트랜스포머의 일반화 성능에 영향을 주는 로스 랜드스케이프 연구

최민기 [1], 이소은 [1], 허종욱 [1*]
[1] 한림대학교 소프트웨어학부
chlalsrl98@naver.com, dlth508@naver.com, juhou@hallym.ac.kr

A Study on Loss Landscape Affecting the Performance Generalization of Transformer

MinGi Choi[1], So-Eun Lee[1], Joug-Uk Hou[1*]
[1]Division of Software, Hallym University

요 약

뉴럴 네트워크는 학습에 사용하는 파라미터를 문제에 맞게 최적화하여 일반화 성능을 향상시키는 것이 목적이다. 선행 연구들은 다차원의 로스 랜드스케이프(loss landscape)를 시각화하는 방법을 탐구하며, 모델의 일반화 측면에서 어떤 영향을 주는지 탐구한다. 하지만 아직까지 로스 랜드스케이프가 근본적으로 일반화 성능에 어떠한 영향을 주는지 잘 알려져 있지 않으며, 평평하거나 경사진 로스 랜드스케이프 중 어떤 형태가 일반화 성능에 더 효과적인지 여러 의견이 나뉜다. 따라서 우리는 로스 랜드스케이프가 일반화 성능과 연관 있음을 실험을 통해 파악한다. 나아가 비전 문제에서 MSA(multi-head self-attention) 레이어를 기반으로 구성된 트랜스포머 구조를 사용해 작은 유도 편향(inductive bias)을 가지며 소규모 데이터 셋 체제에서의 단점을 보완한다. 결론적으로 평평한 로스 랜드스케이프가 일반화 성능에 긍정적인 영향을 끼친다는 것을 관찰한다.

1. 서론

인공신경망을 학습시키기 위해서는 로스를 최소화하는 것이 목표이며 이는 일반화 성능과 관련이 있다. 실제 로스는 대부분 고차원의 비볼록 함수이고 이론적으로 이를 최소화하는 경로를 찾기 어렵다. 모델이 전역 최소값으로 수렴할지의 여부는 모델의 구조, 사용된 옵티마이저, 파라미터 초기화 방식에 따라 달라진다. 그러나 모델이 개별적인 요소들에 의해 받는 영향을 파악하기는 쉽지 않으며, 많은 연구들은 로스 랜드스케이프를 시각화하여 여러 변화에 따른 특성을 실험적으로 분석하는 게 도움이 된다고 주장한다 [1, 2]. 여기서는 로스 랜드스케이프가 평평한 최소값이 좋다는 의견 [3]과 경사진 최소값이 좋다는 의견으로 나뉘며 [4], 아직까지 어떤 형태의 로스 랜드스케이프가 좋은지 명확히 답을 내릴 수 없다. 본 논문에서는 평평한 로스 랜드스케이프의 형태가 좋다는 주장을 실험을 통해 뒷받침하며, MSA(multi-head self attention) 메커니즘을 사용하여 견고한 네트워크를 구성하기 위해 선행 연구를 자세히 조사한다. 그 중 평평한 형태의 로스 랜드스케이프를 가지는 대표적인 네트워크로 트랜스포머가 존재한다.

자연어 처리 분야에서 처음으로 제안된 트랜스포머는 현재까지 각광받고 있는 구조이다 [5]. 2021 년 Dosovitskiy 연구팀이 제안한 비전 트랜스포머(ViT)는 이미지를 패치 단위로 자르고, 대규모 데이터 셋으로 학습하여 비전 분야에서의 성공을 거뒀다 [6]. 하지만, 트랜스포머는 매우 약한 유도 편향(inductive bias)을 가지게 되어 소규모 데이터 셋에서 학습이 잘 되지 않는다는 단점이 존재한다.

이와 동시에, 트랜스포머의 MSA 와 컨볼루션 레이어를 결합하는 네트워크 구조도 다양하게 연구되었다 [1, 7, 8]. 우리는 앞선 연구들을 기반으로 로스 랜드스케이프를 볼록과 비볼록, 평평함과 경사짐, 두 가지 기준으로 나누어 모델의 일반화 성능을 평가한다 [4].

