Int. J. Pure Appl. Sci. 8(1);53-60 (2022)

IJPAS
ijpas@munzur.edu.tr
ISSN: 2149-0910

Research article/Araştırma makalesi
DOI:10.29132/ijpas.995579

Gömülü Sistemler İçin Performansı Arttırılmış HBONet CNN Yaklaşımı

Gürkan Doğan[1]*, Burhan Ergen[2]
[1]*Munzur Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Tunceli, Türkiye*
[2]*Fırat Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Elazığ, Türkiye*
**gurkandogan@munzur.edu.tr , bergen@firat.edu.tr*
Makale gönderme tarihi:16.09.2021, Makale kabul tarihi: 26.03.2022

Öz

Son yıllarda, evrişimli sinir ağlarının (CNN) kullanım alanları dikkate değer bir şekilde artmıştır. İş istasyonlarından gömülü cihazlara varıncaya kadar birçok platformda yaygın olarak kullanılmaktadır. Bununla birlikte, her CNN modeli farklı miktarda hafıza, işlemci, depolama birimi kullanmaktadır ve nesne tanımada farklı doğruluk oranlarına sahiptir. Gömülü sistemlerde kullanılacak CNN'lerin daha az maliyetli olması, daha az kaynak tüketmesi ve daha fazla doğruluk oranını başarması gibi bazı zorlukları vardır. Bu zorlukların en iyi üstesinden gelen CNN modellerinden biri de HBONet modelidir. Ancak, bu model gömülü sistemlerde yeterince iyi performans sağlamamaktadır. Bu çalışmada, gömülü sistemler için kullanılan HBONet modelinin kaynak tüketimi ve doğruluk gibi performans metriklerinin daha da iyileştirilmesi amaçlanmıştır. Bu amaçla, HBONet modelini temel alan bir model olan A-HBONet modeli önerilmiştir. CIFAR-10 veri seti kullanılarak gerçekleştirilen deneyler sonucunda, önerilen modelin doğruluğu HBONet modeline göre %3 arttırılırken hafıza ve depolama birimi kullanımı da yaklaşık olarak %80 oranında azaltılmıştır. Bu sonuçlar, önerilen modelin gömülü cihazlarda daha etkin ve verimli çalıştığı göstermektedir.

Anahtar Kelimeler: Gömülü sistemler, evrişimli sinir ağları, A-HBONet, CIFAR-10

Performance Enhanced HBONet CNN Approach for Embedded Systems

Abstract

In recent years, the usage areas of convolutional neural networks (CNN) have increased remarkably. It is widely used on many platforms, from workstations to embedded devices. However, each CNN model uses a different amount of memory, processor, storage and has different object recognition accuracy rates. CNNs to be used in embedded systems have some difficulties such as being less costly, consuming less resources and achieving higher accuracy. One of the CNN models that best overcomes these difficulties is the HBONet model. However, this model does not perform well enough in embedded systems. In this study, it is aimed to increase the performance of the HBONet model for embedded systems. For this purpose, the A-HBONet model, which is based on the HBONet model, is proposed. As a result of the experiments performed, the accuracy of the proposed model was increased by 3% compared to the HBONet model, while the memory and storage unit usage was reduced by approximately 80%. These results show that the proposed model works more effectively and efficiently in embedded devices.

Keywords: Embedded systems, convolutional neural networks, A-HBONet, CIFAR-10

GİRİŞ

Son yıllarda, nesne tanıma, tespit etme ve semantik bölütleme uygulamaları için derin öğrenme teknolojileri yaygın olarak kullanılmaktadır. Özellikle, Evrişimli Sinir Ağları (CNN), görsel tanıma görevi için akıllı kameralar (Rinner & Wolf, 2008), robotik navigasyon ve sanal gerçeklik (Weibin Liu, Chao Zhang, & Baozong Yuan, 2002) vb. gibi çeşitli bilgisayar görmesi uygulamalarında oldukça başarılı bir şekilde kullanılmaktadır. Evrişimli Sinir Ağları, etkili öğrenme yetenekleri sayesinde, iş istasyonlarından gömülü cihazlara kadar çok geniş bir platform yelpazesine nüfuz etmiştir.

