VAE 기반 모델의 Batch 단위 Top-K 희소 잠재 표현을 위한 EMA 기반 추론 기법 연구

변성훈, 이민식*

한양대학교

goldpig0625@hanyang.ca.kr *mleepaper@hanyang.ac.kr

A Study on Batch Top-K Sparse Latent Representations with EMA-Based Inference in VAE-Based Models

Byeon Seong Hoon, Lee Min Sik*

Hanyang Univ.

요 약

본 논문은 VAE 모델의 잠재 표현을 효율적으로 희소화하기 위해 batch 단위 Top-K 기반 잠재 선택 기법을 적용하고, 테스트 단계에서 지수 이동 평균(EMA)을 활용한 추론 방식을 제안한다. 제안한 방법은 활성 잠재 차원 수를 크게 제한함에도 불구하고 생성 이미지 분포의 유사성 측면에서 기존 방법 대비 개선된 성능을 보였으며, 재구성 강건성 측면에서도 비슷하거나 더 나은 수준의 성능을 달성하였다.

Ⅰ. 서 론

변분 오토인코더(VAE)[1]는 확률적 잠재 공간을 통해 데이터의 핵심적인 표현을 학습하는 생성 모델로, 이미지 생성 및 표현 학습 분야에서 널리 활용되고 있다. VAE의 잠재 표현을 희소하게 구성하는 방법은 모델의 해석 가능성과 계산 효율성을 높일 수 있다는 점에서 지속적으로 연구되어 왔다.[2][3]

한편, batch 단위로 특징의 중요도를 판단하여 Top-K 선택을 수행하는 방법은 Sparse Autoencoder(SAE) 분야에서 이미 제안된 바 있으며[4][5], 배치 전체의 통계 정보를 활용함으로써 개별 샘플 기준 방법 대비 안정적인 희소 표현을 유도할 수 있음이 보고되었다. 그러나 기존 SAE 기반 연구들은 결정론적 표현 학습을 대상으로 하며, 확률적 잠재 분포와 Kullback - Leibler 발산(KLD)을 포함하는 VAE 구조에 batch Top-K 기법을 직접 적용하기에는 한계가 있다. 특히 VAE에서는 잠재 희소화가 재구성 성능 및 잠재 분포 정합에 미치는 영향을 함께 고려해야 한다.

이에 본 논문에서는 SAE 분야에서 제안된 batch 단위 Top-K 희소화 개념을 VAE의 잠재 공간에 적용하고, 확률적 잠재 표현의 특성을 고려한 희소화 및 추론 기법을 제안한다. 또한 테스트 단계에서 지수 이동 평균(Exponential Moving Average, EMA)을 활용한 추론 방식을 도입하여, batch Top-K 적용 시 발생할 수 있는 추론 변동성을 완화하고자 한다.

제안한 방법은 희소 잠재 표현이 재구성 성능, 생성 분포 특성 및 잡음 환경에서의 강건성에 미치는 영향을 분석하는 것을 목표로 하며, 실험을 통해 그 특성과 한계를 고찰한다.

Ⅱ. 본론

2.1 Batch 단위 Top-K 기반 잠재 희소화

기존 Top-K 기반 희소화 기법은 주로 개별 샘플의 잠재 값 또는 고정된 기준에 따라 활성 차원을 선택한다. 반면, batch 단위 Top-K 기법은 하나의 샘플이 아닌 배치 전체에서 잠재 값의 분포를 고려하여 중요도를 산정함으로써, 보다 안정적인 희소 표현을 유도한다. 본 논문에서는 VAE의 잠재 변수 z에 대해 배치 전체에서의 잠재 값 크기를 기반으로 중요도를 계산하고, 해당 중요도에 따라 Top-K 기준으로 잠재 차원을 선택한다. 이를 통해 배치 구성에 따른 변동성을 완화하고, 잠재 공간의 활용을 보다 구조적으로 제한할 수 있다.

그러나 VAE는 확률적 잠재 분포를 기반으로 학습되기 때문에, 단순한 batch Top-K 적용은 재구성 성능 저하나 잠재 분포 불안정성을 유발할 수 있다. 특히 잠재 차원의 강제적인 비활성화는 Kullback - Leibler 발산(KLD)에 직접적인 영향을 미치며, 학습 과정의 불안정성을 초래할 가능성이 있다. 이러한 점에서 VAE 구조에 적합한 희소화 적용 방식이 필요하다.

2.2 확률적 잠재 분포를 고려한 학습 전략

본 연구에서는 batch 단위 Top-K 희소화를 적용하되, KLD는 잠재 마스크 이전의 평균과 분산을 기준으로 계산하여 확률적 잠재 분포의 정합성을 유지한다.