2. 연구 배경

2.1. 로스 랜드스케이프

로스 랜드스케이프는 모델의 파라미터 ω 에 따른 로스 값의 변화를 의미한다. 이는 고차원이기 때문에 시각화를 위해 저차원으로 차원을 축소하는 PCA 기법을 사용한다. 로스 랜드스케이프의 시각화를 통해 ω 에 따른 로스의 등고선을 파악할 수 있으며, 로스 랜

* 교신 저자

드스케이프를 잘 파악한다면 뉴럴 네트워크 학습을 이해하는데 직관적인 도움을 준다 [2, 3, 9].

2.2. 로스 랜드스케이프와 일반화 성능의 관련성 파악

2017 년 Dinh 연구팀은 동일한 네트워크로 가중치 θ 를 변경하며 동일한 결과를 내뱉는 모델이 다른 가중치(θ)을 가질 수 있다고 설명한다. 다시 말해, 하나의 모델이 경사지거나 평평한 로스 랜드스케이프을 가질 수 있기 때문에 일반화와 관련이 없다고 주장한다 [4]. 반대로 2018 년 Li 연구팀은 로스 랜드스케이프의 경사진 정도를 평가하기 위해 모델의 파라미터 ω 를 토대로 2 차 미분 값들로 구성된 헤센 행렬(hessian matrix)의 고윳값(eigenvalue)을 이용하여 로스 랜드스케이프의 볼록함과 경사진 정도를 확인한다 [1]. 헤센 고윳값이 모두 양수 값을 가지면 아래로 볼록한 형태가 되고, 음수 값이 존재한다면 비볼록한 로스 랜드스케이프가 형성된다. 또한 헤센 고윳값의 크기가 클수록 경사지며, 크기가 작을수록 평평한 형태가 된다. 최근 많은 연구들은 로스 랜드스케이프가 평평할 때 일반화 성능이 좋아진다고 말하며, 최적화, 배치 정규화 등을 사용하여 개선된 일반화 성능을 낼 수 있는 방법을 제안한다 [2, 3, 9].

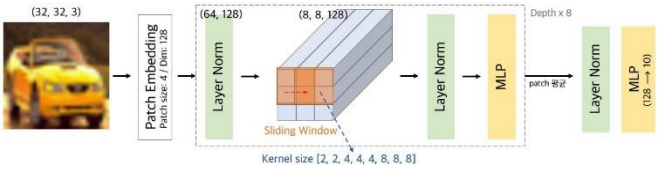
2.3. 멀티헤드 셀프 어텐션(Multi-head self attention)

셀프 어텐션(self-attention)은 컨볼루션 기반 네트워크와 다르게 전체 입력 정보를 가지고 정보를 추출하기 때문에 약한 유도 편향(inductive bias)를 가진다. Park 연구팀은 트랜스포머와 Resnet 을 비교하며, MSA 레이어가 고주파 신호를 감소시키기 때문에 저주파 필터라고 주장하고, MSA 레이어를 사용했을 때 헤센 고유값이 음의 값을 가지므로 로스 랜드스케이프가 비볼록하다고 말한다 [7]. 반대로, 컨볼루션 레이어는 고주파 신호를 증폭시키는 고주파 필터이며, 결론적으로 MSA 레이어와 서로 상호보완적 관계라 주장한다.

3. 본론

3.1. 슬라이딩 윈도우 멀티헤드 셀프 어텐션

우리는 ViT 에서 사용된 전역적인 MSA 와 달리 지역적인 MSA 기반의 슬라이딩 윈도우(sliding window) MSA 방법을 제안하며, 컨볼루션 레이어의 진행 과정을 토대로 고안하였다. 슬라이딩 윈도우 MSA 는 기존 전역적인 MSA 보다 적은 계산 복잡도를 가진다. 윈도우 개념은 컨볼루션 레이어의 커널 크기와 동일하며, 이는 네트워크의 끝단으로 갈수록 전역적인 정보를 볼 수 있도록 윈도우 크기를 증가시킨다. 이 방법은 윈도우 크기를 증가시킬 필요 없이, 깊이가 깊어질수록 자동적으로 전역적인 정보를 파악할 수 있다. 하지만 이는 임베딩 된 특징들의 사이즈가 매우 중요하며, 네트워크를 깊게 쌓을 수 없다는 단점이 존재한다. 이러한 단점은 학습 시 성능 저하의 원인이 된다.



