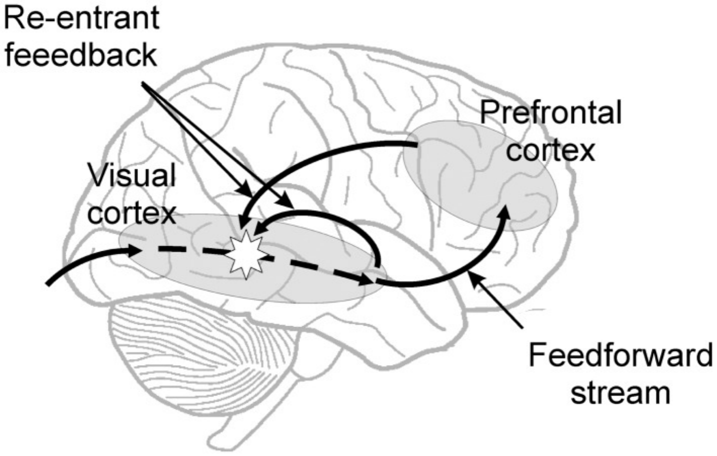
My Current Research

Jacob Fein-Ashley

Contextual Feedback Loops Amplifying Deep Reasoning with Iterative Top-Down Feedback

Your Brain is not a Strictly Feedforward Mechanism

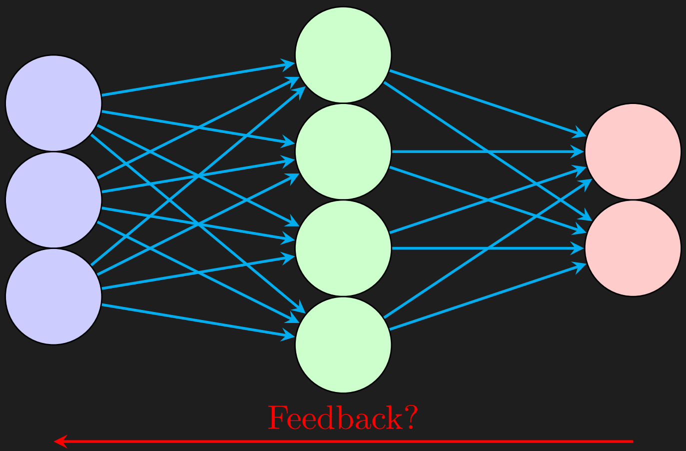


Human brains are made up of about 60% feedback connections.

Fire and Pain: Learning Through Feedback

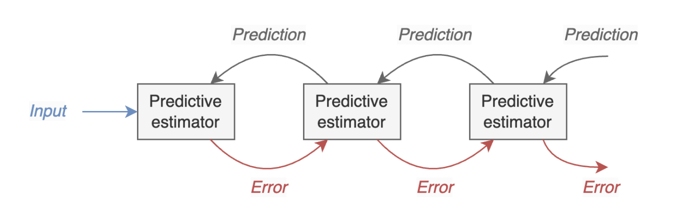
When you get burned, your brain uses feedback signals to correct mistakes and avoid touching fire in the future. This learning process is guided by constant error-checking and adapting to new information.

The Single Pass Dilemma — Feed Forward Networks



Standard neural networks rely on a single forward pass—Feedback?

Related: Predictive Coding & Generative Feedback



- Overly optimistic model assumptions.
- High computational complexity.
- Limited empirical evidence.

These models lack contextual grounding, performance on benchmarks, and high memory/compute requirements.

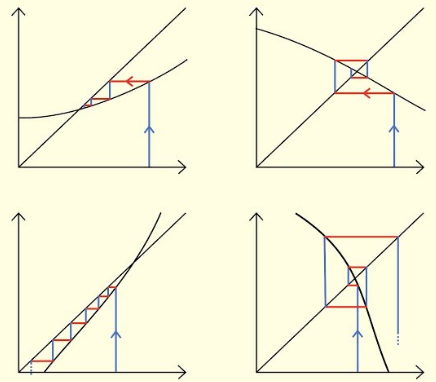
Stop Signs and Context

You don't actually read the stop sign—you recognize it by its context: its placement, distinctive color, and shape.

