Better Lightweight Network for Free: Codeword Mimic Learning for Massive MIMO CSI Feedback

Zhilin Lu, Xudong Zhang, Rui Zeng, and Jintao Wang, Senior Member, IEEE

Abstract—The channel state information (CSI) needs to be fed back from the user equipment (UE) to the base station (BS) in frequency division duplexing (FDD) multiple-input multiple-output (MIMO) system. Recently, neural networks are widely applied to CSI compressed feedback since the original overhead is too large for the massive MIMO system. Notably, lightweight feedback networks attract special attention due to their practicality of deployment. However, the feedback accuracy is likely to be harmed by the network compression. In this letter, a cost-free distillation technique named codeword mimic (CM) is proposed to train better feedback networks with the practical lightweight encoder. A mimic-explore training strategy with a special distillation scheduler is designed to enhance the CM learning. Experiments show that the proposed CM learning outperforms the previous state-of-the-art feedback distillation method, boosting the performance of the lightweight feedback network without any extra inference cost.

Index Terms—Massive MIMO, CSI feedback, deep learning, lightweight network, codeword mimic, distillation.

I. INTRODUCTION

MASSIVE multiple-input multiple-output (MIMO) is an important technique to increase the spectrum and energy efficiency in modern wireless communication systems [1]. Nevertheless, the base station (BS) needs to obtain the downlink channel state information (CSI) in order to make use of the large-scale antennas with beamforming. Since the uplink and downlink channels are asymmetric in frequency division duplexing (FDD) systems, the downlink CSI needs to be estimated at the user equipment (UE) and reach the BS via uplink feedback. Due to the large feedback overhead brought by the huge antenna scale in massive MIMO systems, the CSI compressed feedback becomes an important task [2].

The deep learning (DL) aided CSI compression has become the mainstream feedback solution since CsiNet [3] proved its superiority over traditional compressed sensing. Following the pipeline in CsiNet, many influential works are proposed to acquire better feedback performance. For instance, DS-NLCsiNet [4] enhances the feedback performance with non-local blocks. CRNet [5] and MRFNet [6] introduce the multi-resolution architecture to capture the features of different scales. However, most of these advanced feedback networks achieve better performance at the cost of higher complexity, which harms the practicality of the design. Specifically, a lightweight encoder design is very necessary since the hardware resource of communication systems is strictly limited especially for the UE.

In fact, researchers have designed many methods to reduce the network cost. The fully connected (FC) layer is binarized in [7] so that the parameter size is greatly reduced. Dilated convolution is applied in [8] to enlarge the receptive field with limited complexity. Weight pruning is utilized in [9] to drop the redundant connections. Besides, [10] utilizes the information theory to propose a preprocessing pipeline with compressed feedback models. In [11], a single lightweight network is trained to process the real and imaginary parts of CSI separately, so that the overall complexity is greatly reduced. However, network compression is likely to cause performance degradation. It is significant to explore ways of enhancing lightweight networks.

Knowledge distillation (KD) [12] is a popular technique to boost the performance of lightweight networks in DL. It utilizes a powerful but heavier teacher model to guide the training of a lighter student model. A vanilla KD algorithm is first introduced into CSI feedback in [13]. Nevertheless, experiments show that the vanilla KD algorithm does not perform well with a practical extremely lightweight encoder.

In this letter, a new KD scheme is specially designed for the CSI feedback task and proved to be effective for the distillation with the extremely lightweight encoder. The main contributions of this letter are listed as follows.

• A new distillation strategy named codeword mimic (CM) is proposed for lightweight encoders. Codewords generated by the encoder are used as the transferred knowledge. Experiments show that our new scheme outperforms the previous state-of-the-art feedback KD strategy [13].

• A two-stage training pipeline named mimic-explore is designed for further performance boosting. Experiments show that the CM-aided feedback benefits from the balance between the codeword mimic and the CSI recovery.

• In order to unlock the full potential of the CM strategy, a cosine annealing distillation scheduler is designed based on ablation studies.

