

Tam Bağlı İleri Beslemeli Sinir Ağlarda Softplus Aktivasyon Fonksiyonu İçin Yeni Bir Yol Ağırlık Başlatma Düzeni

A New Weight Initialization Scheme for Softplus Activation Function in Fully Connected FeedForward Neural Networks

¹Mehmet Murat TURHAN , ²Emrah DÖNMEZ , ³Muhammed Fatih TALU

¹Muş Alparslan Üniversitesi, Mühendislik-Mimarlık Fakültesi, Yazılım Müh., Muş, Türkiye

²Bandırma Onyedi Eylül Üniversitesi, Mühendislik ve Doğa Bilimleri Fakültesi, Yazılım Müh., Bandırma/Balıkesir, Türkiye

³İnönü Üniversitesi, Mühendislik Fakültesi, Bilgisayar Müh., Battalgazi/Malatya, Türkiye

¹mm.turhan@alparslan.edu.tr, ²emrahdonmez@bandirma.edu.tr,

³fatih.talu@inonu.edu.tr

Araştırma Makalesi/Research Article

ARTICLE INFO

Article history

Received : 11 June 2025

Accepted : 6 July 2025

Keywords:

Activation Function, Deep Neural Networks, Softplus, Weight Initialization

ABSTRACT

The trainability of deep neural networks becomes challenging as the number of layers increases, primarily due to issues such as vanishing or exploding gradients. In this study, two new weight initialization schemes based on the central limit theorem are proposed for fully connected feedforward neural networks using the Softplus activation function. The first scheme aims to preserve signal statistics only in the forward direction, while the second aims to preserve them in both forward and backward directions. Experiments conducted on the CIFAR-10 and CIFAR-100 datasets demonstrate that preserving only forward signal statistics is insufficient in deep architectures, whereas preserving statistics in both directions significantly improves trainability. Particularly in architectures with 25 hidden layers, successful training was achieved only with the bidirectional initialization scheme. The results reveal that initialization strategies compatible with the dynamics of the activation function play a critical role in enabling the effective training of deep neural networks.

© 2025 Bandırma Onyedi Eylül University, Faculty of Engineering and Natural Science. Published by Dergi Park. All rights reserved.

MAKALE BİLGİSİ

Makale Tarihleri

Gönderim : 11 Haziran 2025

Kabul : 6 Temmuz 2025

Anahtar Kelimeler:

Aktivasyon Fonksiyonu, Derin Sinir Ağları, Softplus, Yol Ağırlık Başlatma

ÖZET

Derin sinir ağlarının eğitilebilirliği, özellikle katman sayısı arttığında, gradyan sönümlenmesi veya patlaması gibi sorunlar nedeniyle zorlaşmaktadır. Bu çalışmada, Softplus aktivasyon fonksiyonu kullanılan tam bağlı ileri beslemeli sinir ağları için, merkezi limit teoremine dayalı iki yeni yol ağırlık başlatma düzeni önerilmiştir. İlk düzen yalnızca ileri yönde sinyal istatistiklerini, ikincisi ise hem ileri hem de geri yönde istatistikleri korumayı hedeflemektedir. CIFAR-10 ve CIFAR-100 veri kümeleri üzerinde yapılan deneylerde, derin mimarilerde yalnızca ileri yöndeki istatistik korumanın yeterli olmadığı, iki yönlü korumanın ise ağırlık eğitilebilirliğini anlamlı şekilde artırdığı gözlemlenmiştir. Özellikle 25 gizli katmanlı ağlarda, yalnızca iki yönlü koruma sağlayan başlatma düzeniyle başarılı eğitim gerçekleştirilebilmiştir. Elde edilen sonuçlar, aktivasyon fonksiyonu dinamiklerine uygun başlatma stratejilerinin, derin sinir ağlarının eğitilebilmesinde belirleyici rol oynadığını göstermektedir.

© 2025 Bandırma Onyedi Eylül Üniversitesi, Mühendislik ve Doğa Bilimleri Fakültesi. Dergi Park tarafından yayınlanmaktadır. Tüm Hakları Saklıdır.

ORCID ID: ¹0000-0003-4497-9102
²0000-0003-3345-8344
³0000-0003-1166-8404

1. GİRİŞ

Derin sinir ağları, sık ağlara kıyasla daha fazla katman içererek özellik uzayını hiyerarşik biçimde dönüştürmekte ve bu sayede problem çözümünde daha etkili sonuçlar elde etmektedir [1]. Görüntü sınıflandırma, nesne tanıma ve doğal dil işleme gibi birçok karmaşık problemin çözümünde derin öğrenme yöntemleri yüksek başarı sağlamaktadır [2-4]. Bu nedenle, zorlu problemlerin çözümünde derin ağlar sık ağlara kıyasla bir gereklilik haline gelmiştir [5]. Yeni algoritmaların geliştirilmesi, donanım teknolojilerindeki özellikle GPU'ların işlem kapasitesindeki artış ve veri kümelerinin çeşitlenmesi, derin öğrenme alanındaki ilerlemeleri hızlandırmıştır [6]. Sinir ağlarında eğitim süreci; giriş verisinin katmanlar boyunca dönüştürülerek çıkış verisinin elde edilmesi, bu çıktının hedef etiketle kıyaslanması ve oluşan hatanın geri yayılım algoritmasıyla ağı parametrelerine gradyan bilgisi aracılığıyla aktarılmasıyla gerçekleşir. Ağı derinliğinin artmasıyla başarı arasında pozitif bir ilişki bulunsa da, gradyan sönümlenmesi veya gradyan patlaması gibi problemler derin ağların eğitilebilirliğini olumsuz etkileyebilir [7, 8]. Gradyan sönümlenmesi, geri yayılım sürecinde gradyanların çok küçülerek anlamsız hale gelmesi; gradyan patlaması ise bunun tersine, gradyanların kontrolsüz biçimde büyümesidir. Bu gibi durumlar ağı yerel minimumlara takılmasına, çözüm uzayından uzaklaşmasına veya hiç eğitilememesine yol açabilir. Bu nedenle derin ağların yukarıda bahsedilen durumlarla karşılaşmaması için başlangıç parametrelerinin dikkatli bir şekilde ayarlanması gerekir.

Gradyan problemlerinin temelinde, ileri ve geri yöndeki sinyallerin ağ boyunca dengeli bir şekilde iletilmemesi yatmaktadır. Bu problemi aşmak için çeşitli yaklaşımlar geliştirilmiştir: Yeni mimari tasarımlar [9, 10], normalizasyon teknikleri [11, 12], yeni aktivasyon fonksiyonları [13-15] ve bu fonksiyonlara özgü ağırlık başlatma stratejileri [13, 16-18]. Tüm bu çalışmaların ortak amacı, eğitim süresince ileri ve geri yönde kararlı sinyal iletimini sağlamak ve anlamlı gradyan bilgisinin sürdürülebilirliğini temin etmektir.