Diğer taraftan, CNN temelli bilgisayar görmesi uygulamaları, iş istasyonları gibi enerjisini şebekeden sağlayan cihazlarda, maliyetli olmasına rağmen donanım kısıtlaması bulunmadığından herhangi bir CNN modeli (Örn:

IJPAS
ijpas@munzur.edu.tr
ISSN: 2149-0910

Research article/Araştırma makalesi
DOI:10.29132/ijpas.995579

AlexNet(Krizhevsky, Sutskever, & Hinton, 2012), VGG (Simonyan & Zisserman, 2015) vb. gibi) ayırt edilmeden kullanılabilir. Ancak, enerjisini şebekeden almayan cihazlarda; işlemci, hafıza, depolama ve enerji birimi gibi donanım kısıtlamaları söz konusudur. Dolayısıyla, gömülü cihazlarda en az kaynak tüketen ve en yüksek başarıya sahip bir CNN modelini seçmek gerekir. Örneğin, ImageNet (ILSVRC-2012) veri seti ile eğitildiğinde VGG-16 modeli, 138 Milyon parametre sayısı ve 30.9 Milyar tane saniyede kayan nokta işlemine (FLOPs) sahipken MobileNetV2 (Sandler, Howard, Zhu, Zhmoginov, & Chen, 2018b), 3.4 Milyon parametre sayısına ve 0.3 Milyar tane FLOPs işlemine sahiptir (Han, Pool, Tran, & Dally, 2015; Sandler, Howard, Zhu, Zhmoginov, & Chen, 2018a). Bir CNN modelinde parametre sayısı hafıza kullanımının yoğunluğunu belirtirken FLOPs sayısı da hesaplama merkezi olan işlemci kullanımının yoğunluğunu belirtmektedir (Shawahna, Sait, & El-Maleh, 2019). Dolayısıyla, gömülü cihazlarda, daha az parametre sayısı ve FLOPs sayısına sahip bir CNN modeli kullanılmalıdır.

Mobil veya gömülü cihazların kaynak tüketimini dikkate alan çalışmalar içinde HBONet modelinin iyileştirilmesine yönelik herhangi bir çalışma olmamasına rağmen diğer modellerle ilgili çalışmalara rastlamak mümkündür. Bunlardan MobileNetV1(Koonce, 2021),istiflenen derinlemesine ayrıştırılabilir evrişim katmanları üzerine inşa edilmiştir. Bu sayede, mobil cihazlar için daha hafif (lightweight) bir CNN omurgası elde edilmiştir. ShuffleNetV1 (Zhang, Zhou, Lin, & Sun, 2018), noktasal evrişimlerin karmaşıklığını azaltmak için noktasal grup evrişimlerden yararlanan artıklıdarboğazları (residualbottlenecks) ve kanallar arası korelasyonları geliştirmek için kanal karıştırma işlemlerini kullanır. ShuffleNetV2 (Ma, Zhang, Zheng, & Sun, 2018) ise ShuffleNetV1 (Zhang ve ark., 2018)'deki kanal karıştırma işlemlerini sürdürmektedir. Ek olarak, bu modelde, öznitelik kanallarının özel konfigürasyonunun ve temel işlem sırasının önerilen pratik kılavuzlara daha iyi uyacak bir şekilde düzenlendiği için kaynak tüketimini daha etkin kullanabilecek tasarımlar sunmaktadır. MobileNetV1 (Koonce, 2021)'i temel alarak geliştirilen MobileNetV2 (Sandler ve ark., 2018a), doğrusal darboğazlara sahip ters artıklı yapısına (invertedresidualstructure) dayanmaktadır. MobileNetV2, doğruluk ve verimlilik açısından önceki sürüme göre daha iyi bir denge sağlamaktadır.

MobileNetV1 (Howard ve ark., 2017), MobileNetV2 (Sandler ve ark., 2018b) ve HBONet (Li, Zhou, & Yao, 2019) CNN modelleri, gömülü cihazlar için tanıtılan birkaç modelden biridir. Bu modeller içinde kaynak tüketimini ve doğruluğu en iyileştirilen model, HBONet'tir. Ancak, gömülü cihazlar için kaynak tüketimi ve doğruluk açısından HBONet modelinin daha da iyileştirilerek performansının arttırılması bu çalışmanın ilham kaynağı olmuştur. Bu çalışmada, HBONet modelini temel alarak türettiğimiz A-HBONet modelini önerdik. Önerilen model, HBONet'in kullandığı katman ve blok yapılarını baz alarak türetilmiştir. A-HBONetmodelinin HBONet modelinden en büyük farkı, bazı hiper parametrelerin ve yığın (stack) yapısının farklı olmasıdır. Önerilen ağ modeli olan A-HBONet 'in mevcut HBONet modeline kıyasla katkıları şunlardır;

- Nesne tanıma görevlerinde performans ölçütü olarak kullanılan doğruluk oranı yaklaşık %3 arttırılmıştır,
- Gömülü sistemlerde en önemli kaynaklardan biri olan hafıza tüketimi, yaklaşık olarak %80 oranında azaltılmıştır, bir diğer önemli kaynak olan depolama birimi tüketimi ise yaklaşık olarak %79 oranında azaltılmıştır.