Batch 단위 Top-K 기법은 학습 단계에서는 효과적이나, 테스트 단계에서는 배치 구성에 따라 희소 패턴이 달라질 수 있다는 한계를 가진다. 이를 해결하기 위해 본 논문에서는 학습 과정에서 계산된 잠재 차원 중요도에 EMA를 적용하고, 테스트 단계에서는 해당 EMA 기반 중요도를 사용하여 잠재 차원을 선택한다. 이 방식은 단일 배치에 의존하는 추론 방식 대비 희소 패턴의 변동성을 완화하며, 보다 일관성 있는 잠재 표현을 생성할 수 있도록 한다.

결과적으로 제안한 방법은 SAE 분야에서 제안된 batch Top-K 개념을 VAE의 확률적 잠재 구조에 맞게 확장하고, EMA 기반 추론을 통해 희소 잠재 표현의 안정성을 확보하고자 한다.

2.3 Batch 단위 Top-K 적용에 따른 실험 결과 분석

제안한 방법의 특성을 분석하기 위해 MNIST 데이터셋을 사용하여 실험을 수행하였다. 비교 알고리즘으로는 대표적인 VAE 변형 모델인 β-VAE를 사용하였다.

2.3.1 생성 및 재구성 성능 분석

본 실험에서 두 방법 모두 잠재 차원 수는 64로 통일하였으며, β-VAE의 β 값은 0.001로 설정하였다. Batch 단위 Top-K 기반 VAE의 경우 이 중 상위 8개의 잠재 차원(K=8)만을 활성화하도록 하였다.

성능 평가지표로는 FID, PSNR, SSIM을 사용하였다. FID는 생성 이미지의 분포 유사성을 평가하기 위한 지표로 사용되었으며, PSNR과 SSIM은 입력 이미지와 이에 대응하는 재구성 이미지 간의 품질과 구조적 유사성을 정량적으로 분석하기 위해 사용되었다.

	Beta-VAE	Batch Top-K
FID ↓	41.307	34.373
PSNR ↑	21.538	14.171
SSIM ↑	0.9171	0.749

표 1. MNIST Dataset 생성/재구성 정량 평가

표 1은 MNIST 데이터셋을 대상으로 β-VAE와 제안한 Batch 단위 Top-K 기반 VAE의 정량적 성능을 비교한 결과이다.

실험 결과, 제안한 Batch 단위 Top-K 기반 VAE는 FID 지표에서 β-VAE 대비 더 낮은 값을 기록하여, 생성 이미지의 분포 유사성 측면에서는 개선된 성능을 보였다.

반면, PSNR과 SSIM 지표에서는 β-VAE가 Batch 단위 Top-K 기반 VAE 대비 높은 값을 나타내어, 재구성 이미지의 품질 및 구조적 유사성 측면에서는 우수한 성능을 보였다. 이는 Batch 단위 Top-K 희소화로 인해 잠재 표현의 자유도가 감소하면서, 입력 이미지의 세부 정보를 정확히 복원하는 데 제약이 발생한 것으로 분석하였다.

2.3.2 잡음 환경에서의 잠재 표현 강건성 분석

잡음 환경에서는 재구성 이미지의 픽셀 수준 품질이 잠재 표현의 의미적 정보 보존 능력을 충분히 반영하지 못할 수 있다. 이는 PSNR이나 SSIM과 같은 지표가 국소적인 픽셀 차이나 구조적 유사성에 민감한 반면, 객체의 형태나 클래스와 같은 고수준 의미 정보를 직접적으로 평가하지는 않기 때문이다. 특히 잡음이 추가된 입력의 경우, 재구성 과정에서 일부 세부 픽셀 정보가 손실되더라도 전체적인 구조나 의미가 유지될 수 있으며, 이러한 경우 픽셀 기반 지표만으로는 잠재 표현의 정보 보존 특성을 정확히 판단하기 어렵다. 따라서 본 연구에서는 재구성 결과가 분류 성능에 미치는 영향을 분석함으로써, 잡음 환경에서 희소 잠재 표현이 의미적 정보를 얼마나 안정적으로 유지하는지를 간접적으로 평가하였다. 이러한 평가는 희소 잠재 표현의 잡음 강건성을 분석한 기존 연구[2]의 실험 설정과도 일관된다.