Aktivasyon fonksiyonları, sinir ağlarında sinyalin akışını ve işlenmesini düzenleyerek karmaşık problemlerin çözümüne olanak tanıyan temel bileşenlerdir [19]. Genellikle doğrusal olmayan yapıda olup, evrensel yaklaşıklama teoremini [20] sağlayabilecek nitelikte seçilirler. Uzun süre yaygın olarak kullanılan sigmoidal fonksiyonlar [21, 22] giriş değeri çok büyük veya küçük olduğunda doyumluğa ulaşarak gradyan sönümlenmesine sebep olabilir. Bu sınırlamaların üstesinden gelebilmek için daha güncel aktivasyon fonksiyonları geliştirilmiştir. Özellikle ReLU ve varyantlarının kullanımı, derin ağların eğitilebilirliğini önemli ölçüde artırmıştır [2]. Ancak, aktivasyon fonksiyonlarının ağı eğitilebilirliğine katkısı, ağırlık başlatma yöntemiyle birlikte ele alındığında daha anlamlı hale gelir. Nitekim He ve arkadaşları, 30 katmanlı bir konvolüsyonel mimaride ReLU fonksiyonu ve Xavier başlatma yöntemi [16] kullanarak başarılı bir eğitim sağlayamazken, önerdikleri yeni başlatma tekniğiyle bu sorunu aşabilmişlerdir [13]. Bu durum, aktivasyon fonksiyonunun yapısal özelliklerini dikkate alan özel ağırlık başlatma stratejilerinin derin ağların eğitilebilirliğinde belirleyici olduğunu göstermektedir.

Bu çalışma, Softplus [23] gibi sürekli ve türevlenebilir bir aktivasyon fonksiyonu için, derin sinir ağlarının eğitilebilirliğini artırmayı hedefleyen özel ağırlık başlatma yöntemleri sunması açısından literatüre özgün bir katkı sağlamaktadır. Literatürde ReLU ve benzeri parçalı doğrusal fonksiyonlara yönelik başlatma stratejisi yer almakta iken, Softplus gibi yumuşatılmış doğrusal olmayan fonksiyon için yol ağırlık başlatma stratejisi bulunmamaktadır. Bu amaçla, ileri beslemeli tam bağlı sinir ağı mimarisinde Softplus fonksiyonu kullanıldığında iki farklı ağırlık başlatma yöntemi önerilmiştir. Bu yöntemlerden ilki yalnızca ileri yönde sinyal istatistikini, diğeri ise hem ileri hem de geri yönde sinyal istatistiklerini korumayı amaçlamaktadır. 10 katmanlı ağlarla yapılan deneylerde, iki yönlü istatistik koruma sağlayan başlatma düzeni daha başarılı sonuçlar verirken; 25 katmanlı ağlarda ise yalnızca ileri yönlü korumanın yeterli olmadığı ve iki yönlü korumanın ağı eğitilebilirliğini sağladığı gözlemlenmiştir. Sonuç olarak, bu çalışmada Softplus aktivasyon fonksiyonu kullanılan tam bağlı ileri beslemeli ağlarda, ağı derinleşmesiyle birlikte sinyal istatistiklerini hem ileri hem de geri yönde koruyan ağırlık başlatma stratejisinin, eğitilebilirlik açısından kritik bir öneme sahip olduğu ortaya konmuştur.

Makalenin devamı şu şekilde yapılandırılmıştır: Bölüm 2'de ilgili çalışmalar sunulmakta, Bölüm 3'te Softplus aktivasyon fonksiyonu ve önerilen başlatma yöntemleri detaylı biçimde açıklanmaktadır. Bölüm 4 deneysel çalışmalara, Bölüm 5 ise elde edilen sonuçların tartışmasına ayrılmıştır.

2. İLGİLİ ÇALIŞMALAR

Sinir ağlarının kararlı biçimde iterasyonlar boyunca eğitilebilmesi için en kritik iki parametre aktivasyon fonksiyonu ve başlangıç katman yol ağırlıklarıdır. Özellikle normalizasyon gibi, aktivasyon öncesi ya da sonrası sinyal dağılımını dengeleyen yöntemlerin kullanılmadığı durumlarda, bu iki parametrenin birlikte uygun şekilde ayarlanması daha da önem kazanmaktadır. Bazı aktivasyon fonksiyonları, giriş verisinin katmanlar boyunca 0 ortalama ve 1 standart sapma özelliklerini koruyacak şekilde tasarlanmıştır ve bu amaçla özel yol ağırlık başlatma yöntemleri önerilmiştir. LeCun ve arkadaşları, yeni bir sigmoidal aktivasyon fonksiyonu önermiş ve buna uygun olarak bilinen LeCun başlatma yöntemini geliştirmiştir [24, 25]. Klambauer ve arkadaşları ise, ELU (Exponential Linear Unit) fonksiyonunu modifiye ederek, normalizasyona ihtiyaç duymadan giriş-çıkış dağılımının istatistiksel dengesini sağlayabilen SELU (Scaled Exponential Linear Unit) fonksiyonunu önermiştir [15, 26]. Bu fonksiyonun kendinden normalizasyon özelliğine uygun yol ağırlık başlatması da LeCun yöntemiyle gerçekleştirilmiştir. Zhang ve arkadaşları, SELU'nun bu özelliğinden ilhamla, monoton olmayan yeni bir aktivasyon fonksiyonu önermiştir [27]. Öte yandan Xavier ve arkadaşları, 0 noktasında simetrik, türevlenebilir ve lineer davranış gösteren aktivasyon

fonksiyonları için bir yol ağırlık başlatma formülü geliştirmiştir [16]. He ve arkadaşları ise ReLU ve PReLU gibi fonksiyonlara özgü başlatma düzenleri önermiştir [13].

Bazı çalışmalar işlem kolaylığından ötürü lineer olmayan aktivasyon fonksiyonunun 0 noktasında Taylor açılımını hesaplayarak elde edilen yeni fonksiyonla uyumlu yol ağırlık başlangıç formülü üretmişlerdir. Kumar ve arkadaşları sigmoidal fonksiyonlar [18], Li ve arkadaşları ise parametrik eksponansiyel fonksiyonlar [17] için bu yaklaşımı benimsemiştir.

Literatürdeki bu çalışmaların çoğu, yalnızca ileri yöndeki sinyal istatistiklerinin korunmasını hedeflemiştir. Ancak, geri yayılım algoritmasıyla öğrenme gerçekleştiren sinir ağlarında, geri yöndeki sinyal akışının istatistiksel özelliklerinin de korunması, özellikle derin ağlarda eğitilebilirliğin sürdürülebilmesi açısından kritik öneme sahiptir. Bu çalışmada ise Softplus aktivasyon fonksiyonu için herhangi bir lineerleştirme yapılmaksızın, hem ileri yönde hem de ileri ve geri yönde birlikte sinyal istatistiklerini koruyabilen yeni yol ağırlık başlatma formülleri önerilmiştir. Yapılan deneyler, bu fonksiyon özelinde çift yönlü istatistiksel koruma sağlayan yol ağırlık başlatma düzeninin, gradyan sönümlenmesi veya patlaması gibi sorunları önemli ölçüde azalttığını ve daha derin mimarilerin etkin bir şekilde eğitilebildiğini ortaya koymuştur.

3. MATERYAL VE METOT

Bu bölümde öncelikle Softplus aktivasyon fonksiyonu tanıtılmakta, ardından merkezi limit teoremi temel alınarak bu fonksiyon için yalnızca ileri yönde ve hem ileri hem de geri yönde sinyal istatistiklerini koruyabilen yol ağırlık başlatma formülleri türetilmektedir.