Gerçekleştirilen deneysel testler sonucunda önerilen modelin parametre sayısı ve model boyutu muazzam bir şekilde azaltılırken modelin nesne tanıma başarısı da önemli ölçüde arttırılmıştır. Dolayısıyla, önerilen model, gömülü cihazlarda diğer mobil CNN modellerine göre daha etkin ve verimli bir şekilde çalışabilmektedir.

Bu makalenin geri kalanı şu şekilde organize edilmiştir; İkinci bölümde ilgili çalışmalara yer verilirken, üçüncü bölümde temel mimari ele alınmıştır. Dördüncü bölümde önerilen yaklaşım, öğrenme oranı planlaması ve veri artırma sunulurken, beşinci bölümde deneyler ve sonuçlar sunulmaktadır. Son olarak, altıncı bölümde genel sonuçlar paylaşılmaktadır.

HBONET MİMARİSİ

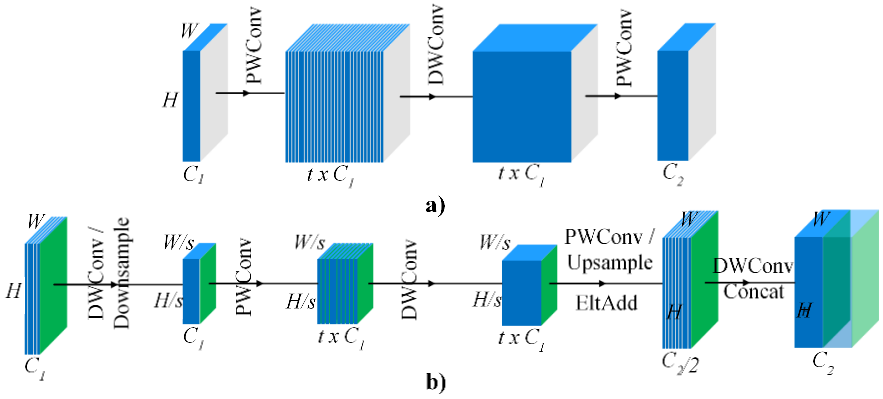
HBONet (Li ve ark., 2019) CNN modeli, MobileNetV2 (Sandler ve ark., 2018a) referans alınarak geliştirilmiştir. Özellikle, MobileNetV2 'in darboğaz blokları,iki ortogonal boyut (HBO) boyunca uzanan HarmoniousBottleneck

Int. J. Pure Appl. Sci. 8(1);53-60 (2022)

IJPAS
ijpas@munzur.edu.tr
ISSN: 2149-0910

Research article/Araştırma makalesi
DOI:10.29132/ijpas.995579

(HarmoniousBottleneck on two Orthogonaldimensions - HBO) blokları,evrişimli katman ve ters artıklı blokların oluşturduğu



Şekil 1. Darboğaz katmanlarını kıyaslama; a) Derinlemesine Ayrılabilir Evrişim Katmanı, b) HarmoniousBottleneck Katmanı (Li ve ark., 2019).

yığın yapısı ile değiştirilerek inşa edilmiştir. Bu sayede, hesaplama maliyetinin düşürülmesi amaçlanmıştır. HBO, uzamsal bir daralma-genişletme bileşeni ve bir kanal genişletme-daralma bileşeninden oluşan iki taraflı simetrik bir yapı aracılığıyla, derinlemesine evrişimsel özelliklerin uzamsal ve kanal boyutları arasındaki karşılıklı bağımlılıkları ortaklaşa kodlayarak meydana gelmektedir. Hem uzamsal hem de kanal boyutlarına odaklanmak için HarmoniousBottleneck yaklaşımında derinlemesine ayrıştırılabilir evrişim kullanılır. Bu yöntem iki bölüme ayrılmıştır: ilk olarak, $h/s \times w/s \times c_1$kanallarını sabit tutarken uzamsal boyutun aşağı doğru örneklenmesi ve daha sonra $h/s \times w/s \times t \times c_1$ kanallarının genişletilmesi ve ikincisi, $h \times w \times c_2/2$kanal indirgemesinin yarıya indirilmesi sırasında uzamsal boyutların yukarı doğru örneklenmesi ve son olarak $h \times w \times c_2/2$giriş tensörünün veya birleştirilmiş(pooled) versiyonunun kısmi kanallarıyla birleştirilmesi. Uzamsal daralma-genişletme bileşenini mevcut bloklarla birleştirdikten sonra, toplam hesaplama maliyeti şu şekilde olur:

$$Cost = \frac{B}{s^2} + \left[\left(\frac{h}{s} \times \frac{w}{s} \times c_1\right) + (h \times w \times c_2)\right] \times k^2 \qquad (1)$$

Denklem (1) 'de *B*, uzamsal daraltma ve genişletme işlemleri arasına yerleştirilen blokların orijinal hesaplama maliyetini göstermektedir. Bunun yanı sıra, s, adımı (stride) ifade ederken k, çekirdek boyutunu (kernel size), h ve w sırasıyla öznitelik haritasının (featuremap) yükseklik ve genişliğini, c ise kanal sayısını ifade etmektedir.