Iterative Feedback in Action

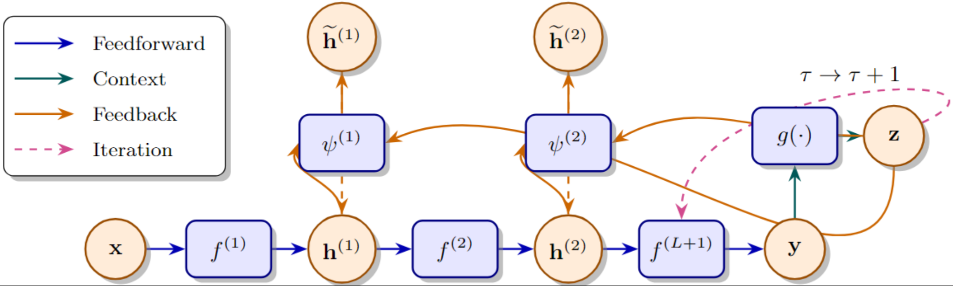
When you see a Ford Explorer on the side of the road, you might initially mistake it for a cop. But by checking for details—like the push bumper, the lights on top, and other distinct markers—you iteratively refine your expectation until your hypothesis is corrected.

Fixed Point Iteration & Banach's Theorem



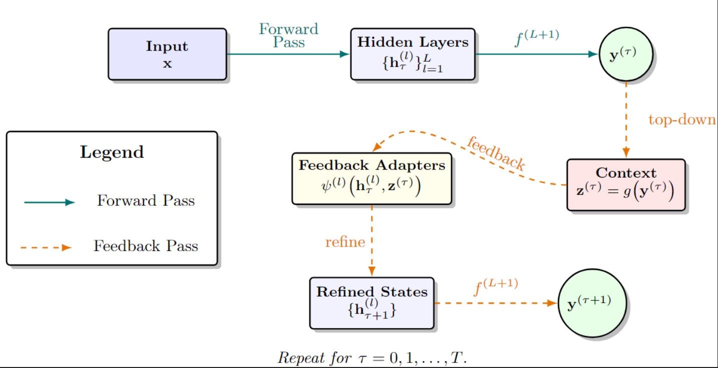
Iterative refinement can be viewed as repeatedly applying a function until convergence. Under the Banach fixed point theorem, if that function is a contraction, the iteration is guaranteed to converge to a unique fixed point.

A Contextual Feedback Loops Framework



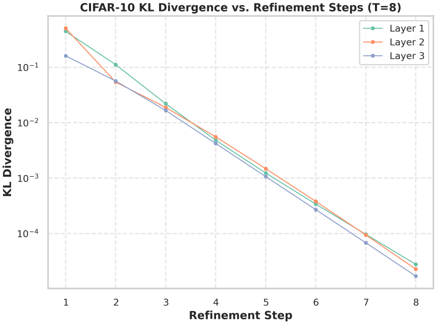
This framework merges feedforward signals with top-down context and iterative feedback steps, allowing each layer to refine its representation over multiple passes instead of relying on a single forward propagation.

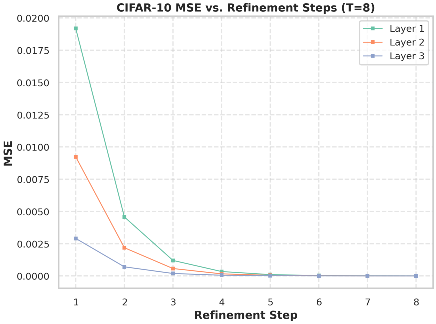
Contextual Feedback Loops: Training vs. Inference



During training, each top-down feedback iteration refines the model's parameters to reduce error, while at inference time, the same iterative process helps the network converge on a stable representation that aligns with the input and context.

Per-Layer Analysis on CIFAR-10





The left image shows the KL divergence per layer on CIFAR-10, while the right displays the MSE per layer. These metrics indicate that our feedback loops balance divergence and reconstruction error at each layer.

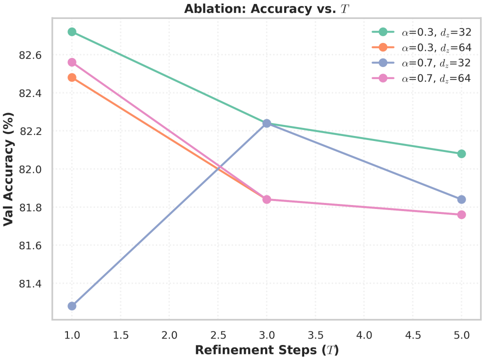
Overall Performance: CIFAR-10, ImageNet, and Speech Commands

Model	Run 1	Run 2	Run 3	Run 4	Run 5
CNN	75.1	75.4	74.8	76.0	75.3
FeedbackCNN	79.2	78.7	79.9	79.6	80.1