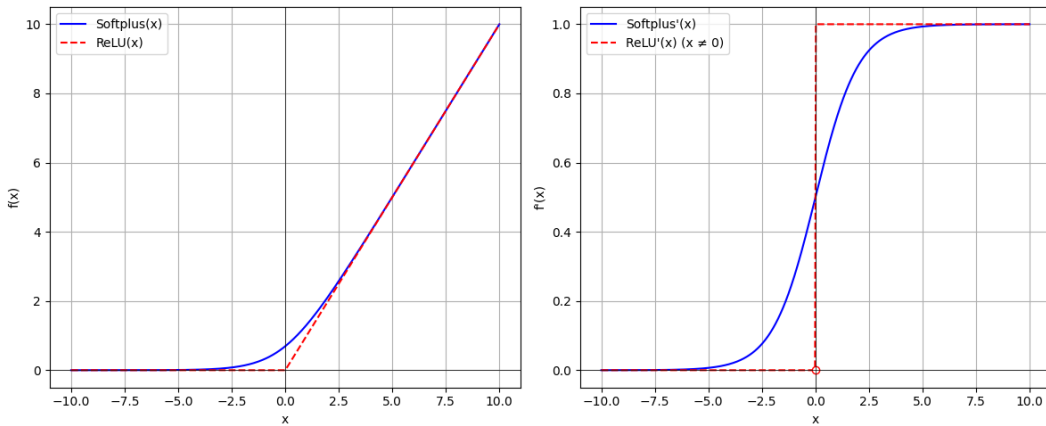
3.1. Softplus Aktivasyon Fonksiyonu

Softplus, reel sayı uzayında tanımlı, sürekli ve türevlenebilir bir aktivasyon fonksiyonudur. Bu fonksiyonun türevi, sıkça kullanılan standart lojistik sigmoid fonksiyonuna karşılık gelmektedir.

$$\text{Softplus}(x) = \ln(e^x + 1) \quad (1)$$

$$\text{Softplus}'(x) = \frac{1}{1 + e^{-x}} \quad (2)$$

Softplus ve literatürde yaygın olarak kullanılan ReLU ($\text{ReLU}(x) = \max(0, x)$) fonksiyonlarına ve bu fonksiyonların türevlerine ait grafikler Şekil 1'de sunulmaktadır. ReLU fonksiyonu, negatif giriş değerleri için sıfır çıkışı üretirken, pozitif değerler için doğrusal (birim) bir fonksiyon gibi davranır. Softplus fonksiyonu da benzer şekilde, giriş değeri büyüdükçe ReLU fonksiyonuna yaklaşan bir çıkış üretmektedir. Türev grafiklerine bakıldığında, Softplus fonksiyonunun türevinin her noktada tanımlı ve süreklilik gösterdiği; buna karşılık ReLU fonksiyonunun türevinin $x=0$ noktasında tanımsız olduğu görülmektedir.



Şekil 1. Softplus ve ReLU fonksiyonları ile türevlerinin grafikleri.

3.2. Softplus Aktivasyon Fonksiyonu İçin Yeni Bir Yol Ağırlık Başlatma Düzeninin Önerilmesi

Bu bölümde, tam bağlı ileri beslemeli sinir ağı mimarisi temel alınarak işlemler gerçekleştirilmiştir. Bu mimarilerde, her katmanda vektör biçimindeki giriş verileri ilgili yol ağırlık matrisleriyle çarpılır. Elde edilen bu çarpım sonucu, yanlılık (bias) vektörü ile toplanarak aktivasyon giriş vektörü elde edilir. Bu aşamaya kadar yapılan işlemler, giriş vektörüne uygulanan lineer bir dönüşüm olarak değerlendirilebilir. Daha sonra bu aktivasyon giriş vektörü, doğrusal olmayan bir aktivasyon fonksiyonundan geçirilerek bir sonraki katmanın giriş vektörü elde edilir. Bu işlem ise doğrusal olmayan dönüşüm olarak adlandırılır. Tüm bu işlemler, nihai çıkış vektörü elde edilene kadar her katmanda tekrarlanır. İlgili işlemler, i . katmanda Softplus aktivasyon fonksiyonu kullanılarak gerçekleştirildiğinde, aşağıdaki matematiksel tanımlamalar ortaya çıkar:

$$y_i = W_i x_i + b_i \quad (3)$$

$$x_{i+1} = \text{Softplus}(y_i) \quad (4)$$

x_i , b_i , y_i ve W_i sırasıyla giriş vektörü, yanlılık vektörü, aktivasyon giriş vektörü ve yol ağırlık matrisi olarak tanımlanır. x_{i+1} ise i . katmanın çıkış vektörü ve aynı zamanda $(i + 1)$. katmanın giriş vektörüdür. n_i , i . katmandaki nöron sayısını; n_{i+1} ise $(i + 1)$. katmandaki nöron sayısını temsil etsin. Bu bağlamda, x_i vektörü $n_i \times 1$ boyutunda, b_i vektörü $n_{i+1} \times 1$ boyutunda, W_i matrisi $n_{i+1} \times n_i$ boyutunda; y_i ve x_{i+1} vektörleri ise $n_{i+1} \times 1$ boyutundadır. Yol ağırlık matrisleri genellikle tekdüze (uniform) ya da normal dağılımdan örneklenmekte olup sıfır ortalamaya sahip olacak şekilde başlatılır. Literatürde, ileri yönde sinyal akışını analiz edebilmek amacıyla bazı önvarsayımlar altında istatistiksel incelemeler gerçekleştirilmiştir [13, 16]. Bu önvarsayımlara göre, x_i vektörünün ve W_i matrisinin elemanları kendi içlerinde bağımsız ve özdeş dağılımdan gelmekte ve ayrıca birbirlerinden bağımsız kabul edilmektedir. Yanlılık vektörü b_i ise başlangıçta sıfır vektörü olarak alınmaktadır. Bu koşullar altında, i ilk katmandan farklı olmak üzere y_i vektörünün varyansı aşağıda analiz edilmiştir:

$$\text{Var}[y_i] = n_i \text{Var}[w_i x_i] \quad (5)$$

x_i , w_i , y_i sırasıyla x_i , W_i ve y_i değerlerinin rassal değişkenleri olsun.

$$\text{Var}[y_i] = n_i (E[w_i^2] E[x_i^2] - E[w_i]^2 E[x_i]^2) \quad (6)$$

$E[w_i] = 0$ ve pozitif çıkış üreten Softplus fonksiyonundan ötürü $E[x_i] \neq 0$ olur. Bu bilgiler doğrultusunda, Denklem 6 şu şekilde düzenlenir:

$$\text{Var}[y_i] = n_i \text{Var}[w_i] E[x_i^2] \quad (7)$$

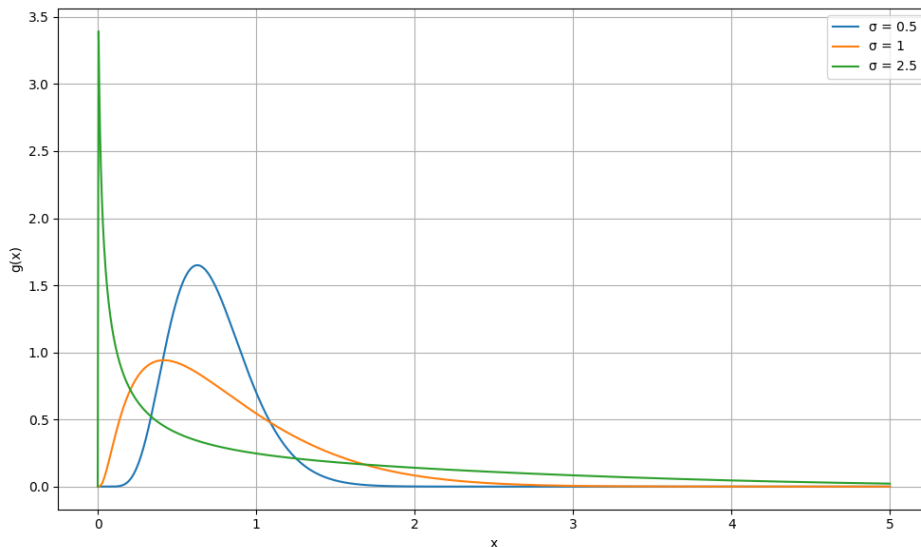
W_i matrisinin sıfır ortalama olması nedeniyle y_i vektörü de sıfır ortalama ve simetrik bir dağılıma sahiptir. Bunun yanında y_i vektörü, n_i değerinin artmasıyla merkezi limit teoremine göre normal dağılım özelliği gösterir [18, 26]. Gradyan patlaması veya sönümlenmesi problemlerinden kaçınmak için, katmanlar boyunca giriş sinyallerinin ve aktivasyon giriş sinyallerinin kendi içerisinde benzer istatistiklere sahip olması kritik öneme sahiptir. Bu nedenle, eğitim başlangıcında L katman boyunca aktivasyon giriş sinyallerinin varyanslarının sabit kalması sağlanmalıdır.