İki ortogonal boyut (HBO) boyunca uzanan HarmoniousBottleneck, Şekil 1'de görüldüğü gibi; iki taraflı simetrik bir biçimde düzenlenmiş olan uzamsal boyutların (h, w) daralması-genişlemesi ve kanal boyutlarının (c) genişlemesi-daralması olmak üzere iki bölümden oluşur. İlk olarak, kanal boyutları sabit tutulur ve uzamsal boyutlardan öznitelikler çıkarılır. İkinci olarak, uzamsal boyutlar sabit tutulur ve kanal boyutlarından öznitelikler çıkarılır. Bu işlemler, modelin doğruluğunu artırmaktadır.

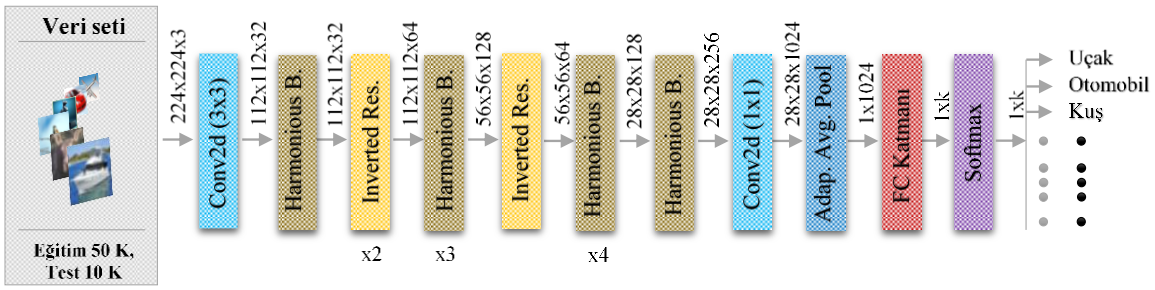
HBONet, öznitelik çıkarmak için evrişimli katmanlarını, mobil terslenmiş (inverted)darboğaz ve harmoniousbottleneck katmanlarını kullanır. Tablo 1 'de orijinal HBONet mimarisi sunulmuştur.

Tablo 1. *HBONet Mimarisi (Li ve ark., 2019).*

Girdi B.	Blok/Katman	*t*	*c*	*n*	*s*
224^2x3	Conv2d 3x3	-	32	1	2
112^2x32	Harmonious Bottleneck	1	20	1	1
112^2x20	Harmonious Bottleneck	2	36	1	1
112^2x36	Harmonious Bottleneck	2	72	3	2
56^2x72	Harmonious Bottleneck	2	96	4	2
28^2x96	Harmonious Bottleneck	2	192	4	2
14^2x192	Harmonious Bottleneck	2	288	1	1
14^2x288	Conv2d 1x1	-	144	1	1
14^2x144	Inverted Residual	6	200	2	2
7^2x200	Inverted Residual	6	400	1	1
7^2x400	Conv2d 1x1	-	1600	1	1
7^2x1600	Avgpool	-	-	1	-
1^2x1600	Conv2d 1x1	-	k	-	-

IJPAS
ijpas@munzur.edu.tr
ISSN: 2149-0910

Research article/Araştırma makalesi
DOI:10.29132/ijpas.995579



Şekil 2. Önerilen yaklaşımın şematik tasarımı

ÖNERİLEN YAKLAŞIM: A-HBONET

Orijinal HBONet(Li ve ark., 2019) CNN modelinin doğruluğunu arttırmak ve kaynak tüketimini azaltmak için evrişimliveya darboğaz katmanlarını verimli bir şekilde kullanarak öznitelikleri çıkarmak gerekir. Önerilen mimari olan A-HBONet 'te doğruluğu arttırıp kaynak maliyetini düşürmek için orijinal HBONet 'teki ters artıklı(InvertedResidual), HarmoniousBottleneckve Noktasal (Pointwise) evrişimli katmanlarının genişleme faktörü (t), kanal sayısı (c), tekrar faktörü (n) ve adım (stride) boyutu (s) gibi bazı hiper parametreler ve istiflenme düzeni değiştirilmiştir.

