대규모 다중입출력 통신 시스템을 위한 딥러닝 기반 CSI 피드백: 벡터 양자화 접근법

신준용, 전요셉
포항공과대학교

{sjyong, yoseb.jeon}@postech.ac.kr

Deep-Learning-Based CSI Feedback for Massive MIMO Systems: A Vector Quantization Approach

Junyong Shin, Yo-Seb Jeon
Pohang Univ. of Science and Technology (POSTECH)

요 약

본 논문은 대규모 다중입출력 통신 시스템을 위한 딥러닝 기반 채널 상태 정보(Channel state information, CSI) 피드백 기법을 제안한다. 제안된 기법은 CSI 정보를 압축하고 복원시키기 위해 벡터 양자화 variational autoencoder (VAE) 기법을 활용한다. 특별히, VAE 기법에서 발생하는 복잡도 문제를 해결하기 위해 잠재 벡터의 크기와 방향을 분리하여 각각 양자화 시키는 Shape-Gain 벡터 양자화 기법을 적용한다. 모의 실험을 통해, 제안된 기법이 동일한 피드백 오버헤드 조건 하에서 기존 기법들 보다 높은 CSI 복원 성능과 적은 계산 복잡도를 가진다는 것을 입증한다.

Ⅰ. 서 론

주파수 분할 이중화 다중 입출력(Multiple-input multiple-output, MIMO) 시스템의 전송 속도를 최대화하기 위해서는 기지국이 사용자들의 채널 상태 정보(Channel state information, CSI)를 정확하게 알고 있어야 한다. 이를 실현하기 위해서 각 사용자들은 파일럿 신호를 통해 추정한 CSI 를 주어진 코드북을 이용하여 기지국에 전달하는 CSI 피드백 기술이 활용되고 있다. 그러나, 최근 차세대 통신 기술로 주목받고 있는 대규모(Massive) MIMO 시스템의 경우, 많은 수의 안테나로 인해 CSI 의 차원이 크게 증하며, 이로 인해 CSI 피드백 과정에서 상당한 오버헤드가 발생하게 된다. 이 문제에 대한 해결책으로써, 최근 Autoencoder 를 활용한 딥러닝 기반 CSI 피드백 기법이 큰 주목을 받고 있다. 본 기법에서는 encoder 신경망을 활용하여 CSI 를 더 낮은 차원의 잠재 공간 상의 벡터로 압축하고, decoder 신경망을 활용하여 CSI 를 다시 복원한다. 본 기법을 활용할 경우 잠재 벡터의 모든 요소는 임의의 실수 값으로 얻어지므로, 이를 유한 비트로 변환하기 위한 양자화 기법이 함께 고려되어야 한다.

기존 Autoencoder 의 잠재공간 상에서 벡터 양자화 모델을 공동 학습시킨 선행 연구로는 VQ-VAE (Vector quantized variational autoencoder) 가 존재한다 [1]. 해당 모델은 잠재공간 상에서 학습이 가능한 벡터 코드북을 활용하여 압축, 양자화, 복구 과정을 동시에 최적화한다. 그러나 VQ-VAE 를 CSI 피드백에 그대로 적용할 경우 잠재 벡터와 모든 코드 벡터 간의 거리를 계산하는 과정에서 큰 계산 복잡도가 발생한다. 따라서, 딥러닝 기반 CSI 피드백을 위해 VQ-VAE 기술을 활용하기 위해서는 양자화 과정의 계산 복잡도를 줄일 수 있는 기술 개발이 필수적이다.

본 논문에서는 VQ-VAE 기술의 복잡도 문제를 해결할 수 있는 딥러닝 기반 CSI 피드백 기법을 제안한다. 제안된 기법은 Shape-Gain 벡터 양자화를 통해 잠재 벡터의 크기와 방향을 따로 양자화하여 양자화 계산 복잡도를 크게 감소시킨다. 모의 실험을 통해 제안된 기법이 기존의 기법들 보다 더 높은 CSI 복원 성능과 적은 계산 복잡도를 가진다는 것을 입증한다.