$$\text{Var}[y_1] = \text{Var}[y_2] = \dots = \text{Var}[y_L] = \sigma^2 \quad (8)$$

Ardışık iki aktivasyon giriş sinyali arasındaki ilişkiyi inceleyebilmek için, aralarında bulunan Softplus dönüşümünün analiz edilmesi gerekmektedir. Ortalama değeri sıfır ve standart sapması σ olan normal dağılıma sahip bir rassal değişkene Softplus dönüşümü uygulandığında, elde edilen dağılımın olasılık yoğunluk fonksiyonu aşağıdaki gibidir.

$$g(x) = \frac{1}{\sqrt{2\pi}\sigma} \frac{1}{(1 - e^{-x})} e^{\frac{-(\ln(e^x - 1))^2}{2\sigma^2}}, \quad \infty > x > 0 \quad (9)$$

Ortalaması sıfır ve farklı standart sapma değerlerine sahip normal dağılmış rassal değişkenlere Softplus fonksiyonu uygulandığında elde edilen olasılık yoğunluk fonksiyonları, Şekil 2’de yer gösterilmiştir.



Şekil 2. Sıfır ortalama ve farklı σ değerli normal dağılıma sahip rassal değişkenlerin Softplus dönüşümü sonrasındaki olasılık yoğunluk fonksiyonları.

Şekil 2 incelendiğinde, σ değerinin artmasıyla olasılık yoğunluk fonksiyonlarının sağa çarpıklığının arttığı gözlemlenmektedir. Denklem 9'daki olasılık yoğunluk fonksiyonunun ikinci momenti olan $E[x_i^2]$ değerini hesaplayarak, $Var[y_i]$ ile $Var[y_{i-1}]$ arasındaki ilişki bulunabilir. Ancak $E[x_i^2]$ için analitik bir çözüm mevcut olmadığından, yaklaşık sonuçlar için nümerik yöntemlere başvurulması gerekmektedir. Birçok çalışmada aktivasyon giriş vektörü standart normal dağılım özelliğine sahip kabul edilip işlemler gerçekleştirildiğinden, bu çalışmada da ileri yönde sinyal istatistiği koruma amacıyla aktivasyon giriş bilgisinin 0 ortalama ve 1 standart sapmalı olması üzerinden formül üretimi gerçekleştirilecektir [26, 27]. $\sigma = 1$ için yapılan nümerik çözüm sonucunda $E[x_i^2] \approx 0.9212$ değeri elde edilmiştir. Bu değer ve $Var[y_i] = 1$ ifadesi Denklem 7'ye yerleştirildiğinde, $Var[w_i]$ değeri hesaplanabilir.

$$Var[w_i] \approx \frac{1.0855}{n_i} \quad (10)$$

Yukarıdaki denklem, ilgili katman yol ağırlık varyansının katman nöron sayısı ile ters orantılı olduğunu göstermektedir. Bu başlatma düzeni ile ileri yönde aktivasyon giriş sinyallerinin benzer ve yaklaşık 1 varyanslı olması sağlanır. Ayrıca, farklı σ değerleri için farklı yol ağırlık başlatma katsayılarının elde edilebileceği açıktır. İleri yönde sinyal iletiminden sonra, geri yayılım algoritması kullanılarak sinir ağınn parametreleri uygun şekilde güncellenir. Öğrenmenin gerçekleşebilmesi için, yol ağırlıklarının tüm katmanlarda iterasyonlar boyunca “canlı” kalması, yani güncellenebilirliğini koruması gereklidir. Bu nedenle, ileri yönde sinyal istatistiğinin korunmasının geri yöndeki sinyal istatistiklerine olan etkisinin incelenmesi önem taşır. Burada amaç, yol ağırlıklarının gradyan istatistiklerinin katmanlar boyunca benzerliğini sürdürebilecek bir başlangıç formülü geliştirmektir. İleri yöndeki gibi, geri yönde de sinyal kararlılığının sağlanması bazı önvaryanslımlar altında analiz edilmiştir [13, 16].

Δx_i , Δy_i sırasıyla $\frac{\partial Loss}{\partial x_i}$ ve $\frac{\partial Loss}{\partial y_i}$ olarak tanımlansın. Burada, Δx_i , i . katmanın giriş gradyan vektörünü temsil eder ve boyutu $n_i \times 1$ 'dir. Δy_i ise i . katman aktivasyon giriş gradyan vektörüdür ve boyutu $n_{i+1} \times 1$ 'dir. Bu iki vektör arasındaki ilişki aşağıdaki şekilde ifade edilir:

$$\Delta x_i = \hat{W}_i \Delta y_i \quad (11)$$

$\hat{W}_i$, $n_i \times n_{i+1}$ boyutunda bir matristir ve bu matris, ileri yönde sinyal istatistiklerini koruma amacıyla tanımlanan W_i matrisinin yeniden düzenlenmiş biçimidir. Dolayısıyla, $\hat{W}_i$ ve W_i aynı istatistiksel özelliklere sahiptir.

$\hat{W}_i$ matrisinin ve Δy_i vektörünün elemanları kendi içerisinde bağımsız ve özdeş dağılımdan gelsin. Ayrıca $\hat{W}_i$ matrisi ve Δy_i vektörü birbirinden bağımsız olsun. Benzer şekilde, $(i + 1)$. katman giriş gradyan vektörü Δx_{i+1} ile i . katman aktivasyon giriş vektörü y_i birbirinden bağımsız olsun. Δy_i aşağıdaki şekilde hesaplanır:

$$\Delta y_i = \text{Softplus}'(y_i) \Delta x_{i+1} \quad (12)$$

$\hat{W}_i$ matrisinin sıfır ortalamalı bir dağılımdan geliyor olması nedeniyle, $E[\Delta x_i] = 0$ olur. Bu durum her bir katman için geçerlidir. Dolayısıyla, $E[\Delta x_{i+1}] = 0$ ve buna bağlı olarak $E[\Delta y_i] = 0$ olduğu sonucuna varılır. Bu bilgiler doğrultusunda, Δx_i vektörünün varyansı aşağıdaki gibi incelenebilir:

$$Var[\Delta x_i] = n_{i+1} Var[\hat{W}_i] Var[\Delta y_i] \quad (13)$$

Δx_i , $\hat{W}_i$, Δy_i sırasıyla Δx_i , $\hat{W}_i$, Δy_i değerlerinin rassal değişkenleri olsun.

