparse parity check matrix leads to higher decoding performance.

↓

with sparse node activation (neural network decoder).

∴ The proposed sparse loss term is obtained by using a L_p norm on variable, check nodes over the student network.

$$L_s(t) = \sum_{v=1}^{n} \sum_{c' \in N(v)} \| 6(\mu^{student}_{c',v}(t)) \|_p^p + \sum_{c=1}^{n-k} \sum_{v' \in M(c)} \| 6(\mu^{student}_{v',c}(t)) \|_p^p$$

$$= \sum_{t=1}^{T_{student}} L_s(t) \qquad ③$$

∴ we observe that the p parameter is important for successful training since larger p increases the gradient.

$$\therefore L = L_{ce} + \alpha L_{kd} + \gamma \cdot L_s$$

「$\alpha = 1$, $\gamma = 0.01$」

$$L_{ce} = -\frac{1}{n} \cdot \sum_v B_v \cdot \log(6(S_v)) + (1 - B_v) \cdot \log(1 - 6(S_v))$$

$$B_v = \frac{1 - x_v}{2} \quad : \text{BPSK}$$

「bit corresponding to transmitted symbol x_v」

training exploding of phenomena = neural decoder가 전송된 codeword에 잘못된 예측치를 주는 CN, VN에 zero activation을 출력하기 위해서 train 되기에.

't' iteration: soft output vector $S_v(t)$

$$= l_v + \sum_{c' \in N(s)} \mu_{c'v}(t).$$

↓

$$\hat{x}_v(t) \begin{cases} 1, & \text{if } S_v(t) > 0. \\ 0 \end{cases}$$

$\beta_{c,v}(t)$ = SGD를 통해 train.

「results」

↳ $T_{teacher} = 30$ iterations, $T_{student} = 5$ iterations.

look-ahead parameter : $t_0 = 25$

※ 낮은 dB에서의 성능 하락
= knowledge distillation loss term does not use the ground truth bits at training.

Ablation Analysis = Fig.2.(a)

① knowledge distillation

② sparse node loss + cross entropy

③ green = min-sum (T=30, teacher).

↳ This shows that the ② (red) sparse node activation loss gives more improvement than that ① (cyan) knowledge distillation loss.

저 dB 에서 성능 하락은 cross-entropy loss로 해결 가능. (low)

Fig.2(d) = only sparse node activation

epoch 2 = 성능↑↑ (극적으로)

↳ for a low SNR regime, it gives high degradation in the BER performance.

knowledge distillation

ㄴ one ① uses a teacher network to guide the training of a smaller ② student neural network.

↓

제안 된 neural decoder의 node를 제한(constraint)하기 위해서, 전문가 teacher에 의해 knowledge distillation method를 사용하는 것 제시. (propose to use the knowledge distillation method/ by an expert $\boxed{\text{teacher}}$ network / to constraint the nodes (of the proposed neural decoder.)) (teacher)

The $\boxed{\text{student}}$ tries to Ⓐ decode the

「student는 teacher node activation을 흉내 냄으로서 새로운 loss term을 가진 전송된 codeword를 decode하기 위해 노력」

transmitted codeword with a novel loss term that mimics the teacher node activations.

* we present ① novel sparse node activation loss, ② knowledge distillation loss term.

↓

We demonstrate that each (of the new loss terms) improves the results of the baseline methods by a large margin without adding computational complexity.

• previous work have tried to improve the results by changing the neural architecture, the loss function at the output layer or finding ~~better~~ sparse parity check matrix.

↓

two novel loss terms.

「① sparse node loss, ② knowledge distillation loss」

• $\boxed{\text{Sparse}}$ activation regularization = imposes sparse 「activation에 희소제약을 부과하는 새로운 loss function term」 constrained on the activation with new loss function term.

Since we know that $\boxed{\text{sparse}}$ parity check matrices lead to better decoding performance, / we propose teaching the neural network decoder / with sparse nodes.

• knowledge distillation is a $\boxed{\text{basic}}$ technique in deep learning / where one uses a $\boxed{\text{teacher}}$ network / to guide the training of a smaller $\boxed{\text{student}}$ neural network.