표 2는 MNIST 데이터셋에 가우시안 잡음을 추가한 후, 재구성된 이미지를 기반으로 분류 정확도를 측정하여 재구성 강건성을 평가한 결과를 나타낸다. 잡음의 강도는 표준편차 σ=0.3, 0.5, 0.7로 설정하였으며, 잡음 수준이 증가할수록 재구성 난이도가 높아지도록 구성하였다. 비교 모델로는 β-VAE를 사용하였다.

본 실험에서 두 방법 모두 잠재 차원 수는 256으로 통일하였으며, β-VAE의 β 값은 0.001로 설정하였다. Batch 단위 Top-K 기반 VAE의 경우 상위 4개의 잠재 차원(K=4)만을 활성화하였다.

	σ = 0.3	σ = 0.5	σ = 0.7
Beta-VAE	0.9125	0.8493	0.6486
Batch Top-K	0.8992	0.8329	0.7841

표 2. 잡음 추가 MNIST 재구성 강건성 성능(분류 정확도)

실험 결과, 두 방법 모두 잡음 강도가 증가함에 따라 분류 정확도가 감소하는 공통적인 경향을 보였다. 상대적으로 낮은 잡음 수준인 σ=0.3 및 σ=0.5 조건에서는 β-VAE가 Batch 단위 Top-K 기반 VAE 대비 더 높은 분류 정확도를 기록하여, 재구성된 이미지의 정보 보존 측면에서 우수한 성능을 보였다. 이는 batch 단위 Top-K 희소화로 인해 잠재 공간의 활성 차원이 제한되면서, 비교적 낮은 잡음 환경에서는 일부 세부 정보 손실이 분류 성능에 영향을 미친 결과로 해석할 수 있다.

반면, 잡음 강도가 가장 높은 σ=0.7 조건에서는 Batch 단위 Top-K 기반 VAE가 β-VAE 대비 더 높은 분류 정확도를 기록하였다. 이는 강한 잡음 환경에서는 불필요하거나 노이즈에 민감한 잠재 차원이 제거된 희소 잠재 표현이, 의미적 정보 중심의 안정적인 표현을 유지하는 데 유리하게 작용했음을 보여준다.

Ⅲ. 결론

본 논문에서는 SAE 분야에서 제안된 batch 단위 Top-K 희소화 개념을 VAE 모델의 잠재 공간에 적용하고, 확률적 잠재 분포의 특성을 고려한 학습 및 추론 방법을 제안하였다. 제안한 방법은 배치 전체의 잠재 값 분포를 기반으로 잠재 차원을 선택함으로써, 기존 개별 샘플 기반 희소화 기법 대비 보다 구조적인 잠재 표현을 형성하고자 하였다.

또한 batch Top-K 적용 시 발생할 수 있는 추론 단계의 변동성을 완화하기 위해, 학습 과정에서 추정된 잠재 차원 중요도에 EMA를 적용하고 이를 테스트 단계의 잠재 선택 기준으로 활용하였다. 이러한 설계를 통해 확률적 잠재 분포를 갖는 VAE 구조에서도 batch 단위 희소화를 적용할 수 있는 하나의 방법을 제시하였다.

실험 결과, 제안한 기법은 기존 방법 대비 더 나은 생성 품질을 달성하였으나 재구성 품질은 저하되는 것을 확인하였다. 또한, 강한 희소 제약 조건 하에서 잡음이 포함된 환경에서도 제안한 방법은 잠재 표현의 강건성 측면에서 기존 방법과 유사하거나 일부 조건에서는 더 우수한 성능을 보였다. 향후 연구에서는 batch Top-K 보다 정교한 중요도 정의와, 다양한 생성 모델 구조로의 확장을 통해 희소 잠재 표현의 활용 가능성을 추가적으로 탐구할 수 있을 것으로 기대된다.

ACKNOWLEDGMENT

"이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-지역지능화혁신인재양성사업의 지원을 받아 수행된 연구임 (IITP-2026-RS-2020-II201741)"

참 고 문 헌

[1] Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes." arXiv preprint arXiv:1312.6114 (2013).

[2] Li, Hanao, and Tian Han. "Enforcing sparsity on latent space for robust and explainable representations." Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2024.

[3] Tonolini, Francesco, Bjørn Sand Jensen, and Roderick Murray-Smith. "Variational sparse coding." Uncertainty in Artificial Intelligence. PMLR, 2020.

[4] Bussmann, Bart, Patrick Leask, and Neel Nanda. "Batchtopk sparse autoencoders." arXiv preprint arXiv:2412.06410 (2024).

[5] Leask, Patrick, et al. "Sparse autoencoders do not find canonical units of analysis." arXiv preprint arXiv:2502.04878 (2025).