(그림 1) 제안하는 네트워크 구조

3.2. 전체적인 네트워크

제안된 네트워크는 그림 1 와 같다. 이미지를 기존 ViT 와 같이 패치 임베딩을 적용한 후, 트랜스포머의 인코더의 입력으로 사용한다. 기존 MSA 와 다르게 슬라이딩 윈도우를 사용하고, 각 MSA 의 윈도우 사이즈를 [2, 2, 4, 4, 4, 8, 8, 8] 깊이 순으로 경험적으로 설정하였다. 이는 지역적인 영역부터 전역적인 영역까지 순차적으로 학습이 가능하기 위함이다. 슬라이딩 윈도우 MSA 는 기존 전역적인 MSA 보다 유도 편향(inductive bias)가 강하고, 소규모 데이터 셋 체제에서 보다 높은 견고성을 갖기를 기대한다.

4. 실험 결과

우리는 먼저 선행 연구의 실험들을 기반으로 로스 랜드스케이프와 일반화 성능이 얼마나 관련이 있는지 확인한다. 또한, 일반화 성능 개선을 위해 옵티마이저의 역할을 관찰한다. 최종적으로 제안한 네트워크가 소규모 데이터 셋 체제에서도 일반화 성능이 높아질 수 있는지 확인한다. 실험은 공개된 소규모 데이터 셋 체제인 CIFAR-10 을 사용하고, 대표적으로 많이 사용되는 네트워크인 Resnet 과 ViT 를 관찰한다. 모든 환경을 동일하게 세팅하였으며, 학습 시 batch size=256, learning rate=3e-4, epochs=200, single NVIDIA RTX 3060 GPU 으로 실험을 진행한다.

4.1. 네트워크의 로스 랜드스케이프와 일반화

<표 1> 일반화 성능 결과

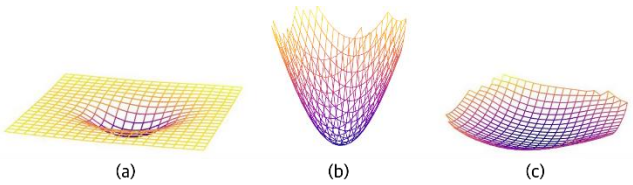
model	Train loss	Valid loss	acc
ViT_patch2	0.707	0.901	70.7
ViT_patch4	0.597	**0.739**	**76.4**
ViT_patch8	0.683	0.844	71.1
Resnet-18	0.448	**0.505**	**83.3**

본 논문에서는 Park 연구팀이 제안한 여러 모델 간의 스케일에 불변하는 특징을 보완하는 filter-wise normalization 기법을 사용하여 시각화한다 [7]. 그림 2 은 소규모 데이터 셋으로 학습한 ViT 의 로스 랜드스케이프가 Resnet 보다 경사진 것을 확인할 수 있으며, 표 1 에서 트랜스포머 기반 모델의 일반화 성능이 더 낮고 학습이 어렵다는 사실을 확인했다.

4.2. 옵티마이저에 따른 실험 결과

본 실험에서는 일반화를 극대화하기 위해 기존에 제안된 옵티마이저에 따른 성능 차이를 확인한다. 표 2

를 보았을 때, 흔히 사용되는 Adam 옵티마이저보다 경사진 정도를 로스 값에 반영한 SAM 옵티마이저가 더 개선된 성능을 보인다. 이는 소규모 데이터 셋 체제에서 학습이 어려운 트랜스포머 구조의 일반화 성능을 높여주는데 큰 역할을 하는 것이라고 볼 수 있다 [4].