Geri yönde gradyan yok olması veya patlaması problemlerinin önüne geçebilmek için, ardışık katman giriş gradyan vektörleri Δx_{i+1} ile Δx_i arasında istatistiksel bir ilişkinin kurulması gerekmektedir. Bu amaçla, öncelikle Δy_i ile Δx_{i+1} arasındaki istatistiksel ilişkinin incelenmesi gereklidir.

$$Var[\Delta y_i] = Var[\text{Softplus}'(y_i) \Delta x_{i+1}] \quad (14)$$

$$Var[\Delta y_i] = E[\text{Softplus}'(y_i)^2] E[\Delta x_{i+1}^2] - E[\text{Softplus}'(y_i)]^2 E[\Delta x_{i+1}]^2 \quad (15)$$

$E[\Delta x_{i+1}] = 0$ 'dır. Softplus fonksiyonunun türevi, standart lojistik sigmoid fonksiyondur. Bu fonksiyonun yalnızca pozitif çıkış üretmesi nedeniyle $E[\text{Softplus}'(y_i)] \neq 0$ olur. Bu bilgiler doğrultusunda ifade aşağıdaki şekilde düzenlenebilir:

$$Var[\Delta y_i] = E[\text{Softplus}'(y_i)^2] Var[\Delta x_{i+1}] \quad (16)$$

Denklem 16'daki $Var[\Delta y_i]$ ifadesi, Denklem 13'te yerine yazılıp gerekli düzenlemeler yapılırsa aşağıdaki sonuç elde edilir.

$$\frac{Var[\Delta x_i]}{Var[\Delta x_{i+1}]} = n_{i+1} Var[\hat{W}_i] E[\text{Softplus}'(y_i)^2] \quad (17)$$

$\hat{W}_i$ ile w_i aynı istatistiksel özelliklere sahiptir. Denklem 7'deki $Var[w_i]$ ve $Var[y_i] = \sigma^2$ değeri, Denklem 17'de yerine yazıldığında, ardışık iki gradyan giriş vektörünün varyans oranı elde edilir.

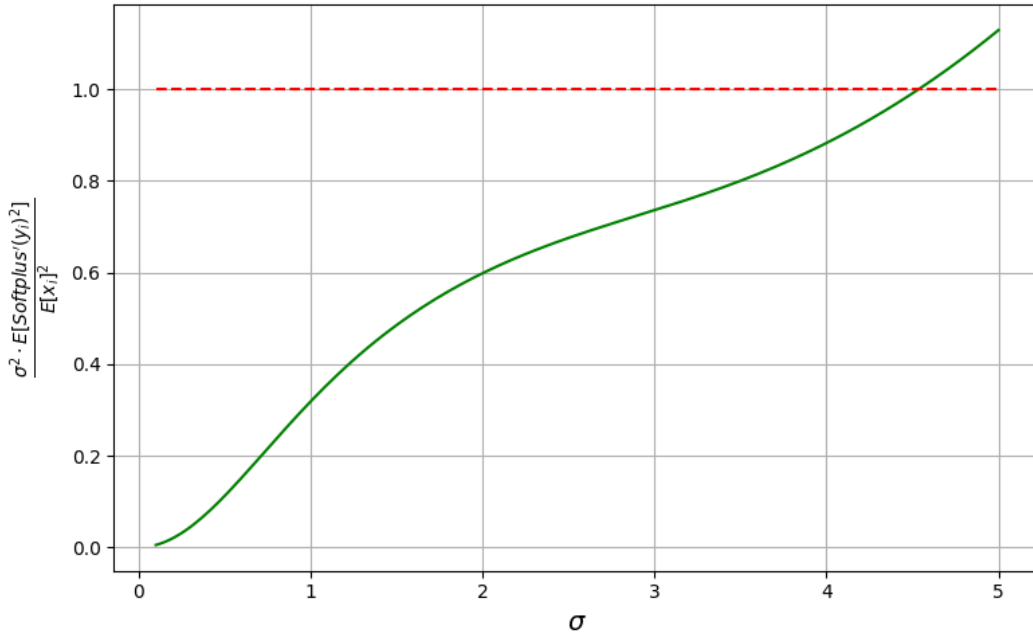
$$\frac{Var[\Delta x_i]}{Var[\Delta x_{i+1}]} = \frac{\sigma^2 n_{i+1} E[Softplus'(y_i)^2]}{n_i E[x_i^2]} \quad (18)$$

$E[Softplus'(y_i)^2]$ değerinin hesaplanabilmesi için, normal dağılım özelliği gösteren y_i vektörünün, standart lojistik sigmoid fonksiyonu ile dönüşümünün incelenmesi gerekmektedir. Ortalama değeri 0 ve standart sapması σ olan normal dağılıma sahip bir rassal değişkene, standart lojistik sigmoid fonksiyon dönüşümü uygulandığında, elde edilen dağılımın olasılık yoğunluk fonksiyonu aşağıdaki gibidir. [28].

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \frac{1}{x(1-x)} e^{-\frac{(\ln(\frac{x}{1-x}))^2}{2\sigma^2}}, \quad 1 > x > 0 \quad (19)$$

Softplus fonksiyonunun ileri ve geri yöndeki dinamik etkilerini inceleyebilmek amacıyla, tüm katmanlardaki nöron sayıları eşit kabul edilerek, σ değerine bağlı olarak $\frac{\sigma^2 E[Softplus'(y_i)^2]}{E[x_i^2]}$ oranı hesaplanmalıdır. Buradaki $E[Softplus'(y_i)^2]$, Denklem 19'da verilen olasılık yoğunluk fonksiyonunun ikinci momentidir. Bu değer analitik olarak hesaplanamaz; bu nedenle nümerik yöntemlerle yaklaşık çözümler elde edilir. Söz konusu oranı yaklaşık 1 yapan σ değeri bulunduğunda, hem ileri hem de geri yönde sinyal istatistiği korunmuş olur. Şekil 3 incelendiğinde, bu oranın 1 olduğu durum $\sigma \approx 4.536$ olarak gözlemlenmektedir. Böylece, hem ileri hem de geri yönde istatistiksel bilgiyi koruyabilen σ değeri elde edilmiş olur. Bu durumda, $\sigma \approx 4.536$ iken $E[x_i^2] \approx 8.602$ olarak bulunur. Bu iki değer Denklem 7'de yerine konulduğunda, ileri ve geri yönde sinyal istatistiğini aynı anda koruyabilen yol ağırlık başlangıç formülü elde edilir.

$$Var[w_i] \approx \frac{2.3888}{n_i} \quad (20)$$



Şekil 3. Softplus fonksiyonu için ileri ve geri yönde istatistiksel kararlılığı sağlayan kritik σ değerinin belirlenmesi.

$\sigma \approx 4.536$, Softplus fonksiyonu için hem ileri hem de geri yönde sinyal istatistiğini koruyabilen kritik değerdir. Bu değer, başlangıç aşamasında aktivasyon giriş vektörlerinin standart sapması olarak ayarlanmalıdır. Eğer σ , bu kritik değerden küçük seçilirse, ilk katmanlardaki giriş gradyanlarının büyüklüğü, son katmanlara göre anlamlı şekilde azalır. Bu durum, ilk katmanlardaki yol ağırlık gradyan varyanslarının, son katmanlardakilere göre çok daha küçük olmasına neden olur ve katman sayısı arttıkça bu fark üstel olarak büyür. Sonuç olarak, ağırlık başlangıç katmanları yeterince öğrenemez ve ağırlık eğitimi başarısız olur. Hem ileri hem de geri yönde istatistik korumayı sağlayan başlatma yaklaşımının avantajı, yalnızca ileri yönde istatistik koruyabilen yöntemle göre bir sonraki bölümde deneysel olarak incelenmiştir.