Ⅱ. 본 론

본 논문에서는 단일 사용자 안테나 및 N_t 개의 기지국 안테나를 상정한다. 또한 기지국은 N_c 개의 subcarrier 를 통해 직교 주파수 분할 다중 방식(Orthogonal frequency-division multiplexing)을 수행한다고 가정한다. 따라서 공간-주파수 영역에서 CSI 행렬은 $\mathbf{H}_{\mathrm{sf}} \in \mathbb{C}^{N_c \times N_t}$ 와 같이 나타낼 수 있다. 여기서, 해당 행렬의 압축을 위해 각도-지연 영역에서의 희소 특성을 활용할 수 있다. Massive MIMO 시스템에서는 안테나 수에 비해 반사체의 수와 다중 경로 지연의 가짓수가 적기 때문에, 각도-지연 영역에서 희소 특성을 보이고, 특히 지연 영역에서는 작은 값에 몰려있는 경향을 보인다. 따라서 위의 행렬을 2 차원 이산 푸리에 변환을 통해 각도-지연 영역으로 변환하고, 첫 $\tilde{N}_c$ 만큼의 지연 영역만을 취하면 전처리가 완료된 실수 상의 CSI 행렬 $\tilde{\mathbf{H}}_{\mathrm{ad}} \in \mathbb{R}^{2\times \tilde{N}_c \times N_t}$이 유도된다. $\tilde{\mathbf{H}}_{\mathrm{ad}}$는 encoder 신경망 f_{enc}를 거쳐 총 M 차원의 잠재벡터 $\mathbf{z} = f_{\mathrm{enc}}(\tilde{\mathbf{H}}_{\mathrm{ad}})$ 로 변환된다. 그리고 $\mathbf{z}$ 는 양자화 모델 Q 를 통해 $\mathbf{z}_q = Q(\mathbf{z})$ 로써 양자화된다. 이후 $\mathbf{z}_q$는 사용자로부터 기지국에게 피드백 되며, 기지국에서 decoder 신경망 f_{dec} 를 통해 $\hat{\mathbf{H}} = f_{\mathrm{dec}}(\mathbf{z}_q)$로써 복원된다.

양자화 모델 Q로 잠재벡터를 양자화하는 일반적인 방식은 잠재벡터 $\mathbf{z}$를 D 차원을 가지는 N 개의 하위 벡터들 $\mathbf{z}_i = [z_{(i-1)D+1}, \cdots, z_{iD}]$로 나눈 후 이들을 독립적으로 D 차원의 codebook 을 통해 양자화하는 것이다. 기존의 VQ-VAE 모델에서는, B 비트를 사용하여 2^B개의 코드벡터 $\{\mathbf{b}_k\}_{k=1}^{B}$를 가지는 D차원의 코드북 $\mathcal{B}$를 통해 $\mathbf{z}_i$를 codebook 내 가장 가까운 코드벡터인 $\mathbf{z}_{q,i} = \mathrm{argmin}_{\mathbf{b}_k \in \mathcal{B}} \|\mathbf{z}_i - \mathbf{b}_k\|^2$ 로 양자화 시킨다. 또한, encoder, 코드북, decoder 를 공동 학습시키기 위한 VQ-VAE 의 손실 함수는 다음과 같다.

$$\mathcal{L} = \|\hat{\mathbf{H}} - \tilde{\mathbf{H}}_{\mathrm{ad}}\|_F^2 + \|sg(\mathbf{z}) - \mathbf{z}_q\|^2 + \beta\|\mathbf{z} - sg(\mathbf{z}_q)\|^2. \quad (1)$$

여기서 sg(·)는 'stop-gradient' 연산을 의미하며, 역전파 과정에서 계산되는 gradient 를 무시하고, 해당 입력 값을 변수가 아닌 상수취급을 하도록 한다. 위 손실 함수에서 첫

번째 항은 전체 모델의 복원 손실 정도를 나타내고, 두 번째와 세 번째 항은 양자화 모델의 양자화 손실 정도를 나타낸다. VQ-VAE 모델은 벡터 양자화를 통해 잠재 벡터의 각 요소간의 상관관계를 고려한 양자화가 가능하다는 장점이 있지만, 잠재벡터와 모든 코드 벡터 간의 거리를 계산해야 하기 때문에 계산 복잡도가 크다는 단점이 있다.