Tablo 2. *Arttırılmış HBONet (A-HBONet) Mimarisi*

Girdi B.	Blok/Katman	*t*	*c*	*n*	*s*
224^2x3	Conv2d 3x3	-	32	1	2
112^2x32	Harmonious Bottleneck	2	32	1	1
112^2x32	Inverted Residual	1	64	2	1
112^2x64	Harmonious Bottleneck	2	128	3	2
56^2x128	Inverted Residual	1	64	1	1
56^2x64	Harmonious Bottleneck	1	128	4	2
28^2x128	Harmonious Bottleneck	2	256	1	1
28^2x256	Conv2d 1x1	-	1024	1	1
28^2x1024	Avgpool	-	1024	1	-
1x1024	FC Layer	-	k	1	-
1xk	Sofmax	-	k	1	-

Diğer taraftan, ters artıklı ve harmonious bottleneck gibi farklı blok gruplarından elde edilen ara öznitelikleri, düşük boyutlu temsil edebilmek için noktasal evrişimli katmanı da kullanılmaktadır. Tablo 2 'de önerilen modelin mimarisi verilirken Şekil 2 'de de önerilen yaklaşımın şematik tasarımı verilmiştir. Bu model ile orijinal HBONet modelinin doğruluğu yaklaşık olarak %3 arttırılırken model boyutu 9 MB kadar azaltılmıştır.

Öğrenme Oranı Planlaması (LR Scheduling)

Orijinal HBONet modeli ve önerilen A-HBONet modelinin doğruluğunu arttırmak için farklı öğrenme oranları (LR) iteratif olarak kullanılmıştır. Bunlar içerisinde en ünlü olanlar; Step LR, Multistep LR, Exponential LR, Cosine Annealing LR 'dır. Bu çalışmada, diğer LR'lerden daha iyi bir sonuç ürettiği için Cosine Annealing Öğrenme Oranı(Loshchilov & Hutter, 2017) kullanılmıştır. CosineAnnealing LR yöntemi ile her bir batch 'te öğrenme oranı aşağıdaki denklem ile hesaplanır;

$$n_t = n_{min}^i + \frac{1}{2}(n_{max}^i - n_{min}^i)\left(1 + \cos\left(\frac{T_{cur}}{T_i}\pi\right)\right) \quad (2)$$

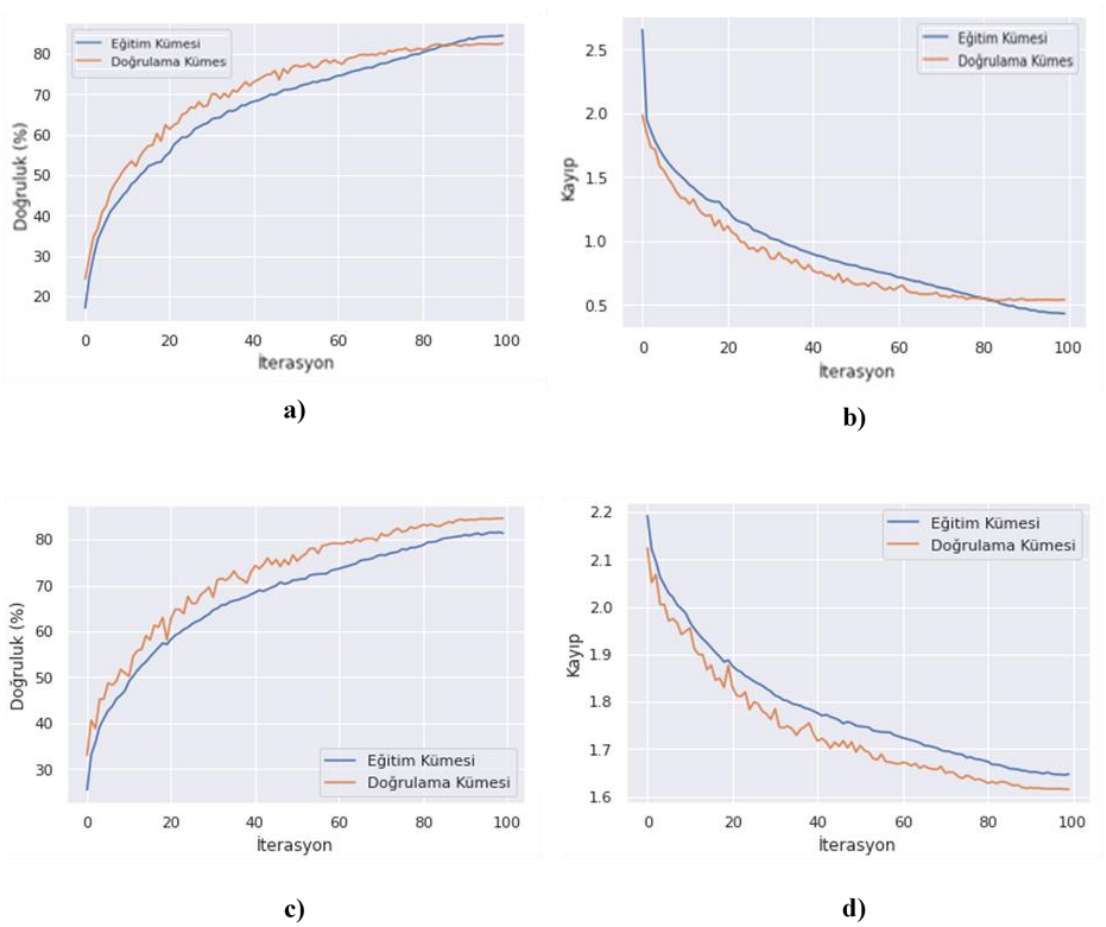
Denklem (2) 'de n_{min}^i ve n_{max}^i, öğrenme oranı aralığıdır. T_{cur}, son yeniden başlatmadan beri gerçekleşen iterasyon sayısıdır. i, yürütme indeksi ve t, herbir batch iterasyonunu ifade etmektedir.