II. SYSTEM MODEL

In this letter, a single-cell massive MIMO system in FDD modes is considered. There are $N_t$ transmitting antennas at the BS and $N_t \gg 1$ is assumed. Meanwhile, the UE is equipped with $N_r$ receiving antennas and we have $N_r = 1$ to
simplify system settings. The orthogonal frequency division multiplexing (OFDM) with \( N_c \) sub-carriers is adopted. The received signal \( y_i \) at the \( i^{th} \) sub-carrier can be described as follows.

\[
y_i = \text{hi}^H \cdot \text{pi} \cdot x_i + \eta_i,
\]

where \( \text{hi} \in \mathbb{C}^{N_t \times 1} \), \( \text{pi} \in \mathbb{C}^{N_c \times 1} \), \( x_i \in \mathbb{C} \), and \( \eta_i \in \mathbb{C} \) represent the downlink channel vector, beamforming vector, transmitted symbol, and additive Gaussian noise at the \( i^{th} \) sub-carrier, respectively. \((\cdot)^H\) denotes conjugate transpose.

In the FDD system, there is a lack of reciprocity between uplink and downlink channels. Thus, the UE needs to feed the downlink CSI matrix \( \text{H} = [\text{h}_1, \ldots, \text{h}_{N_c}]^H \in \mathbb{C}^{N_c \times N_t} \) back to the BS for beamforming design. However, the downlink CSI matrix \( \text{H} \) contains \( 2N_cN_t \) real numbers, which is unacceptably large for the digital feedback.

Considering the channel sparsity in the angular-delay domain, we can first reduce the feedback overhead by transferring the downlink CSI matrix \( \text{H} \) into the angular-delay domain with discrete Fourier transform (DFT) as follows.

\[
\text{H} = \text{AHB}^H,
\]

where \( \text{A} \in \mathbb{C}^{N_c \times N_c} \) and \( \text{B} \in \mathbb{C}^{N_t \times N_t} \) are the DFT matrices. As mentioned above, the angular-delay domain CSI matrix \( \text{H} \) is sparse. Specifically, only the first \( N_c \) rows contain relatively large values. The elements in the rest of \( (N_c - N_c) \) rows are very close to zero. Therefore, we can truncate the original CSI matrix \( \text{H} \) to get a submatrix \( \text{H}_c \), which consists of only the first \( N_c \) rows. The feedback overhead can be reduced to \( 2N_cN_t \) real numbers if we only transmit the submatrix \( \text{H}_c \) and fill the rest of the matrix \( \text{H} \) with zero.

To further reduce the feedback cost, the DL-based CSI compression scheme is utilized. As shown in Fig. 2, the submatrix \( \text{H}_c \) is first compressed into a codeword \( \text{v} \in \mathbb{R}^{M \times 1} \) by the encoder network. Then, the decoder at the BS recovers the submatrix \( \hat{\text{H}}_c \) according to the codeword \( \text{v} \) fed back from the UE. Notably, we define the compression ratio as \( \eta = \frac{M}{N_cN_t} \). The whole feedback process can be described as follows.

\[
\hat{\text{H}}_c = \mathcal{D}(\mathcal{E}(\text{H}_c, \Theta_E), \Theta_D),
\]

where \( \mathcal{E}(\cdot) \) and \( \mathcal{D}(\cdot) \) denote the encoding and decoding procedures. \( \Theta_E \) and \( \Theta_D \) represent the parameters of encoder and decoder networks. Once the recovered \( \hat{\text{H}}_c \) is obtained, the zero-filling and inverse discrete Fourier transform (IDFT) are conducted to recover the original CSI matrix.

III. CODEWORD IMIC LEARNING FOR CSI FEEDBACK

A. Codeword Mimic Based Knowledge Distillation

In order to train lightweight networks better, vanilla distillation is first introduced to CSI feedback in [13]. As depicted in Fig. 1-(a), a heavy but more powerful teacher network takes part in the training and a distillation loss is added to narrow the gap between the output of the teacher \( \hat{\text{H}}_t \) and that of the student \( \text{H}_s \). However, knowledge transfer between decoder outputs can not guide the training of the encoder well due to the long distance. Considering that the UE is much more resource sensitive, a proper distillation scheme designed for lightweight encoder training is needed.