(그림 2) 각 모델의 로스 랜드스케이프 시각화 결과
(a) 대규모 데이터 셋(ImageNet)에서 사전 학습된 ViT
(b) 소규모 데이터 셋에서 학습된 ViT-patch4
(c) 소규모 데이터 셋에서 학습된 Resnet-18

<표 2> 옵티마이저에 따른 실험 결과

Ours	Optimizer	Train loss	Valid loss	acc
ViT_patch4	Adam	0.597	0.739	76.4
ViT_patch4	SAM	0.629	**0.661**	**78.8**
Our_patch4	Adam	1.229	1.108	60.7
Our_patch4	SAM	1.135	**1.058**	**62.3**

소규모 데이터 셋을 기반으로 실험한 표 2 에서는 기존 전역적인 MSA 를 사용하는 ViT 와 일반화 성능을 비교해 본 결과, 윈도우 크기가 작아질수록 성능이 떨어지는 것을 확인하였다. 추가적으로 제안한 슬라이딩 윈도우 MSA 를 사용했을 때, 성능이 현저하게 저하된다. 이는 이미지 크기가 32 x 32 로 저화질 데이터이기 때문에 지역적인 영역을 관찰하는 슬라이딩 윈도우 MSA 를 사용했을 때 성능이 떨어지는 것으로 추정된다.

5. 결론

먼저 우리는 평평한 로스 랜드스케이프가 일반화 성능에 긍정적인 영향을 주는 것을 시각화를 통해 확인한다. 다음으로 학습에 사용하는 옵티마이저에 따른 실험을 제공하여 학습 파라미터와 모델 구조가 일반화 성능에 미치는 영향을 조사한다. 결과적으로 여러 실험과 선행 연구의 조사를 통해 로스 랜드스케이프와 학습 파라미터에 따른 일반화 성능 향상에 대해 약간의 통찰을 얻는다.

하지만 여전히 소규모 데이터 셋 체제를 갖는 비전 문제에서의 트랜스포머 구조로 일반화 성능을 개선하는 것은 어려우며, 제안한 슬라이딩 윈도우 MSA 는 기존의 성능보다 향상되지 않음을 알 수 있다. 게다가 저화질 데이터 셋에서 지역적인 영역을 보는 것이 큰 의미를 가지지 않는 것을 실험을 통해 관찰한다. 시각화한 로스 랜드스케이프는 다차원의 로스 표면(surface)을 축소하기 때문에 제한적이다. 또한 아직까지 각 레이어가 정확히 어떤 역할을 수행하는지 탐지하는 것은 불가능하다. 향후 연구로는 차원 축소로 제한적인 로스 랜드스케이프의 단점을 보완하는 다양한 방법을 모색하고, MSA 레이어가 가진 특성에 대해 면밀히 조사할 계획이다.

사사

이 논문은 2022 년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No.2022R1A4A1033600). 또한, 본 연구는 2022 년 과학기술정보통신부 및 정보통신기획 평가원의 SW 중심대학사업의 연구결과로 수행되었음 (20180002160301001).

참고문헌

[1] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. Advances in neural information processing systems, 31, 2018.
[2] Zhewei Yao, Amir Gholami, Kurt Keutzer, and Michael W Mahoney. Pyhessian: Neural networks through the lens of the hessian. In 2020 IEEE international conference on big data (Big data), pages 581–590. IEEE, 2020.
[3] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. arXiv preprint arXiv:2010.01412, 2020.
[4] Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In International Conference on Machine Learning, pages 1019–1028. PMLR, 2017.
[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017.
[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.
[7] Namuk Park and Songkuk Kim. How do vision transformers work? arXiv preprint arXiv:2202.06709, 2022.
[8] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 22–31, 2021.
[9] Yaoqing Yang, Liam Hodgkinson, Ryan Theisen, Joe Zou, Joseph E Gonzalez, Kannan Ramchandran, and Michael W Mahoney. Taxonomizing local versus global structure in neural network loss landscapes. Advances in Neural Information Processing Systems, 34:18722–18733, 2021.