Model	# Params (M)	FLOPs (G)	Top-1 (%)	Top-5 (%)
ViT-B/16 (baseline)	86.6	17.5	83.5	96.5
FeedbackViT-B/16	115.0	22.4	**84.2**	**96.9**
ViT-L/16 (baseline)	304.2	61.7	85.1	97.1
FeedbackViT-L/16	398.8	78.2	**85.8**	**97.4**
ViT-H/14 (baseline)	632.1	132.9	85.7	97.6
FeedbackViT-H/14	820.8	210.4	**86.3**	**97.8**

Our approach drastically improves accuracy on CIFAR-10 and ImageNet. We also see significant gains on the Speech Commands dataset, demonstrating broad applicability.

Ablation Study



The ablation study highlights how different parameter settings affect performance. Small changes yield significant differences, underscoring the importance of our model's design choices.

Conclusion and Future Directions

We achieve drastic improvements in accuracy with only a slight overhead. Although I currently lack the computational resources to fully explore these avenues, this approach shows great promise for scaling to large language models and generative models in the future.

The FFT Strikes Back

An Efficient Alternative to Self-Attention

Introduction to FFTNet

- **FFTNet:** A new project that **efficiently** mixes tokens using the Fast Fourier Transform.
- **Scalability:** Scales **much better** than standard self-attention.
- **Efficiency:** Achieves global token mixing in $\mathcal{O}(n \log n)$ time.

Motivation & Convolution Theorem

Key Idea: Convolution in the Token (Time) Domain
Corresponds to
Multiplication in the Frequency Domain.

- **Self-attention** can be seen as a form of global mixing with $\mathcal{O}(n^2)$ complexity.
- By **moving to the frequency domain**, we leverage the convolution theorem to handle these interactions more cheaply:

  Convolution $\longleftrightarrow$ Element-wise Multiplication in Fourier Space.

- **Result:** We obtain global mixing in $\mathcal{O}(n \log n)$ vs. $\mathcal{O}(n^2)$ for standard self-attention.

Global Token Mixing with FFT

Key Idea:

- A Fourier Transform decomposes a sequence of tokens into waves at different frequencies.
- Global interactions arise by combining tokens at all frequencies, revealing overall patterns.

Discrete Fourier Transform (DFT):

$$X_k = \sum_{n=0}^{N-1} x_n \, e^{-i\,2\pi k\, n/N} \quad (k = 0, \ldots, N-1).$$

FFT:

- Efficient algorithm ($O(N \log N)$) for computing the DFT.
- Enables fast global token mixing in large sequences.

Parseval's Theorem and Self-Attention

Parseval's Theorem:

$$\sum_{n=0}^{N-1} |x_n|^2 = \frac{1}{N} \sum_{k=0}^{N-1} |X_k|^2.$$

- Equates energy (sum of squares) in time domain with that in frequency domain.
- Preserves inner products, meaning no global information is lost under the transform.

Connection to Self-Attention:

- In self-attention, token interactions rely on dot products.
- Parseval's Theorem implies that moving to frequency space retains these similarities.

Overview

- **Goal:** Efficiently capture both global and local token interactions.
- **FFT:** Summarizes long-range patterns in $O(N \log N)$.
- **Wavelets:** Zooms in on short-range details.
- **Hybrid:** Combines both for multi-scale context.
- In self-attention, the dot product between tokens measures how much each token attends to every other token. Preserving inner products under the transform means FFT-based mixing retains these same similarities, thereby mimicking global self-attention.

FFT (Global Mixing)

Discrete Fourier Transform (DFT):

$$X_k = \sum_{n=0}^{N-1} x_n \, e^{-\,i\,\frac{2\pi kn}{N}} \quad (k = 0, \ldots, N-1).$$

Intuition:

- Each token x_n is broken down into frequency components.
- Reveals broad periodic patterns spanning the entire sequence.
- **FFT** is a fast algorithm ($O(N \log N)$) to compute these components.

Wavelets (Local Mixing)

Wavelet Function:

$$\psi_{a,b}(t) = \frac{1}{\sqrt{b}}\,\psi\Big(\frac{t-a}{b}\Big),$$

where a is position, b is scale.

