# Catastrophic Forgetting: An Extension of Current Approaches

Dhrupad Bharadwaj, Evaristus Ezekwem, Angela V. Teng
Advised by: Dr. Julia Kempe, Artem Vysogorets

DSGA1006: Capstone
Fall 2020

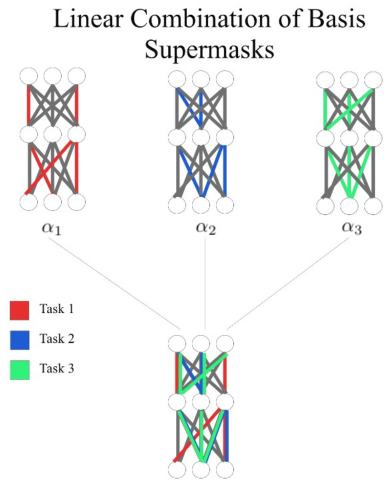
NYU Center for Data Science

## Abstract

Catastrophic forgetting is the phenomenon whereby the performance of learned tasks degreades as new, unseen tasks are learned, specifically in Neural Networks. We extend Wortsman et al.'s (2020) work on continual learning, *Supermasks in Superposition*, by adapting their masking technique to learn new tasks while using significantly fewer additional parameters.

## Introduction and Background

Catastrophic forgetting is an active area of research in continual learning. In order to achieve artificial general intelligence (AGI) it is crucial that learning models are able to learn and remember a wide variety of tasks. Deep learning models have a tendency to forget older tasks once new ones are learnt.




Research done by Wortsman et al. propose the use of Supermasks in Superposition (SupSup), which is "capable of sequentially learning thousands of tasks without catastrophic forgetting." [Wortsman et al., 2020]

Our research extends this approach of using randomly initialized and fixed base networks for each task by optimizing over a linear combinations of supermasks, referred to as the basis masks, to learn new tasks.

**In general, we find that:**
- a mask applied to an arbitrary task t performs no better than the maximum entropy distribution over the output space
- a combination of masks are able to effectively learn new tasks

- while performance is slightly inferior to SupSup, our approach only requires *O(nd)* parameters for each new task, where *d* is the depth of the network and *n* is the number of masks used.

This is a considerable improvement in parameter efficiency compared to SupSup, where each additional task requires storing *O(w)* parameters, where *w* is the number of trainable weights in the network.

## Methods and Algorithm

Our problem to find the optimal $\alpha_i^*$ for a given task $i \in T'$ can be formulated as follows:

$$\alpha_i^* = \arg\min_{\alpha_i} \mathcal{L}\left(y, f\left(x, \left(\frac{1}{|B|}\sum_{t\in T^B}\sum_{d\in D}\alpha_i^{td}\Delta(M_t, d)\right)\oplus W\right)\right)$$

Extending the SupSup model, we take a set of masks $B \subset M^*$, our set of basis masks, and define the set of tasks for which we have a trained mask set, $T^B := \{t \mid M_t^* \in B, \forall\, t \in T\}$ and $T' := T \setminus T^B$. Extending the SupSup model, we take a set of masks B ⊂ M*, our set of basis masks, and define the set of tasks for which we have a trained mask set, TB := {t | Mt* ∈ B,∀ t ∈ T} and T′ := T\TB.

We define $\alpha_i \in \mathbb{R}^{d\times|B|}$ where we have one parameter per $M_t^* \in B$ per layer of the network. Let $\alpha_i^{td}$ be the parameter for task $i \in T'$ on $\Delta(M_t^*, d)$, where $\Delta(M,d)$ is the bit-mask from $M$ in corresponding layer *d* of the network. $D$ is the depth of the network.

**Algorithm 1:** Learning a new task $i \in T'$

**Data:** task $i \in T'$ (random 5-way classification task from CIFAR100)
**Result:** $\hat{\alpha}_i \to \alpha_i^*$
$\hat{\alpha}_i^{t,d} = 1/|B|$; i.e. Uniform prior;
**while** *till convergence* **do**
  $L = 0$; // Set loss on epoch to 0
  $\hat{M}_i = \left(\frac{1}{|B|}\sum_{t\in T^B}\sum_{d\in D}\hat{\alpha}_i^{td}\Delta(M_t, d)\right)$ // Linear combination of masks in $B$
  **for** *x, y in Data, Class* **do** // Forward pass
    $L \leftarrow \mathcal{L}\left(y, f\left(x, \hat{M}_i \oplus W\right)\right)$
  **end**
  Calculate $\frac{\partial L}{\partial \hat{\alpha}_i}$
  $\hat{\alpha}_i \leftarrow \hat{\alpha}_i - \phi\frac{\partial L}{\partial \hat{\alpha}_i}$ // $\phi$ is the learning rate, backprop-step
**end**

## Results

## Conclusion and Future Work

## References