4. DENEYSEL SONUÇLAR

Deneyler, Intel(R) Core(TM) i9-10850K 3.60 GHz işlemci, 128 GB RAM ve Nvidia RTX 3090 ekran kartına sahip bir bilgisayarda, Python programlama dili ve TensorFlow kütüphanesi kullanılarak gerçekleştirilmiştir. Mimari olarak tam bağlı ileri beslemeli sinir ağları, görüntü sınıflandırma problemleri için tercih edilmiştir. Deneylerde, CIFAR-10 ve CIFAR-100 veri kümeleri kullanılmıştır [29]. CIFAR-10 ve CIFAR-100 veri kümeleri

sırasıyla 10 ve 100 farklı sınıfa sahip olup, her biri 50.000 eğitim ve 10.000 test örneğinden oluşmaktadır. Bu veri kümelerindeki tüm görüntüler renkli (RGB) ve $32 \times 32 \times 3$ boyutlarındadır. Veri kümeleri, ön işlem adımı olarak kendi içerisindeki eğitim verilerinin kanal ortalama değerleriyle normalize edilerek yaklaşık sıfır ortalamalı hale getirilmiştir. Ayrıca, ileri yönde sinyal istatistiklerini koruyabilen başlatma düzeniyle uyumlu olması için, giriş verilerinin standart sapması yaklaşık 1 olacak şekilde ölçeklendirilmiştir. Buna karşın, hem ileri hem de geri yönde sinyal istatistiklerini koruyabilmek amacıyla, giriş verilerinin standart sapması yaklaşık 4.536 olacak şekilde ölçeklendirilmiştir. Bu iki farklı başlatma yöntemi için giriş verilerinin standart sapma değerleri başlangıçta bu şekilde ayarlanmıştır. Deneysel çalışmalar, gizli katman sayısı (10 ve 25), yığın boyutu (64, 128, 256, 512), öğrenme oranı (10^{-1} , 10^{-2} , 10^{-3} , 10^{-4}) ve yol ağırlıklarının başlatılma yöntemi (Denklem 10 ve Denklem 20) gibi parametre kombinasyonları için, her biri üçer tekrar olacak şekilde gerçekleştirilmiştir. Gizli katmanlardaki nöron sayısı 512 olarak belirlenmiş ve optimizasyon için stokastik gradyan iniş (SGD) algoritması 50 epok (epoch) boyunca uygulanmıştır.

Bu çalışmada, katman sayısının artması durumunda eğitimin gerçekleştirilebilme durumu incelenmiştir. Bu nedenle, başarımlar eğitim veri kümesi üzerinden kıyaslanmıştır. CIFAR-10 veri kümesi için, yalnızca ileri yönde sinyal istatistikini koruyan ve hem ileri hem de geri yönde koruma sağlayan yol ağırlık başlatma düzenlerinin ortalama eğitim başarımları Tablo 1’de sunulmaktadır.

Tablo 1. CIFAR-10 veri kümesi üzerinde yalnızca ileri yönde ve hem ileri hem de geri yönde sinyal istatistikini koruyabilen başlatma düzenli mimarilerin başarımlar sonuçları.

Başlatma Düzeni	Yığın Sayısı	Eğitim Başarısı (Ortalama) $\pm$ Standart Sapma	
		Gizli Katman Sayısı	
		10	25
İleri	64	0.9693 $\pm$ 0.0022	-
İleri-Geri	64	0.9992$\pm$0.0002	0.9388$\pm$0.0036
İleri	128	0.9182 $\pm$ 0.0007	-
İleri-Geri	128	0.9869$\pm$0.0092	0.8547$\pm$0.0257
İleri	256	0.7552 $\pm$ 0.0035	-
İleri-Geri	256	0.9800$\pm$0.0141	0.7675$\pm$0.0082
İleri	512	0.6070 $\pm$ 0.0029	-
İleri-Geri	512	0.8639$\pm$0.0542	0.5547$\pm$0.0213

Bu veri kümesi için 10 katmanlı mimarinin epoklar boyunca ortalama eğitim başarımları Şekil 4’te sunulmaktadır. Tüm durumlarda hem ileri hem de geri yönde sinyal istatistiklerini koruyabilen başlatma düzeni, yalnızca ileri yönde sinyal istatistiklerini koruyabilen başlatma düzenine göre daha yüksek başarımlar sağlamıştır. Öte yandan, 25 katmanlı mimarinin başarımlar grafikleri, yalnızca ileri yönde sinyal istatistiklerini koruyabilen başlatma düzeniyle hiçbir durumda eğitim yapılamadığından dolayı burada gösterilmemiştir.

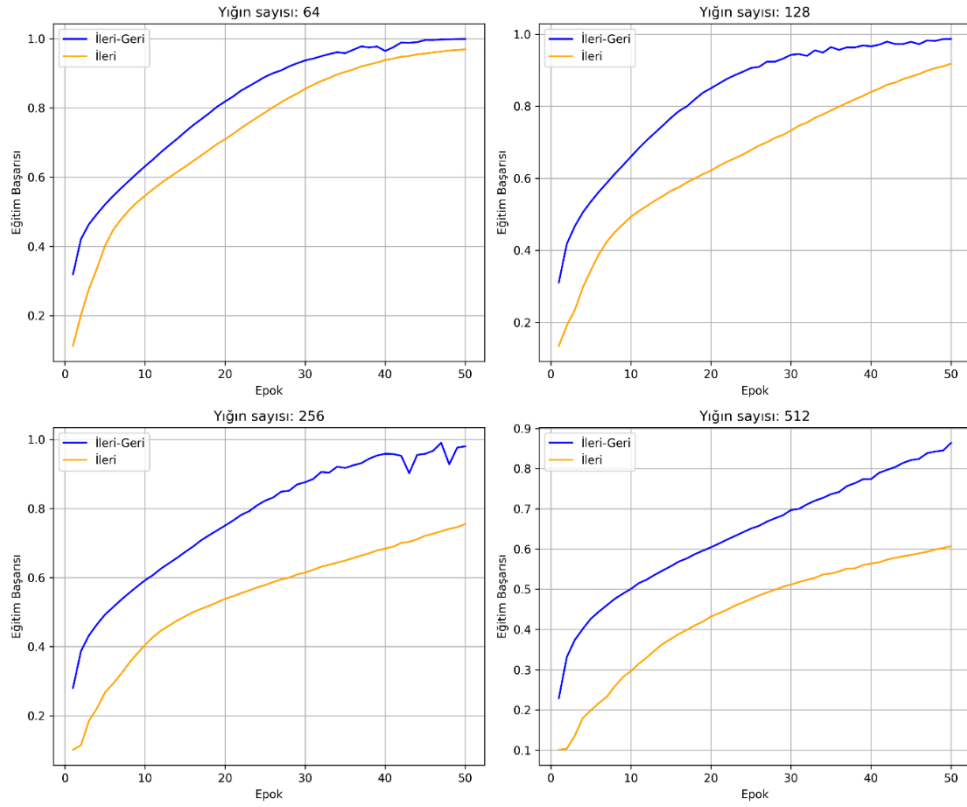
CIFAR-100 veri kümesi için, yalnızca ileri yönde ve hem ileri hem de geri yönde sinyal istatistiklerini korumayı amaçlayan yol ağırlık başlatma düzenli mimarilerin ortalama eğitim başarımları Tablo 2’de sunulmaktadır.

Tablo 2. CIFAR-100 veri kümesi üzerinde yalnızca ileri yönde ve hem ileri hem de geri yönde sinyal istatistikini koruyabilen başlatma düzenli mimarilerin başarımlar sonuçları.

