

# How Do Vision Transformers Work?

HMTanh, N.T.K.Nguyên, N.T.Tài  
N.T.Thái, H.N.Thạch

Slides : <https://hmthanh.github.io/how-vits-work>

Code : <https://github.com/hmthanh/how-vits-work>

## Đặc điểm nào Self-attention cần cải thiện?

**MSAs (multi-head self-attention)** thì làm phẳng hóa nhưng không làm lồi hàm loss. Ngược lại, Convs thì làm lồi nhưng sắc nét hàm loss.

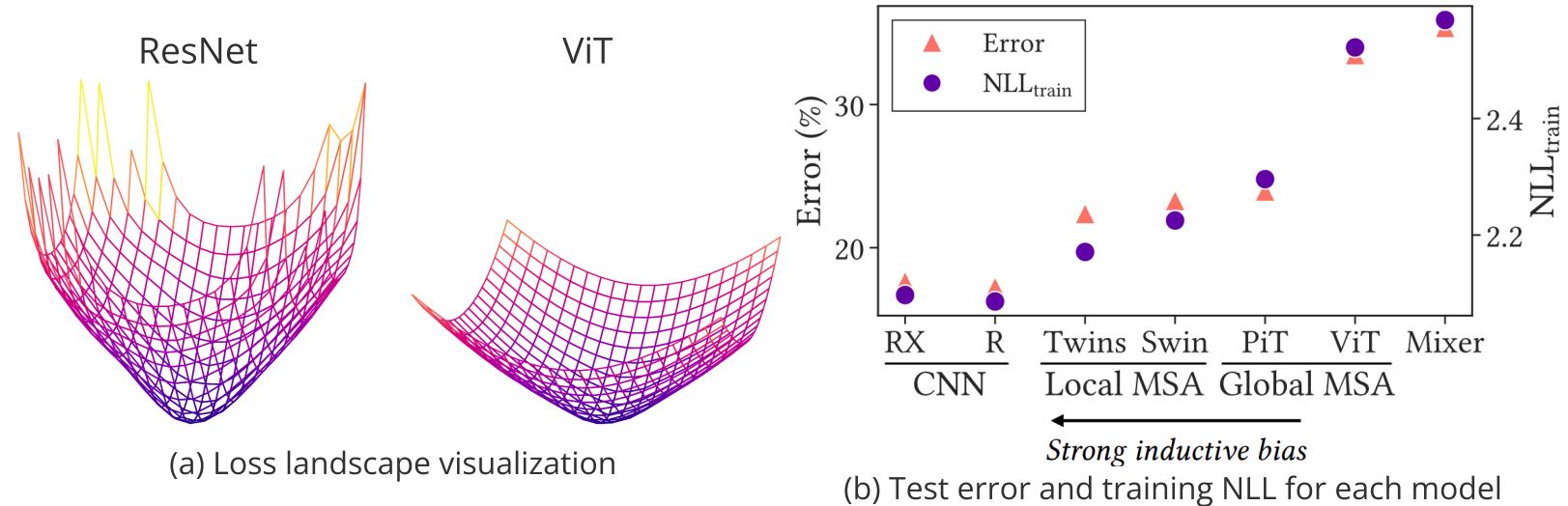


Figure 1: Hình trái hàm thể hiện rằng **ViT** đã làm phẳng hóa hàm loss hơn so với mô hình Resnet. Hình phải: Weak inductive bias (e.g. long-range dependency) làm nhiễu quá trình tối ưu cho NN.