VQ-VAE 에서 요구되는 계산 복잡도를 줄이기 위해, 제안 기법은 Shape-Gain 양자화 방식을 적용한다. 이 방식에서는 하위 잠재 벡터인 $\mathbf{z}_i$의 크기와 방향을 각각 식 (2)와 같이 양자화 한다. 여기서 Q_{mag}와 $\mathbf{Q}_{dir}$은 각각 하위 잠재 벡터의 크기와 방향 양자화 하는 함수이며, 양자화 함수에는 B_{mag} 와 B_{dir} 비트가 각각 할당되어 있다.

$$\mathbf{z}_{q,i} = Q(\mathbf{z}_i) = Q_{mag}(\|\mathbf{z}_i\|) \cdot \mathbf{Q}_{dir}(\mathbf{z}_i/\|\mathbf{z}_i\|). \quad (2)$$

Gain 양자화의 경우, $\|\mathbf{z}_i\|$의 값을 양자화 한다. 본 모델은 encoder 의 출력단에 Tanh 활성함수를 두어 잠재 공간의 각 요소들이 $(-1,1)$ 범위의 값으로 제한되도록 한다. 이에 따라 $\|\mathbf{z}_i\|$의 값은 $(0, D)$ 범위의 값으로 제한된다. Q_{mag}는 이 제한된 범위 내의 스칼라를 양자화한다. 여러 모의 실험으로부터 $\|\mathbf{z}_i\|$ 값의 분포를 확인한 결과, 이 값들은 해당 범위에 균일하게 분포돼 있는 것이 아니라 0 주변의 값에 몰려 있는 특성이 있다. 본 논문은 이를 활용해 다음과 같이 정의된 clipped μ-law quantization 을 제안하고자 한다.

$$Q_{\text{mag}}(\|z_i\|) = g\left(\hat{f}_{\text{quan}}\left(\hat{h}_{\mu-law}(\|z_i\|)\right)\right), \quad (3)$$

$$\hat{h}_{\mu-law}(x) = \begin{cases} ln(1+\mu\frac{x}{\sqrt{D}}), & 0 < x \le \frac{(1+\mu)^A-1}{\mu}\sqrt{D}, \\ A, & \frac{(1+\mu)^A-1}{\mu}\sqrt{D} \le x < \sqrt{D}, \end{cases} \quad (4)$$

$$\hat{f}_{\text{quan}}(x) = \frac{A}{2^{B_{\text{mag}}}-1}\left\{\text{round}\left((2^{B_{\text{mag}}}-1)x/A - 0.5\right) + 0.5\right\}, \quad (5)$$

$$g(x) = \frac{(1+\mu)^x-1}{\mu}. \quad (6)$$

여기서, 식 (5)에 있는 round function 은 역전파 과정에서 gradient 값을 계산할 때 0 을 도출하기 때문에, 전체 모델의 학습과정에 문제가 발생한다. 따라서 역전파 과정에서는 이 대신에 다음과 같이 정의된 함수 $\tilde{f}_{\text{quan}}$를 사용한다. 즉, $\nabla Q_{\text{mag}}(\|z_i\|) = \nabla g\left(\tilde{f}_{\text{quan}}\left(\hat{h}_{\mu-law}(\|z_i\|)\right)\right)$가 만족된다.

$$\tilde{f}_{\text{quan}}(x) = \frac{A}{2(2^{B_{\text{mag}}}-1)}\left\{\sum_{i=1}^{2^{B_{\text{mag}}}-1}\tanh\left(\tau\left(\frac{(2^{B_{\text{mag}}}-1)x}{A} - i\right)\right) + 1\right\}. \quad (7)$$

Shape 양자화의 경우, $\tilde{z}_i = z_i/\|\mathbf{z}_i\|$로 정의된 $\mathbf{z}_i$의 방향을 양자화 한다. Shape 양자화 함수 $\mathbf{Q}_{dir}$는 $2^{B_{dir}}$ 개의 D차원의 단위 벡터들을 코드 벡터로 가지는 학습 가능한 codebook $\mathcal{B}_{\text{dir}}$을 활용한다. 이를 통해 $\tilde{\mathbf{z}}_i$를 codebook 내 가장 가까운 코드 벡터로 양자화하여 $\mathbf{Q}_{dir}(\tilde{\mathbf{z}}_i) = \text{argmax}_{\mathbf{b}_k \in \mathcal{B}_{\text{dir}}} |\tilde{\mathbf{z}}_i^{\text{T}} \mathbf{b}_k|$ 가 성립한다. 모델의 학습을 시작하기 전에, $\mathcal{B}_{\text{dir}}$의 초기화를 위해 Grassmannian line packing problem 을 D차원 단위 구 상의 실수 벡터들에 대해 해결한 결과를 사용하였다. 해당 결과는 D 차원 단위 벡터들을 각각으로부터 최대한 멀리 배치시키고, 결과적으로 구 상에서 균일하게 배치시킨다. 또한, Shape 양자화 모델의 코드 벡터들은 일반적으로 경사 하강법으로 갱신 후에 크기가 1 로 유지되지 않는다. 따라서 코드 벡터들의 변수들을 갱신한 후에, 다시 크기를 정규화하는 작업을 거치게 된다.