Bu çalışmada, The Stochastic Gradient Descent (SGD) algoritması temelli Cosine Annealing yöntemi ile planlama (scheduling) yapılmıştır. Bunun için SDG parametrelerinden öğrenme oranı, 0.1 'e, momentum 0.9, weightdecay10^{-4} 'e ayarlanmıştır ve eğitim 100 iterasyon (epoch) kadar sürmüştür.

Int. J. Pure Appl. Sci. 8(1);53-60 (2022)

IJPAS
ijpas@munzur.edu.tr
ISSN: 2149-0910

Research article/Araştırma makalesi
DOI:10.29132/ijpas.995579



Şekil 3. CNN modellerinin doğruluk ve kayıp grafikleri; a) HBONet doğruluk, b) HBONet kayıp, c) A-HBONet doğruluk and d) A-HBONet kayıp

Veri Arttırma

Cutout(DeVries & Taylor, 2017), veri arttırmada (dataaugmentation) kullanılan tekniklerden biridir. Bu teknik, eğitim esnasında girdi görüntülerinin rastgele bölümlerini maskeleyerek / kapatarak veri kümesini etkin bir şekilde arttırmayı amaçlayan evrişimlisinir ağları için basit bir düzenleme (regularization) tekniğidir. Cutout, kapalı örnekleri simüle eder ve modeli karar verirken birkaç ana özelliğin varlığına güvenmek yerine daha küçük özellikleri dikkate almaya teşvik eder.

Bu çalışmada, ağ modellerinin eğitiminden önce CIFAR-10 veri setine; cutout, rastgele yatay çevirme ve rastgele kırpma teknikleriyle veri arttırma uygulanmıştır. Bu sayede, CIFAR-10 veri setinin daha kararlı olması amaçlanmıştır.

DENEYLER SONUÇLAR ve TARTIŞMA

Bu çalışmada, 50 bin eğitim görüntüsünden ve 10 bin test görüntüsünden oluşan CIFAR10 (Krizhevsky, 2009) veri seti,yine aynı sayıda görüntü içerecek biçimde fakat daha dengeli olması için veri arttırma yapılarak kullanılmıştır. Orijinal HBONet, önerilen A-HBONet CNN modelleri ve yeniden uygulaması yapılan diğer CNN modelleri, veri arttırma teknikleri uygulanan CIFAR-10 veri seti ile 100 iterasyon boyunca eğitime tabi tutulmuştur. Batch size, eğitim için 128 ve doğrulama (validation) için 100 olarak belirlenmiştir. İterasyon sayısı ve batch size, ağ modellerinin öğrenmesini en hızlı sağlayacak ve ezberleme yapmayacak bir şekilde iteratif deneme ile belirlenmiştir. Modellerin eğitimi ve deneysel testleri için aşağıdaki donanım ve yazılımlar kullanılmıştır;

IJPAS
ijpas@munzur.edu.tr
ISSN: 2149-0910

Research article/Araştırma makalesi
DOI:10.29132/ijpas.995579

- Intel(R) Xeon(R) CPU @ 2.30GHz,
- 24 GB RAM,
- Tesla P100-PCIE-16GB,
- Python 3.7.11 versiyonu,
- Pytorch 1.9.0 versiyonu.