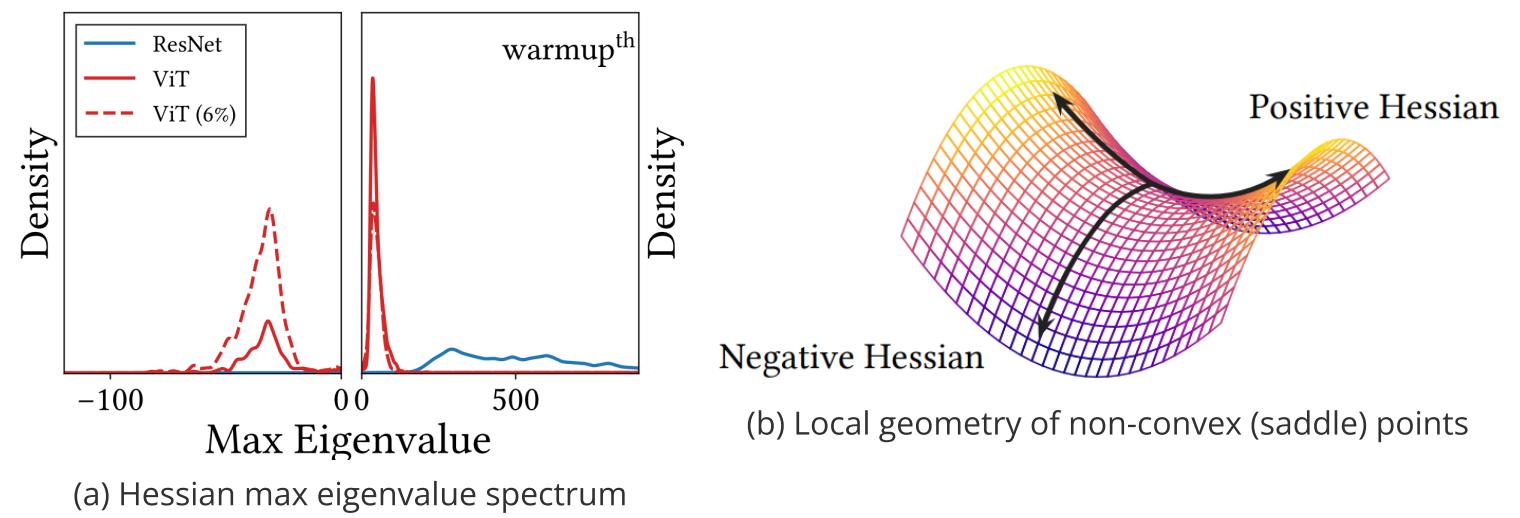


Figure 2: **Hessian max eigenvalue spectra** chỉ ra rằng **MSAs** có ưu và nhược điểm. **ViT** có một số negative Hessian eigenvalues, trong khi **ResNet** chỉ có một ít. Kích thước của **ViT**'s positive Hessian eigenvalues thì nhỏ.

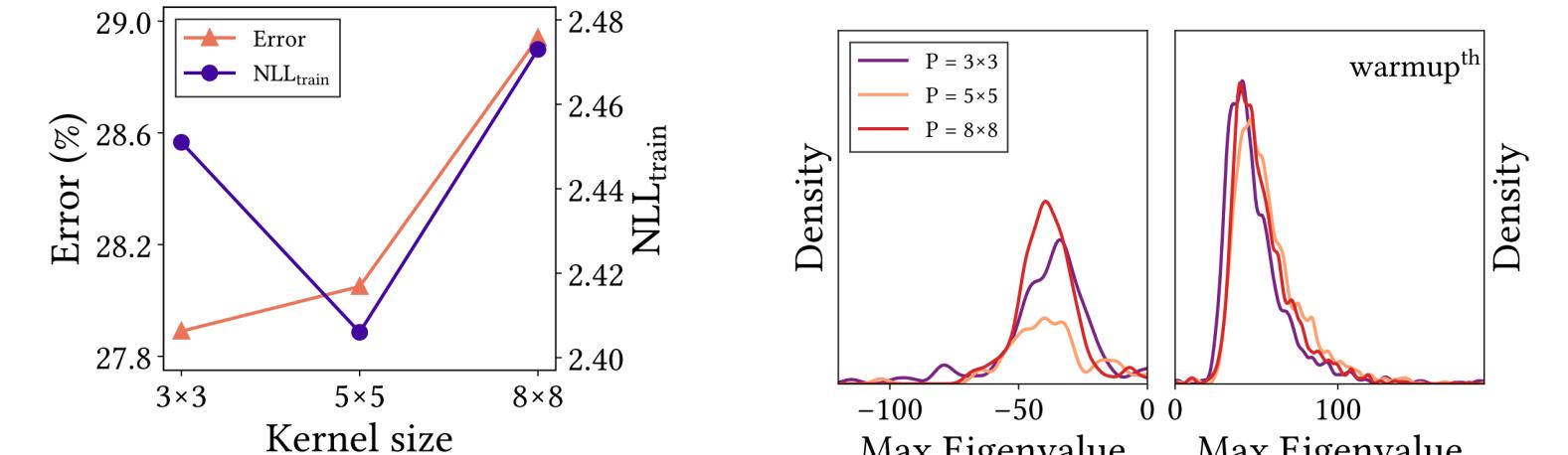


Figure 3: Đặc trưng của **MSA** là data specificity, không phải long-range dependency. **Trái:** Convolutional ViT chứng minh rằng các ràng buộc cục bộ cải thiện ViT. **Phải:** Locality inductive bias bắt dừng lại the negative Hessian eigenvalues.