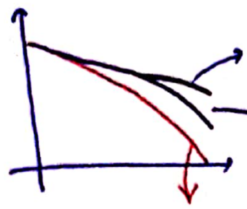
↓

we propose to use the $\boxed{\text{knowledge}}$ $\boxed{\text{distillation}}$ method / by an expert teacher network / to constraint the nodes of the proposed neural decoder. / (student)

ㄴ The student tries to $\boxed{\text{decode}}$ the transmitted codeword (with a novel loss term) ~~that~~ mimics the teacher node activations. /

⑤

Fig 2-(a) BCH (63,45)



Knowledge distillation (T=5)

sparse node Loss
+cross entropy
(T=5)

our
(T=5)

←Min-sum (T=5)

ours(T=5)

→NBP

ours(T=5) 1.1dB

NBP=ours 「 smaller ~~SNR values~~」

①neural decoder that trains only with the knowledge distillation loss (cyan)

② neural decoder that trains with the sparse node activation loss(red)

③ Min-sum (T=30, teacher)

Magenta Curve : ours (T=5)

=> achieves the best results which demonstrates the advantages of our method.

☆ Ablation studies are crucial for deep learning research -- can't stress this enough.

Understanding causality (in your system) is the most straightforward way / to generate reliable knowledge (the goal of any research) / And ablation is a very low-effort way / to look into ☆causality.

「Ablation study」

↳ Machine learning에서, ablation study는 machine learning system의 building blocks「dataset의 feature, model components」을 제거해서 전체 성능에 미치는 효과에 대한 insight을 얻기위한 과학적 실험.

• 각 model ablation trial은 ①1개 혹은 ②그 이상의 components가 제거된 모델을 학습하는 것을 포함한다.

• An ablation study typically refers to removing some features of the model or algorithm) and seeing how that affects performance.

↓ The teacher min-sum decoder can be regarded as lower bound to ①.

☆ training only with the knowledge distillation loss ① leads to degradation in the low SNR regime (3 ~6dB).

Complete

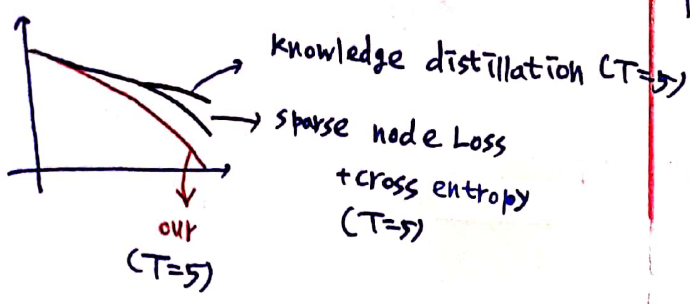
↓

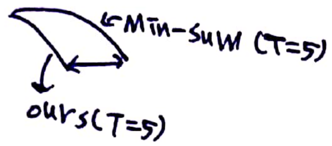
The degradation is alleviated when training with the cross-entropy loss.

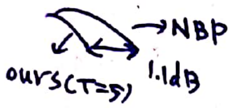
• 제안한 요소가 모델에 어떠한 영향을 미치는지 확인하고 싶을 때, 이 요소를 포함한/포함하지 않은 모델을 비교하는 것. ⇒ 이는 딥러닝 연구에서 매우 중요한 의미를 지니는 데, 시스템의 인과관계를 간단히 알아볼 수 있기 때문.

⑤

Fig 2-(b) BCH (63,45)







NBP=ours ┌ smaller ~~SNR values~~

① neural decoder that trains only with the knowledge distillation loss (cyan)

② neural decoder that trains with the sparse node activation loss (red)

③ Min-sum (T=30, teacher)

Magenta curve: ours (T=5)

=> achieves the best results which demonstrates the advantages of our method.

☆ Ablation studies are crucial for deep learning research -- can't stress this enough.

Understanding causality (in your system) is the most straightforward way / to generate reliable knowledge (the goal of any research) / And ablation is a very low-effort way / to look into ☆ causality.

┌ Ablation study

↳ Machine learning에서, ablation study는 machine learning system의 building blocks ┌ dataset의 feature, model components ┘ 을 제거해서 전체 성능에 미치는 효과에 대한 insight을 얻기위한 과학적 실험.

· 각 model ablation trial은 ① 1개 혹은 ② 그 이상의 components가 제거된 모델을 학습하는 것을 포함한다.

· An ablation study typically refers to removing some (features of the model or algorithm) and seeing how that affects performance.

The teacher min-sum decoder can be regarded as lower bound to ①.

☆ training only with the knowledge distillation loss ① leads to degradation in the low SNR regime (3 ~6dB). complete

↓

The degradation is alleviated when training with the cross-entropy loss.

· 제안한 요소가 모델에 어떠한 영향을 미치는지 확인하고 싶을 때, 이 요소를 포함한 / 포함하지 않은 모델을 비교하는 것. => 이는 딥러닝 연구에서 매우 중요한 의미를 지니는 데, 시스템의 인과관계를 간단히 알 수 있기 때문.