Instead of transferring knowledge from teacher output to student output, it is better for the lightweight encoder training to directly mimic the codeword produced by the teacher encoder. As we can see in Fig. 1-(b), our proposed codeword mimic learning is designed to transfer the knowledge from the teacher encoder to the student encoder. The CM strategy introduces a new term to the original MSE loss and the global loss function \( \mathcal{L} \) is defined as follows.

\[
\mathcal{L} = \alpha(t)\mathcal{L}_{cm} + (1 - \alpha(t))\mathcal{L}_{gt}
\]

\[
= \alpha(t)\text{MSE}(v^T, \hat{v}^S) + (1 - \alpha(t))\text{MSE}(\text{H}_c, \hat{\text{H}}_c),
\]

where \( \mathcal{L}_{cm} \) and \( \mathcal{L}_{gt} \) are the CM loss and the original ground truth loss, respectively. Codewords given by the teacher and student encoder are denoted as \( v^T \) and \( v^S \). \( \alpha(t) \in [0, 1] \) is a balance weight of the two losses determined by the current
training epoch $t$. The detailed design of $\alpha(t)$ will be discussed in section III-B.

Compared with previous KD feedback in [13], the proposed CM strategy has two advantages. For one thing, the lightweight encoder is more likely to converge towards a better local minimum following the guidance of the powerful teacher encoder. This is beneficial when the encoder is compressed to meet the UE deployment constraints. For another, the extra training cost brought by the teacher network is reduced since the teacher decoder with dominant complexity is removed.

B. Mimic-Explore Strategy and Distillation Scheduler Design

In spite of the aforementioned benefits, the CM strategy brings extra challenges to the feedback network training. The main concern is that the codeword mimicking would restrain the network convergence since the optimization is not focused on pure CSI reconstruction. Therefore, a two-stage training strategy named mimic-explore is proposed specially for the CM learning scheme. The main idea is to mimic the codeword only at the early stage. When the student codeword is close enough to the teacher one, the whole network is allowed to explore freely for better CSI reconstruction while keeping the codeword similarity to some extent.

The mimic-explore strategy can be easily applied to feedback network training with the learning rate (LR) design. As it is depicted in Fig. 3, the encoder and the decoder share the same cosine annealing LR at the mimic stage, decaying from $2 \times 10^{-3}$ to $4 \times 10^{-5}$. The loss function of the mimic stage is defined as (4) so that the convergence is guided by codeword mimicking and CSI reconstruction at the same time. After mimicking for hundreds of epochs, the knowledge would be sufficiently transferred from the teacher encoder to the student encoder.

Whereafter, the loss function $L$ degenerates to the ordinary ground truth loss $L_{gt}$, and the network is encouraged to explore freely premised on preserving the codeword similarity. Different LR schemes are set for the encoder and the decoder to achieve such an intention. As we can see in Fig. 3, the decoder LR decays from $4 \times 10^{-3}$ to $4 \times 10^{-5}$, which allows a sharp update for the decoder parameters. Meanwhile, the initial LR of the encoder is set to $2 \times 10^{-4}$, which is so small that the encoder is hardly changed during training. This guarantees the retention of the knowledge learned at the mimic stage for the CM encoder.

In order to balance the $L_{cm}$ and $L_{gt}$ with the best practice at the mimic stage, a cosine annealing distillation scheduler as shown in 5 is chosen based on the ablation studies.

$$
\alpha(t) = \begin{cases} 
\frac{1}{2} \alpha_0 \left(1 + \cos \left(\frac{t}{T_{cm}} \pi\right)\right) & t \leq T_{cm} \\
0 & t > T_{cm}
\end{cases}
$$

where $\alpha_0$ is the maximal value of $\alpha(t)$ and $T_{cm}$ is the epochs of the mimic training stage. It is obvious that $L_{cm}$ is disabled when the current training epoch $t$ reaches the explore stage. Notably, the $\alpha_0$ is set to $10^{-4}$ so that the $L_{cm}$ and $L_{gt}$ can be scaled to a similar size. Compared with the linear decay scheduler, the cosine annealing scheduler provides a varying speed of $\alpha(t)$ evolution during different training stages. In this way, the network can find a better region for the codeword mimicking with longer training at the early stage. Besides, more detailed and thorough exploration is carried out in a near-optimal zone with a longer late training stage.

C. BCRNet: A Typical Lightweight Feedback Network

The FC layer binarization proposed in [7] is a typical feedback encoder compression strategy. As demonstrated in Fig. 4, we take the influential CRNet as an example and apply the FC binarization to it, producing binary CRNet (BCRNet) with an extremely lightweight encoder. It is worth mentioning that the decoder of BCRNet is the same as the original CRNet. The CM learning is a general technique to boost the performance of the feedback network with the lightweight encoder. Since the FC binarization is applicable to a wide range of feedback networks, CRNet and BCRNet can serve as a group of representative teacher and student networks to prove the effectiveness of the proposed CM learning.