## Self-Attention có hoạt động như Convolution?

**MSAs** là một hàm lọc low-pass, nhưng **Convs** là hàm lọc high-pass. Chứng tỏ rằng MSAs thì shape-biased, nhưng trái lại Convs thì texture-biased.

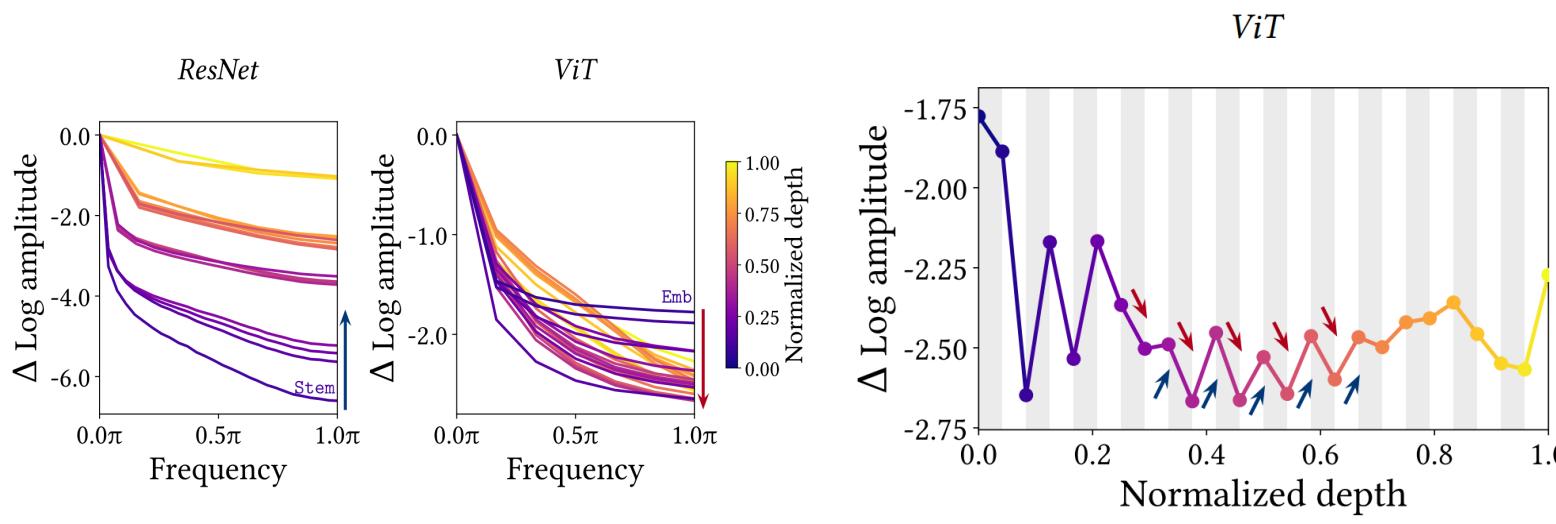


Figure 4: Biên độ log tương đối của **Đặc trưng fourier transformed** map chỉ ra rằng **ViT** có xu hướng giảm các tín hiệu (đặc trưng) high-frequency, trong khi **ResNet** tăng cường chúng. **Phải:** Trong ViT, **MSAs** (vùng xám) generally giảm thành phần high-frequency ( $1.0\pi$ ) của feature map, và Conv/MLPs (vùng trắng) tăng cường chúng.