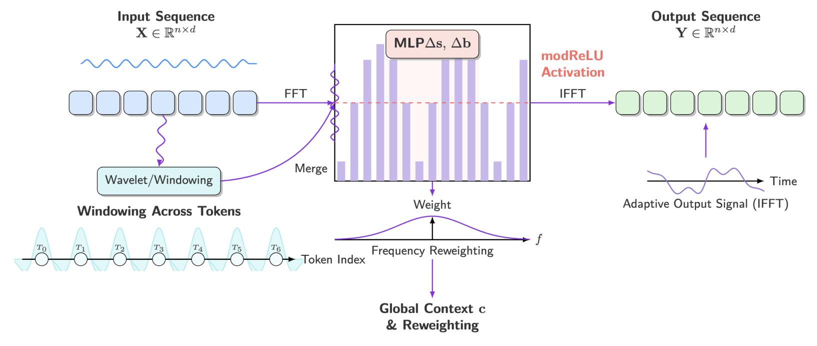
Intuition:

- ψ is a short, localized wave that slides over the sequence.
- Adjusting b changes how zoomed in or out the analysis is.
- Ideal for detecting local features or abrupt changes.

Combining Global and Local

- **FFT:** Captures global context.
- **Wavelets:** Capture local nuances.
- **Result:** A multi-scale representation leveraging both:
  - Long-range patterns,
  - Fine-grained details.

Architecture Diagram



Results

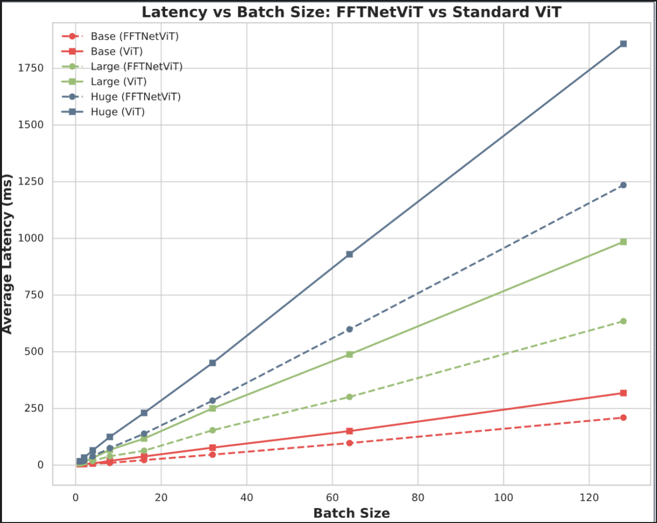
- **Performance:** FFTNet achieves slightly better performance than standard self-attention.
- **Benchmarks:**
  - ImageNet
  - Long Range Arena
- **Comparison:** Outperforms both standard self-attention **and** the baseline FNet.
- **Key Advantage:** Adaptive modulation in the frequency domain allows dynamic token mixing, some studies suggest that operating in the frequency domain allows for better expressivity.

Results

	FFTNetViT (No-Windowing)			FFTNetViT (With Windowing)			ViT		
Variant	FLOPs	Top-1 (%)	Top-5 (%)	FLOPs	Top-1 (%)	Top-5 (%)	FLOPs	Top-1 (%)	Top-5 (%)
Base	22.64	79.6 ↑ 0.2%	94.9 ↑ 0.1%	22.64	79.8 ↑ 0.4%	95.0 ↑ 0.2%	36.65	79.4	94.8
Large	79.92	82.1 ↑ 0.3%	96.2 ↑ 0.2%	79.92	82.3 ↑ 0.5%	96.3 ↑ 0.3%	127.18	81.8	96.0
Huge	166.14	83.2 ↑ 0.3%	96.8 ↑ 0.2%	166.14	83.4 ↑ 0.5%	96.9 ↑ 0.3%	261.39	82.9	96.6

Model	**ListOps**	**Text**	**Retrieval**	**Image**	**Pathfinder**	**Path-X**	**Avg.**
Transformer	36.06	61.54	59.67	41.51	80.38	OOM	55.83
FNet	35.33	65.11	59.61	38.67	77.80	FAIL	55.32
FFTNet (No-Windowing)	37.65	66.01	60.21	42.02	80.71	83.25	58.31
FFTNet (With Windowing)	38.02	66.25	60.64	42.45	80.99	83.64	**58.83**

Latency Comparison



Latency: FFTNet demonstrates lower latency compared to standard self-attention, enabling faster inference.

Conclusion & Future Work

- **Conclusion:** FFTNet demonstrates that global token mixing can be achieved in $\mathcal{O}(n \log n)$ without losing model capacity.
- **Future Directions:**
  - Explore higher-dimensional FFTs for spatial-temporal data.
  - Investigate alternative nonlinear functions in the frequency domain.
  - Apply FFTNet blocks to large-scale language modeling and video tasks.