Why Forward-Forward Hasn't Become the New Paradigm

An Empirical Investigation into Transfer Learning Failures and Bio-Inspired Alternatives

Research Team

FF Algorithm Research

February 2026

Outline

The Biological Implausibility Problem

Backpropagation dominates deep learning, but...

- **Weight Transport Problem**: Requires symmetric forward/backward weights
- **Non-Local Credit Assignment**: Error signals must travel entire network
- **Two-Phase Operation**: Forward pass $\rightarrow$ backward pass separation
- **Biologically Implausible**: No known neural mechanism for exact gradients

The Promise of Forward-Forward

Hinton (2022): A biologically plausible alternative using only local learning rules

The Forward-Forward Algorithm

Core Idea: Replace backward pass with two forward passes

Positive Pass

- Real data + correct label
- Goal: Increase goodness
- $G = \sum_j h_j^2$

Negative Pass

- Real data + wrong label
- Goal: Decrease goodness

Goodness Function

$$G^l = \sum_{j=1}^{n_l} (h_j^l)^2$$

Layer-Local Objective

$$\mathcal{L}^l = \log(1 + e^{-y(G^l - \theta)})$$

$y = +1$ (positive), $y = -1$ (negative)

[Figure: figures/ff-algorithm-diagram.png]

Research Questions

RQ1 How do different **negative sampling strategies** affect FF performance?

RQ2 Can FF models **transfer knowledge** to new tasks?

RQ3 What is the **root cause** of FF's limitations?

RQ4 Can **bio-inspired modifications** improve FF?

Key Finding Preview

Standard FF achieves 94.5% on MNIST but fails catastrophically at transfer learning

RQ1: Negative Sampling Strategies

Which negative sampling strategy works best?

Strategy	MNIST Acc.	Notes
Random Wrong Label	94.5%	Original Hinton method
Hybrid Negatives	93.8%	Mix of strategies
Hard Negatives	91.2%	Confused samples
Gaussian Noise	88.4%	Too easy to detect
Backprop Baseline	*99.2%*	*Upper bound*

- Random wrong label performs best (simple but effective)
- 4.7% gap to backpropagation remains significant

RQ2: The Transfer Learning Paradox

Critical Discovery

Standard FF transfer: 54%

Random initialization: 72%

Experiment: Train on MNIST $\rightarrow$ Transfer to Fashion-MNIST

Method	Fashion-MNIST Acc.
Backprop Transfer	85.2%
Random Initialization	72.0%
Standard FF Transfer	54.0%

FF features are WORSE than starting from scratch!

RQ3: Root Cause — Label Embedding

Why does FF fail at transfer?

The Problem

- FF embeds labels **directly into input**
- Features become coupled to source labels
- Layer 1 learns: "digit 3" not "curved edges"
- New task has different labels $\rightarrow$ features unusable

Evidence

- Label neurons: $10\times$ higher activation variance
- Feature visualization: Label-specific patterns
- Gradient analysis: Label dimensions dominate

[Figure: figures/label-embedding-diagram.png]

Standard FF Input:

$$[\underbrace{x_1, \ldots, x_n}_{\text{image}}, \underbrace{l_1, \ldots, l_k}_{\text{label}}]$$

Features entangled with labels!

Bio-Inspired Variants: Overview

Hypothesis: More biologically realistic $\rightarrow$ better generalization?

Variant	Transfer Change	Status
Three-Factor Hebbian	+1.5%	Marginal
Prospective FF	−13.2%	Failed
PCL-FF	N/A (17.5% acc)	Collapsed
Layer Collaboration	+1.2%	Marginal
CwC-FF	+35%	Success!

- Most bio-inspired changes don't address the core problem
- CwC-FF succeeds by removing label embedding entirely

Three-Factor Hebbian Learning

Idea: Add neuromodulatory signals like dopamine

$$\Delta w_{ij} = \eta \cdot \underbrace{h_i}_{\text{pre}} \cdot \underbrace{h_j}_{\text{post}} \cdot \underbrace{M}_{\text{modulator}}$$

Configurations Tested

- Bottom-up only: -2.1%
- Top-down only: $+1.5\%$
- Bidirectional: -0.8%

Why It Failed

- Modulation doesn't fix label coupling
- Top-down helps marginally
- Still learns task-specific features

Result

Best case: 55.5% transfer (vs 72% random init)

Prospective Configuration (Prospective FF)

Idea: Predictive coding—layers predict future states

Method

- Multiple forward iterations
- Each layer predicts next layer
- Iterative refinement

$$h^{(t+1)} = f(W \cdot h^{(t)} + b)$$

Surprising Result

- 1 iteration: 54.0% (baseline)
- 3 iterations: 48.2%
- 5 iterations: 40.8%

More iterations = worse transfer!