D. CsiNet-RE: A Heavier Teacher Designed for CsiNet

When distilling BCRNet from CRNet, the architecture of the student encoder is the same as the teacher encoder except for the binarized FC layer. To further demonstrate the generality of the proposed CM learning strategy, we also apply it to the plain student network whose encoder architecture is different from the teacher encoder.

The famous CsiNet [3] is adopted as the student network. Meanwhile, a new network named CsiNetRefineEncoder (CsiNet-RE) is introduced as the teacher with a more powerful
Fig. 5. The architecture of the CsiNet encoder and the CsiNet-RE-encoder. The input dimensions (channel × width × height) are given above each layer. Note that the leaky ReLU activation with a negative slope 0.3 is added after each batch normalization (BN) layer. The activation function is omitted in the schematic diagram for simplicity.

encoder to boost the performance under the same decoder. As shown in Fig. 5, the simple 3 × 3 convolutional layer in the CsiNet encoder is upgraded to a refine block, which is originally used only in the CsiNet decoder.

Naturally, the CsiNet-RE can achieve better feedback performance compared with the original CsiNet owing to the richer CSI information in the learned codeword brought by the extra computational complexity. We will show that the proposed CM strategy is able to train better CsiNet from the codeword of the heavier teacher network as expected.

IV. RESULTS AND ANALYSIS

A. Experimental Settings

The COST2100 dataset [14] is adopted following the setting of numerous previous feedback researches including CsiNet [3], CRNet [5], etc. The training and test datasets consist of 100,000 and 20,000 independently generated samples, respectively. Experiments are carried out under the indoor scenario at 5.3GHz and the outdoor scenario at 300 MHz. The number of sub-carriers $N_c$ is set to 1024. $N_c$ and $N_t$ are set to 32 and 32, respectively.

Adam optimizer is used for the network training under the PyTorch framework. Following the original CRNet, warmup-aided cosine annealing LR is adopted for all the benchmark schemes as Fig. 3 shows. All the schemes are trained for 1000 epochs for fair comparison and the batch size is set to 200. The normalized mean square error (NMSE) is used to evaluate the performance of CSI reconstruction as follows.

$$\text{NMSE} = E \left( \frac{\| \hat{H} - H_c \|^2}{\| H_c \|^2} \right).$$ (6)

B. Performance of the Proposed Codeword Mimic Learning

As we can see from Table I, the student network BCRNet suffers from a prominent performance loss compared with the teacher network CRNet due to the encoder compression. Specifically, the encoder size of the BCRNet is over 30× smaller than the CRNet, which is vastly beneficial to the deployment at UE. And the next challenge is to narrow the performance gap between the BCRNet and the CRNet.

Table I shows that the NMSE performance of BCRNet is steadily improved with the help of the proposed CM strategy. For instance, the NMSE of BCRNet is decreased for 1.86 dB and 1.11 dB under the indoor and outdoor scenario when the compression ratio $\eta$ is 1/4. It is worth mentioning that the performance boosting is a result of knowledge transfer from the teacher network and the BCRNet structure is not changed at all. In other words, the improvement is completely cost-free for the network deployment.

Moreover, we compare the CM strategy with the previous state-of-the-art distillation strategy [13]. It can be deduced from Table I that the vanilla KD strategy is not enough to transfer the knowledge to the student with the extremely lightweight encoder. The proposed CM distillation scheme outperforms the previous KD scheme since the codeword mimic provides more direct guidance for the encoder learning.
encoder architecture is different from the teacher encoder. Specifically, we utilize the CsiNet-RE as the teacher model and the original CsiNet [3] as the student model as it is introduced in section III-D. From the results in Table II, we can see that the proposed CM strategy also achieves solid performance improvement and outperforms the vanilla KD method. In this way, its generality and robustness on different model structures can be proved.

For realizing the full potential of the proposed CM strategy, ablation studies on the mimic-explore proportion are designed as Table III. Note that \( \eta \) is set to 1/4 for all experiments and the scheme containing 0 epochs of mimicking and 1000 epochs of exploring is the benchmark without the CM learning. It is clear that the final codeword MSE is reduced by dozens of times after adding the CM strategy, proving its effectiveness on the codeword knowledge transfer. As we can see, the targeted NMSE reaches the minimum with 1000 epochs of mimicking. The shorter mimicking stage harms the codeword learning. Experiments showed that the proposed CM learning outperformed the previous state-of-the-art feedback distillation scheme and improved the CSI reconstruction quality of the lightweight network under different compression ratios.