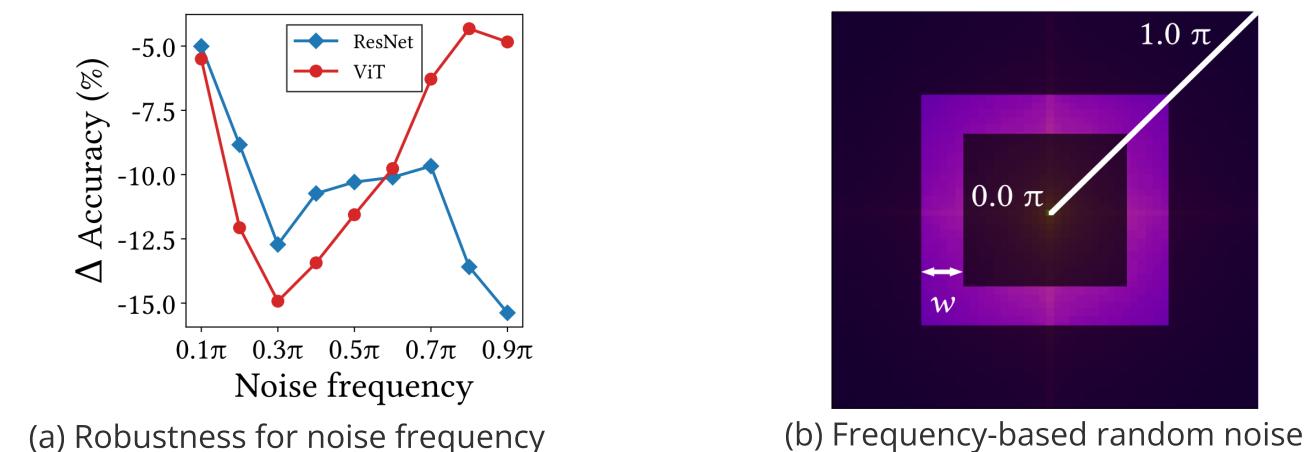


Figure 5: Chúng tôi đo lường **sự giảm độ chính xác đối với nhiều ngẫu nhiên dựa trên tần số (frequency-based random noise)**. **ViT** được tăng cường nhiều high-frequency, trong khi **ResNet** dễ bị tổn thương bởi chúng. Điều này chỉ ra rằng tín hiệu low-frequency và high-frequency được thông tin đến MSAs và Convs.

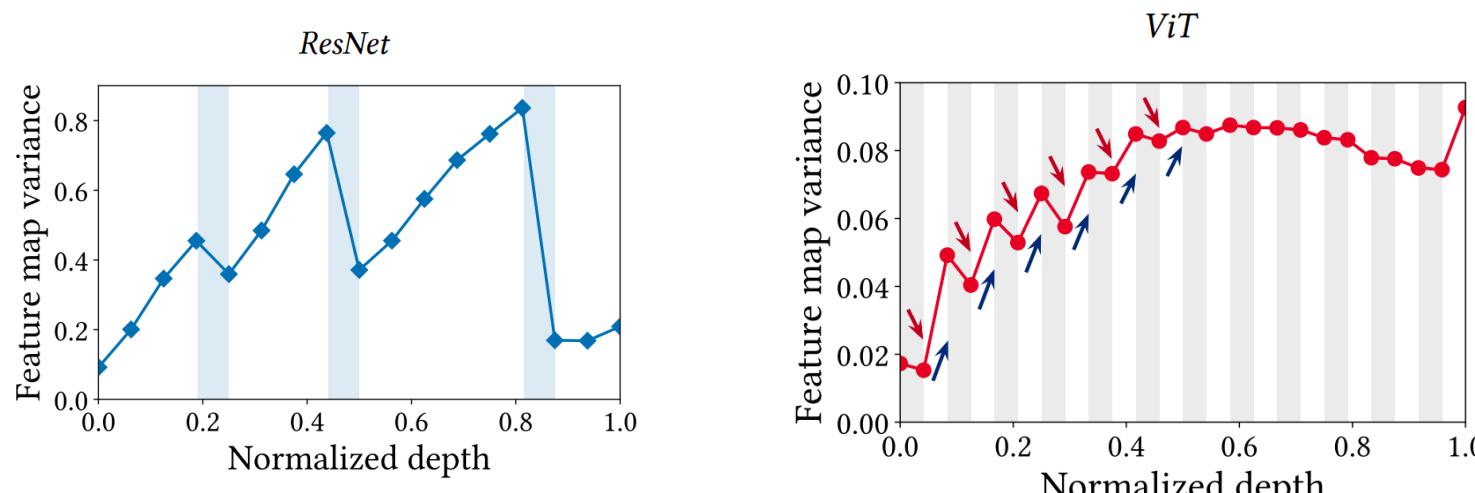


Figure 6: **MSAs** (vùng xám) giảm phương sai của các điểm feature map, nhưng **Convs/MLPs** (vùng trắng) tăng phương sai. Vùng xanh là lớp subsampling. Kết quả ngụ ý rằng **MSAs** tổng hợp feature maps, và **Convs** biến đổi chúng.