Interpretation

Iterative refinement **amplifies** label-specific features, making transfer even harder

Predictive Coding Layer (PCL-FF)

Idea: Full predictive coding with top-down predictions

Architecture

- Prediction: $\hat{h}^l = W^{l+1 \to l} h^{l+1}$
- Error: $e^l = h^l - \hat{h}^l$
- Update based on prediction error

Catastrophic Failure

- Final accuracy: 17.5%
- Neuron death: 100%
- All hidden units saturated

Post-Mortem Analysis

- Prediction errors explode during training
- Neurons saturate to extreme values
- No gradient signal for recovery

Lesson: Naive predictive coding incompatible with FF framework

Layer Collaboration

Idea: Allow layers to share information via lateral connections

$$G^l_{collab} = G^l + \gamma \cdot G^{l-1} + \gamma \cdot G^{l+1}$$

γ	MNIST Acc.	Transfer Acc.
0.0	94.5%	54.0%
0.3	94.1%	54.8%
0.5	93.8%	55.0%
0.7	**93.2%**	**55.2%**
1.0	91.5%	53.1%

- Best result: +1.2% improvement (still below random init!)
- Trade-off: Source accuracy decreases with higher γ

The Solution: Contrastive without Coupling (CwC-FF)

Key Insight: Remove label embedding entirely

Standard FF

- Input: $[image, label]$
- Positive: correct label
- Negative: wrong label
- Features coupled to labels

CwC-FF

- Input: $[image]$ only (no label!)
- Positive: high channel coherence
- Negative: low channel coherence
- Task-agnostic features

Channel-wise Competition Mechanism

- Neurons compete within channels
- Winners (high activation) $\rightarrow$ positive signal
- Losers (low activation) $\rightarrow$ negative signal
- No labels needed $\rightarrow$ pure feature learning

CwC-FF Results

Transfer Learning Success

CwC-FF Transfer: 89%

(vs 54% standard FF, vs 72% random init)

Method	**Source**	**Transfer**	**Δ vs Random**
Random Init	—	72.0%	0%
Standard FF	94.5%	54.0%	−18%
Backprop	99.2%	85.2%	+13.2%
CwC-FF	91.2%	89.0%	+17%

- CwC-FF: Slightly lower source accuracy, **dramatically better transfer**
- Approaches backprop transfer performance!

Key Insights

1. **Label embedding is the root cause of transfer failure**
    - Features become task-specific, not general-purpose
    - More training $\rightarrow$ worse transfer
2. **Bio-inspired modifications don't help (mostly)**
    - They don't address the fundamental coupling problem
    - Some even make it worse (Prospective FF)
3. **Decoupling labels from features is essential**
    - CwC-FF removes labels from input
    - Learns corruption detection $\rightarrow$ general features
4. **Trade-off: Source vs Transfer performance**
    - CwC-FF: 91.2% source (vs 94.5% standard)
    - But: 89% transfer (vs 54% standard)

Why FF Hasn't Become the New Paradigm

Limitations

- 4.7% accuracy gap to backprop
- Catastrophic transfer failure
- Requires careful negative sampling
- Limited to supervised settings
- No large-scale validation yet

Potential

- Truly local learning rules
- Biologically plausible
- CwC-FF fixes transfer issue
- Potential for neuromorphic hardware
- Energy efficiency benefits

The Path Forward

FF can succeed, but requires **architectural changes** (like CwC-FF), not just biological enhancements

Conclusion

Summary

- FF achieves 94.5% on MNIST but fails at transfer learning
- Root cause: **Label embedding couples features to source task**
- Bio-inspired variants provide only marginal improvements
- CwC-FF solves the problem by removing label embedding

Future Work

- Scale CwC-FF to larger datasets (CIFAR-10, ImageNet)
- Explore hybrid approaches (FF + backprop)
- Neuromorphic hardware implementation
- Unsupervised and self-supervised extensions

Take-Home Message

Biological plausibility alone doesn't guarantee good ML properties. **Understanding failure modes** leads to principled solutions.

References

- Hinton, G. (2022). *The Forward-Forward Algorithm: Some Preliminary Investigations*. arXiv:2212.13345
- Lillicrap, T. P., et al. (2016). *Random synaptic feedback weights support error backpropagation for deep learning*. Nature Communications.
- Whittington, J. C. R., & Bogacz, R. (2017). *An approximation of the error backpropagation algorithm in a predictive coding network*. Neural Computation.
- Sacramento, J., et al. (2018). *Dendritic cortical microcircuits approximate the backpropagation algorithm*. NeurIPS.
- Ororbia, A., & Mali, A. (2023). *The Predictive Forward-Forward Algorithm*. arXiv.

Thank You

Questions?

Code and experiments available at:
`github.com/koriyoshi2041/ff-negative-samples`