Deneysel testleri kaynak tüketimi ve doğruluk açısından değerlendirebilmek için modellerin doğruluğu, parametre sayısı, saniyede kayan nokta işlemi(FLOPs) ve model boyutu gibi metrikler kullanılmıştır. Şekil 3 'te orijinal HBONet ve önerilen HBONet (A-HBONet) modellerinin doğruluk ve kayıp grafikleri verilmiştir. Bu grafiklerden de görüldüğü gibi HBONet modelinin doğruluğu maksimum yaklaşık olarak %81 oranında başarı sağlarken A-HBONet 'te bu oran maksimum yaklaşık

Tablo 3. *CIFAR10 seti ile çeşitli genişlik çarpanlarına sahip farklı CNN modelleri için performans karşılaştırması*

GenişlikÇarpanı (Width Multiplier)	Model Adı	Genel Doğruluk (%)	Parametre Sayısı (M)	GFLOPs (224x224)	Model Boyutu (MB)
1.00	HBONet	81.31	2.98	0.31	11.6
	A- HBONet	**84.57**	**0.61**	**0.58**	**2.49**
0.75	HBONet	81.91	1.87	0.20	7.38
	A- HBONet	**83.33**	**0.41**	**0.38**	**1.70**
0.50	HBONet	72.34	0.95	0.10	3.86
	A- HBONet	**81.1**	**0.24**	**0.22**	**1.05**
0.25	HBONet	68.08	0.35	0.04	1.52
	A- HBONet	**75.03**	**0.11**	**0.09**	**0.55**

olarak %84.5 'e çıkmaktadır. Kayıp grafikleri değerlendirildiğinde de HBONet modeli daha başarılı olduğu görülmektedir.

CIFAR10 test seti kullanılarak HBONet ve A-HBONet CNN modellerinin evrişim katmanlarındaki öznitelik çıkarma sayısını belirleyen çeşitli genişlik çarpanlarına (Width Multiplier) göre performans kıyaslaması Tablo 3'te verilmiştir. Performans kıyaslamasında kullanılan metriklerinden FLOPs için girdi boyutu 3x224x224 olarak verilirken diğer metrikler için CIFAR10 veri seti kullanılmıştır. Bu performans metriği önerilen modelde, orijinal modele göre daha fazla hesaplama karmaşıklığı yaratsa bile doğruluk, parametre sayısı ve model boyutunda dikkate değer bir iyileştirme yapıldığı gözlemlenmektedir. Diğer taraftan, tablo 3'te yer alan genel doğruluk ile veri setindeki doğru tahmin edilen görüntülerin tüm görüntülere oranı vurgulanmaktadır. A-HBONet modeli, genişlik çarpanı azalsa dahi genel doğruluk performans metriğindeki yüksek doğruluk başarısını sürdürme eğilimindedir. Dolayısıyla, önerdiğimiz model doğruluk metriği için daha kararlı bir yapıya sahiptir.

Tablo 4. *Literatürdeki diğer çalışmalarla sonuçlarının karşılaştırılması*

AğModeli	Yıl	Dataset	Param. Sayısı (Milyon)	GenelDoğruluk(%)	GFLOPs (224x224)
ShuffleNetV2 (Ma ve ark., 2018)	2018	ImageNet	-	69.4	0.15
MobileNetV1 (Howard ve ark., 2017)	2017	ImageNet	4.2	70.6	0.57
MobileNetV2 (Sandler ve ark., 2018b)	2018	ImageNet	3.4	72.0	0.30
HBONet(Li ve ark., 2019)	2019	ImageNet	-	73.1	0.30
ShuffleNetV2 (Ma ve ark., 2018) (re-imp)	-	Cifar10	2.3	81.9	0.15
MobileNetV2 (Sandler ve ark., 2018b) (re-imp)	-	Cifar10	2.2	81.9	0.31
HBONet(Li ve ark., 2019) (re-imp)	-	Cifar10	2.9	81.3	0.31
A- HBONet	-	Cifar10	**0.6**	**84.5**	**0.58**

Genişlik çarpanı 1.0 için, önerilen modelde, orijinal HBONet'e göre yaklaşık olarak %3 oranında doğruluk değeri daha yüksek çıkmaktadır. Bu oran genişlik faktörünün azalmasıyla daha da artmaktadır.Bir diğer metrik olan parametre sayısı değerlendirildiğinde ise genişlik çarpanı 1.0 'da, önerilen model orijinal modele göre yaklaşık 4 kat daha az parametreye sahiptir. Parametre sayısı, bir

IJPAS
ijpas@munzur.edu.tr
ISSN: 2149-0910

Research article/Araştırma makalesi
DOI:10.29132/ijpas.995579

cihazın kullandığı hafıza (memory) miktarını belirlediği için gömülü cihazlar için çok önemli bir metriktir. Model boyutu metriği ise orijinal modele göre yaklaşık olarak 4 kat azaltılmıştır. Bu metrikte gömülü cihazlarda kullanılan depolama birimini ne kadar kullandığını belirttiği için gömülü cihazlarda çok önemli bir başka metriktir. Önerilen CNN modeli ile gömülü cihazlardaki kısıtlı donanım kaynakları için parametre sayısı ve model boyutu etkin şekilde azaltılırken modelin doğruluk başarısının da dikkate değer bir şekilde arttırıldığı görünmektedir.