Finally, we briefly discuss the disadvantages of the proposed algorithm. Firstly, the CM strategy still involves the inference of the teacher encoder, so the training overhead is inevitably increased. Secondly, the performance gain of the CM learning is constrained by the performance of the teacher model.

V. CONCLUSION

In this letter, a distillation strategy named codeword mimic (CM) was specially designed for the CSI feedback task. With the knowledge transferred from the teacher’s codeword, the feedback network with an extremely lightweight encoder could achieve better performance without any extra inference cost. In addition, the special mimic-explore strategy and the distillation scheduler were designed to boost the performance of the CM learning. Experiments showed that the proposed CM learning outperformed the previous state-of-the-art feedback distillation scheme and improved the CSI reconstruction quality of the lightweight network under different compression ratios.

ACKNOWLEDGMENT

The key results can be reproduced with the following Github link: https://github.com/Kylin9511/CodewordMimicFeedback.

REFERENCES

[1] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, “An overview of massive MIMO: Benefits and challenges,” IEEE J. Sel. Topics Signal Process., vol. 8, no. 5, pp. 742–758, Oct. 2014.
[2] J. Guo, C.-K. Wen, S. Jin, and G. Y. Li, “Overview of deep learning-based CSI feedback in massive MIMO systems,” IEEE Trans. Commun., vol. 70, no. 12, pp. 8017–8045, Dec. 2022.
[3] C.-K. Wen, W. T. Shih, and S. Jin, “Deep learning for massive MIMO CSI feedback,” IEEE Wireless Commun. Lett., vol. 7, no. 5, pp. 748–751, Oct. 2018.
[4] X. Yu, X. Li, H. Wu, and Y. Bai, “DS-NLCsiNet: Exploiting non-local neural networks for massive MIMO CSI feedback,” IEEE Commun. Lett., vol. 24, no. 12, pp. 2790–2794, Dec. 2020.
[5] Z. Lu, J. Wang, and J. Song, “Multi-resolution CSI feedback with deep learning in massive MIMO system,” in Proc. IEEE Int. Conf. Commun. (ICC), Jun. 2020, pp. 1–6.
[6] Z. Hu, J. Guo, G. Liu, H. Zheng, and J. Xue, “MRBNet: A deep learning-based CSI feedback approach of massive MIMO systems,” IEEE Commun. Lett., vol. 25, no. 10, pp. 3310–3314, Oct. 2021.
[7] Z. Lu, J. Wang, and J. Song, “Binary neural network aided CSI feedback in massive MIMO system,” IEEE Wireless Commun. Lett., vol. 10, no. 6, pp. 1305–1308, Jun. 2021.
[8] S. Tang, J. Xia, L. Fan, X. Lei, W. Xu, and A. Nallanathan, “Dilated convolution based CSI feedback compression for massive MIMO systems,” IEEE Trans. Veh. Technol., vol. 71, no. 10, pp. 11216–11221, Oct. 2022.
[9] J. Guo, J. Wang, C.-K. Wen, S. Jin, and G. Y. Li, “Compression and acceleration of neural networks for communications,” IEEE Wireless Commun., vol. 27, no. 4, pp. 110–117, Aug. 2020.
[10] Z. Yin, W. Xu, R. Xie, S. Zhang, D. W. K. Ng, and X. You, “Deep CSI compression for massive MIMO: A self-information model-driven neural network,” IEEE Trans. Wireless Commun., vol. 21, no. 10, pp. 8872–8886, Oct. 2022.
[11] Y. Sun, W. Xu, L. Liang, N. Wang, G. Y. Li, and X. You, “A lightweight deep network for efficient CSI feedback in massive MIMO systems,” IEEE Commun. Lett., vol. 10, no. 8, pp. 1840–1844, Aug. 2021.
[12] G. E. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” 2015, arXiv:1503.02531.
[13] H. Tang, J. Guo, M. Matthaiou, C.-K. Wen, and S. Jin, “Knowledge-distillation-aided lightweight neural network for massive MIMO CSI feedback,” in Proc. IEEE 94th Veh. Technol. Conf. (VTC-Fall), Sep. 2021, pp. 1–5.
[14] L. Liu et al., “The COST 2100 MIMO channel model,” IEEE Wireless Commun., vol. 19, no. 6, pp. 92–99, Dec. 2012.