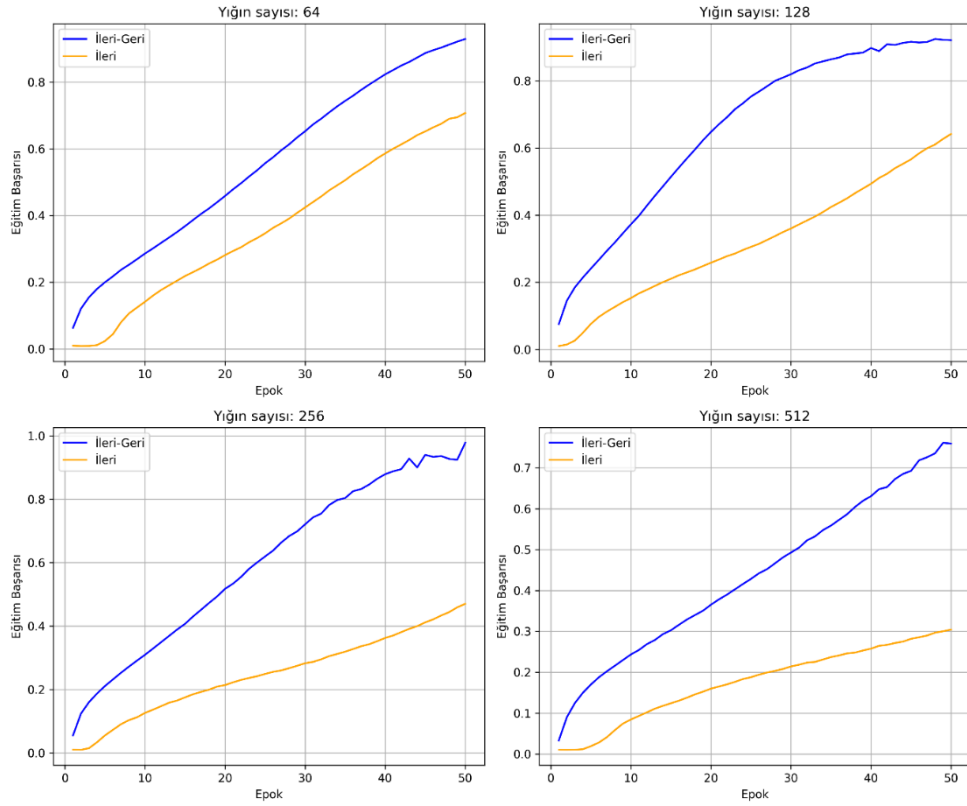
Başlatma Düzeni	Yığın Sayısı	Eğitim Başarısı (Ortalama) $\pm$ Standart Sapma	
		Gizli Katman Sayısı	
		10	25
İleri	64	0.7073 $\pm$ 0.0103	-
İleri-Geri	64	0.9293$\pm$0.0037	0.6996$\pm$0.0037
İleri	128	0.6417 $\pm$ 0.0036	-
İleri-Geri	128	0.9210$\pm$0.0091	0.5740$\pm$0.0145
İleri	256	0.4703 $\pm$ 0.0123	-
İleri-Geri	256	0.9782$\pm$0.0063	0.4231$\pm$0.0086
İleri	512	0.3044 $\pm$ 0.0078	-
İleri-Geri	512	0.7593$\pm$0.0203	0.2727$\pm$0.0092

CIFAR-100 veri kümesi için, 10 katmanlı mimarinin epoklar boyunca ortalama eğitim başarımları Şekil 5’te sunulmaktadır. CIFAR-10 veri kümesinde olduğu gibi, tüm durumlarda hem ileri hem de geri yönde sinyal istatistiklerini koruyabilen başlatma düzeni, yalnızca ileri yönde istatistik koruyabilen başlatma düzenine kıyasla daha başarılı olmuştur. Ayrıca, 25 katmanlı mimarinin başarımlar grafikleri, yalnızca ileri yönde istatistik koruyabilen başlatma düzeniyle hiçbir durumda eğitim gerçekleştirilemediği için burada sunulmamıştır.

Başarımların, giriş verisinin istatistiksel özelliklerinden bağımsız olup olmadığını incelemek amacıyla; yol ağırlıklarının, Denklem 20’de verilen başlatma yöntemiyle belirlendiği ve giriş verisinin standart sapmasının 1 olarak ayarlandığı deneyler, her iki veri kümesi için de 25 katmanlı bir mimari üzerinde yürütülmüştür. Bu hiperparametre kombinasyonunda, tüm durumlarda ağırlar eğitilememiş ve istenen başarımlar elde edilememiştir.



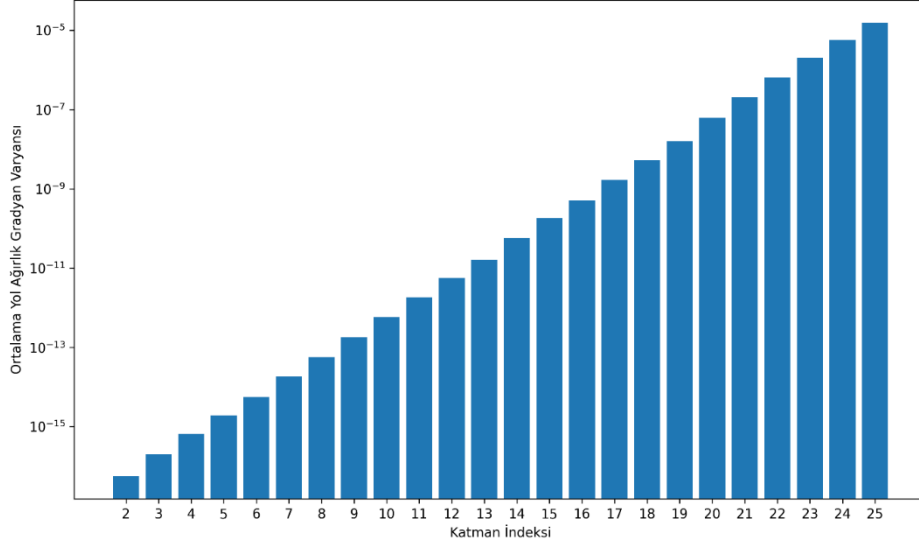
Şekil 4. CIFAR-10 veri kümesi üzerinde, farklı yığın boyutları için yalnızca ileri yönde ve hem ileri hem de geri yönde istatistik koruyabilen başlatma düzenlerinin eğitim kümesi üzerinde başarımları.



Şekil 5. CIFAR-10 veri kümesi üzerinde, farklı yığın boyutları için yalnızca ileri yönde ve hem ileri hem de geri yönde istatistik koruyabilen başlatma düzenlerinin eğitim kümesi üzerinde başarımları.

Bu sonuçlar, katman sayısının artmasıyla derin ağların eğitilebilirliğinin zorlaştığını ve aktivasyon fonksiyonunun dinamiklerini gözetken yol ağırlık başlatma düzenlerinin önemini vurgulamaktadır.

Yalnızca ileri yönde istatistik korumanın, derin ağların eğitimi için neden yeterli olmadığını göstermek amacıyla; her katmanda 512 nöron bulunan, 25 gizli katmanlı ve yığın boyutu 64 olan bir mimarinin, CIFAR-100 veri kümesi üzerindeki 1 epok boyunca katmanlardaki ortalama yol ağırlık gradyan varyansları incelenmiştir. Şekil 6'da bu durum sunulmaktadır. y eksenini logaritmik ölçekte çizdirilmiş olup, ilk katmanlara doğru ortalama yol ağırlık gradyan varyanslarının hızla çok küçük değerlere düştüğü görülmektedir. Bu nedenle, başlangıç aşamasında bu mimari, ilgili hiperparametrelerle eğitilememiştir.



Şekil 6. Katmanlara göre ortalama yol ağırlık gradyan varyansları.