본 논문에서 제안된 Shape-Gain 양자화 모델의 우수성을 보이기 위해 모의 실험을 진행하였다. 실험에서는 COST2100 채널 모델 시뮬레이션의 300MHz 야외 시나리오[2]에서 생성된 데이터 셋을 사용하였다. 또한, 지금까지 언급된 변수들은 $N_t = 32,\ N_c = 1024,\ \tilde{N}_c = 32\ , A = 0.6, B_{\text{mag}} = 4,\ D = 16$와 같이 설정되었다. Encoder, decoder 모델은 [1]의 모델을 차용하였고, 비교 기법으로써 양자화 모델이 VQ-VAE 일 때, 스칼라 양자화 모델(ScalarQ 로 표기)일 때와 제안된 기법을 비교할 것이다.

그림 1 은 주어진 피드백 비트 상에서 제안된 기법과 기존 기법의 $NMSE = \mathbf{E}\{\|\tilde{\mathbf{H}} - \tilde{\mathbf{H}}_{\text{ad}}\|_F^2 / \|\tilde{\mathbf{H}}_{\text{ad}}\|_F^2\}$ 성능을 비교한 것이다. 결과에서 볼 수 있듯이, 제안된 Shape-Gain 양자화 모델이 다른 기법들에 비해 더 좋은 성능을 보인다. 이는 제안된 기법이 주어진 codebook 을 보다 효율적으로 학습할 수 있게 한다는 것을 보여준다.

그림 2 는 $M = 4096$에서 주어진 계산 복잡도 상 제안된 기법과 VQ-VAE 의 성능을 비교한 것이다. 계산 복잡도는 양자화 모델안에서 곱셈 연산의 횟수를 계산한 것이다. 이 결과는 제안된 기법이 벡터 양자화의 계산 복잡도 부담을 덜고, 좋은 성능을 보인다는 것을 나타낸다.

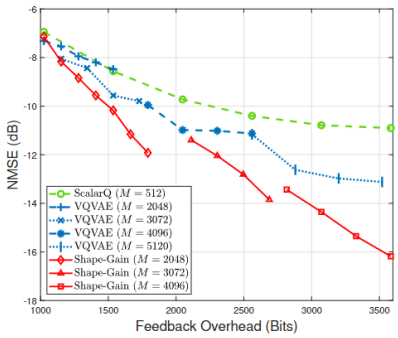

그림 1. 주어진 피드백 비트 상에서 제안된 기법과 기존 기법의 성능 비교

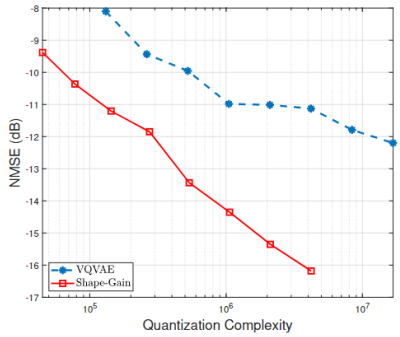

그림 2. 주어진 계산 복잡도 상에서 제안된 기법과 기존 기법의 성능 비교

Ⅲ. 결 론

본 논문에서는 Massive MIMO 통신 환경을 위한 딥러닝 기반 CSI 피드백 기법을 제시하였다. 제시된 기법은 잠재 벡터의 크기와 방향을 분리하여 양자화하는 것으로 양자화 모델이 공동 학습된 딥러닝 모델의 성능을 높이고 계산 복잡도를 줄였다. 모의 실험을 통해, 제안된 기법이 기존의 기법들보다 CSI 복원 성능 및 계산 복잡도 관점에서 우수함을 증명하였다.

ACKNOWLEDGMENT

이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2022R1C1C1010074).

참 고 문 헌

[1] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, “Neural discrete representation learning,” Adv. Neural Inf. Process. Syst., pp. 6306-6315, 2017.

[2] L. Liu et al., “The COST 2100 MIMO channel model,” IEEE Wireless Commun., vol. 19, no. 6, pp. 92-99, Dec. 2012.