## Làm cách nào hòa hợp giữa Conv và Attention

**MSAs** nên ở cuối giai đoạn của mô hình và **Convs** ở đầu giai đoạn của mô hình sẽ cải thiện đáng kể hiệu năng.

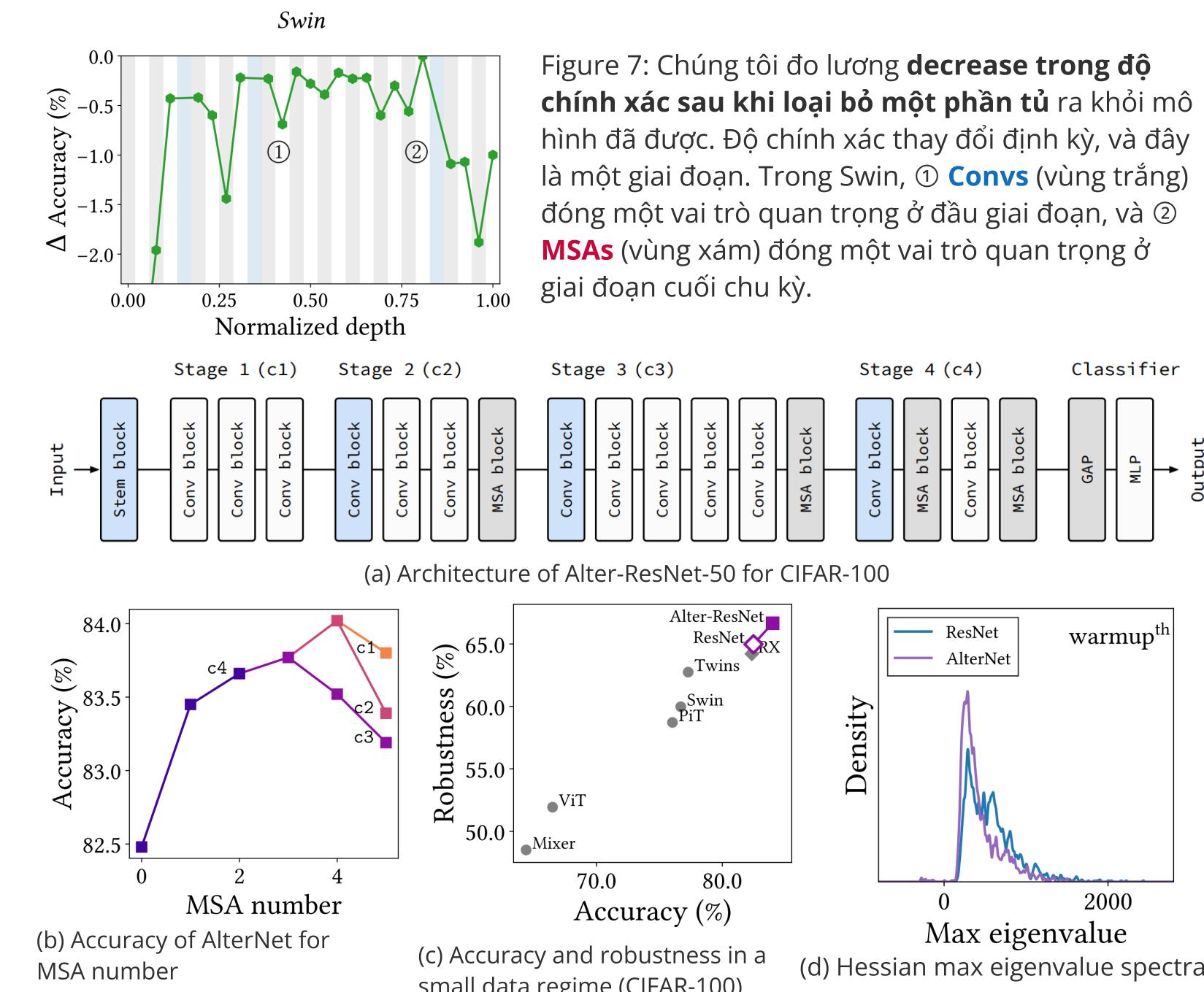


Figure 8: Chúng tôi đề xuất mô hình **AlterNet**, là mô hình mà **Conv** blocks ở cuối chu kỳ được thay thế bằng **MSA** blocks. **AlterNet** vượt trội hơn CNNs dù là ở những tập dữ liệu nhỏ.

**Tổng kết**, những inductive biases thích hợp cải thiện quá trình tối ưu NN, và self-attentions có một chức năng như một spatial smoothing inductive bias.

	Self-Attention	Convolution
Loss Landscape	Flat but not-convex	Convex but sharp
Fourier Analysis	Low-pass filter (shape-biased)	Low-pass filter (shape-biased)
Best practice	Ở cuối mô hình	Ở đầu mô hình