5. SONUÇ

Derin mimarilerde, sığ mimariyle kıyasla katman sayısı ve dolayısıyla parametre sayısı artırılarak daha karmaşık problemlerin çözümüne olanak sağlanır. Ancak, gradyan bilgisiyle parametrelerini güncelleyen bu ağlarda, katman sayısının artması durumunda gradyan sönümlenmesi veya patlaması gibi problemlerle karşılaşılabilir. Bu durum, öğrenmenin durmasına ya da başlangıç aşamasında ağına hiç eğitilememesine neden olur. Bu tür problemlerin önüne geçebilmek için, ağ parametrelerinin başlangıçta dikkatli bir şekilde ayarlanması büyük önem taşır. Aktivasyon fonksiyonu ve yol ağırlık başlatma değerleri, bu parametrelerin en önemlileri arasında yer almakta olup genellikle birlikte ayarlanmaları göz ardı edilmektedir. Bu çalışmada, Softplus aktivasyon fonksiyonu için merkezi limit teoremine dayalı iki farklı yol ağırlık başlatma düzeni geliştirilmiştir. Birincisi, literatürde sıkça kullanılan ve yalnızca ileri yönde istatistik değerlerini koruyabilen başlatma düzenidir. İkincisi ise hem ileri hem de geri yönde istatistik değerlerini koruyabilen başlatma düzenidir. Deneysel çalışmalar, 10 ve 25 gizli katmanlı mimariler için farklı yığın boyutları kullanılarak CIFAR-10 ve CIFAR-100 veri kümeleri üzerinde gerçekleştirilmiştir. Her iki veri kümesinde de, 10 gizli katmanlı mimarilerde hem ileri hem de geri yönde istatistik koruyabilen başlatma düzeniyle daha başarılı sonuçlar elde edilmiştir. Bunun yanında, 25 gizli katmanlı mimarilerde ise yalnızca iki yönde istatistik koruyabilen başlatma düzeniyle ağlar eğitilebilmiştir. Bu çalışma, ağına derinleşmesiyle birlikte Softplus aktivasyon fonksiyonunun dinamiğini dikkate alan ve hem ileri hem de geri yönde istatistik koruyabilen başlatma düzeninin önemini ortaya koymuştur. Gelecek çalışmalarda, Softplus aktivasyon fonksiyonunun kullanıldığı ağlarda, hem ileri hem de geri yönde istatistiksel özellikleri koruyabilen önerilen başlatma düzeninin, Yığın Normalizasyon gibi yaygın normalizasyon yöntemleriyle performans ve eğitim dinamiği bakımından karşılaştırılması planlanmaktadır. Böylece, derin ağlarda aktivasyon dağılımının denetlenmesine yönelik yöntemler arasındaki göreceli katkılar daha güçlü biçimde değerlendirilebilecektir.

Yazar Katkıları

Yazarlar makaleye eşit oranda katkı sağlamışlardır.

Çıkar Çatışması

Makale yazarları aralarında herhangi bir çıkar çatışması olmadığını beyan ederler

KAYNAKÇA

- [1] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," Nat. 2015 5217553, vol. 521, no. 7553, pp. 436-444, May 2015, doi: 10.1038/nature14539.
- [2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in Advances in Neural Information Processing Systems, F. Pereira, C. J. Burges, L. Bottou, and K. Q. Weinberger, Eds.,

- Curran Associates, Inc., 2012. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf
- [3] J. Kaur and W. Singh, "A systematic review of object detection from images using deep learning," *Multimed. Tools Appl.*, vol. 83, no. 4, pp. 12253–12338, 2024.
- [4] D. Guo et al., "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," *arXiv Prepr. arXiv2501.12948*, 2025.
- [5] Y. Bengio, "Learning Deep Architectures for AI," *Found. Trends® Mach. Learn.*, vol. 2, no. 1, pp. 1–127, 2009.
- [6] F. Chollet, *Deep Learning with Python (Second Edition)*. New York: Manning Publications, 2021.
- [7] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994.
- [8] G. Philipp, D. Song, and J. G. Carbonell, "The exploding gradient problem demystified-definition, prevalence, impact, origin, tradeoffs, and solutions," *arXiv Prepr. arXiv1712.05577v4*, 2018.
- [9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2016, pp. 770–778.
- [10] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
- [11] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, in *ICML'15. JMLR.org*, 2015, pp. 448–456.
- [12] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv Prepr. arXiv1607.06450*, 2016.
- [13] K. He, X. Zhang, S. Ren, and J. Sun, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1026–1034.
- [14] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, in *ICML'10. Madison, WI, USA: Omnipress*, 2010, pp. 807–814.
- [15] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and Accurate Deep Network Learning by Exponential Linear Units(ELUs)," in *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2–4, 2016, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2016. [Online]. Available: <http://arxiv.org/abs/1511.07289>
- [16] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, Y. W. Teh and M. Titterton, Eds., in *Proceedings of Machine Learning Research*, vol. 9. Chia Laguna Resort, Sardinia, Italy: PMLR, Jan. 2010, pp. 249–256. [Online]. Available: <https://proceedings.mlr.press/v9/glorot10a.html>
- [17] Y. Li, C. Fan, Y. Li, Q. Wu, and Y. Ming, "Improving deep neural network with Multiple Parametric Exponential Linear Units," *Neurocomputing*, vol. 301, 2018.
- [18] S. K. Kumar, "On weight initialization in deep neural networks," *arXiv Prepr. arXiv1704.08863v2*, 2017.
- [19] S. R. Dubey, S. K. Singh, and B. B. Chaudhuri, "Activation functions in deep learning: A comprehensive survey and benchmark," *Neurocomputing*, vol. 503, pp. 92–108, Sep. 2022.
- [20] M. Leshno, V. Y. Lin, A. Pinkus, and S. Schocken, "Multilayer feedforward networks with a nonpolynomial activation function can approximate any function," *Neural Networks*, vol. 6, no. 6, pp. 861–867, 1993.
- [21] P. F. Verhulst, "Recherches mathématiques sur la loi d'accroissement de la population," *Mémoires l'Académie R. des Sci. B.-lett. Bruxelles*, vol. 18, pp. 1–41, 1845.
- [22] F.-C. Chen, "Back-propagation neural networks for nonlinear self-tuning adaptive control," *IEEE Control Syst. Mag.*, vol. 10, no. 3, pp. 44–48, 1990.
- [23] C. Dugas, Y. Bengio, F. Bélisle, C. Nadeau, and R. Garcia, "Incorporating Second-Order Functional Knowledge for Better Option Pricing," in *Advances in Neural Information Processing Systems*, T. Leen, T. Dietterich, and V. Tresp, Eds., MIT Press, 2000. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2000/file/44968aece94f667e4095002d140b5896-Paper.pdf
- [24] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- [25] L. and O. G. B. and M. K.-R. LeCun Yann A. and Bottou, "Efficient BackProp," in *Neural Networks: Tricks of the Trade: Second Edition*, G. B. and M. K.-R. Montavon Grégoire and Orr, Ed., Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 9–48.
- [26] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, "Self-normalizing neural networks," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, in *NIPS'17. Red Hook, NY, USA: Curran Associates Inc.*, 2017, pp. 972–981.
- [27] G. Zhang and H. Li, "Effectiveness of scaled exponentially-regularized linear units (SERLUs)," *arXiv Prepr. arXiv1807.10117*, Jun. 2018.
- [28] N. L. Johnson, "Systems of Frequency Curves Generated by Methods of Translation," *Biometrika*, vol. 36, no. 1/2, p. 149, Jun. 1949.
- [29] A. Krizhevsky, "Learning Multiple Layers of Features from Tiny Images," 2009.