Tablo 4'te literatürde yer alan diğer mobil CNN yöntemleri ile bizim önerdiğimiz A-HBONet modelinin karşılaştırılması 1.0 genişlik çarpanına göre verilmiştir. Mobil CNN modellerinden ShuffleNetV2, MobileNetV1-V2 ve HBONet, ImageNetveri seti kullanılarak deneysel testleri yapılmıştır. Bizim çalışmamızda ise donanım kaynakları yetersiz olduğu için ImageNetdataseti yerine daha küçük boyutlu olan Cifar10 veri seti kullanılmıştır. Dolayısıyla değerlendirmeyi daha sağlıklı yapabilmek için ShuffleNetV2, MobileNetV2 ve HBONet modelleri, Cifar10 veri seti kullanılarak yeniden uygulaması (re-imp) yapılmıştır. ImageNetdataseti ile yapılan çalışmalarda en yüksek genel doğruluk oranına HBONet, %73.1 ile ulaşmıştır. Cifar10 dataseti ile yapılan çalışmalarda ise en yüksek genel doğruluk oranına %84.5 ile önerdiğimiz A-HBONet modeli ulaşmıştır. Bununla birlikte A-HBONet modeli, diğer modellere göre en az parametre sayısına sahiptir. Dolayısıyla, en az hafıza tüketimini önerdiğimiz model sağlamaktadır. Diğer taraftan önerdiğimiz modelin, diğer modellere göre FLOPs sayısı daha fazladır. Ancak bu, gelişen işlemci teknolojileri sayesinde bir sorun teşkil etmemektedir.

Sonuç olarak, önerdiğimiz model olan A-HBONet'in literatürdeki diğer çalışmalara göre avantajı düşük hafıza ve depolama birimi kullanımı ile yüksek doğruluk oranı sağlamasıdır. A-HBONet'in bu başarısı, mobil cihazlarda daha verimli ve etkin kullanımını mümkün kılmaktadır.

SONUÇ

Bu çalışmada, HBONet CNN mimarisi temel alınarak ve bazı değişiklikler yapılarak yeni bir mimari olan A-HBONet önerilmiştir. Önerilen model, HBONet modelinin bazı katman ve blok yapılarını kullanarak türetilmiştir. A-HBONet modelinin HBONet modelinden en önemli farkı kullanılan hiper parametreler ve yığın (stack) yapısıdır. Önerilen model, diğer modellere göre doğruluk oranı daha yüksek, kaynak tüketimi daha az olan verimli bir modeldir. Doğruluk oranı orijinal modele göre yaklaşık olarak %3 daha yüksektir. Bunun yanı sıra, A-HBONet ile birlikte parametre sayısı, 2.98 Milyondan 0.61Milyona ve model boyutu 11.6 Megabayttan 2.49 Megabayta düşürülmüştür. Bu sayede, önerilen modelin gömülü cihazlarda verimli ve etkin bir şekilde kullanılmasına olanak sağlanmıştır. Önerilen model, mobil cihazlarda görüntü sınıflandırma, nesne tanıma ve semantik bölütleme görevleri için kullanılabilir. Gelecek çalışmalarımızda, gömülü sistemler için sadece modelin hafıza ve depolama birimi değil, işlemci biriminin kullanımını da azaltacak yöntemler üzerinde çalışma yapmayı planlamaktayız.

ÇIKAR ÇATIŞMASI BEYANI

Yazarlar bu makale ile ilgili herhangi bir çıkar çatışması bildirmemektedir.

ARAŞTIRMA VE YAYIN ETİĞİ BEYANI

Yazarlar bu çalışmanın araştırma ve yayın etiğine uygun olduğunu beyan eder.

REFERENCES

DeVries, T., & Taylor, G. W. (2017). *Improved Regularization of Convolutional Neural Networks with Cutout*. Retrieved from http://arxiv.org/abs/1708.04552

Han, S., Pool, J., Tran, J., & Dally, W. J. (2015). Learning both weights and connections for efficient neural networks. *Advances in Neural Information Processing Systems, 2015-Janua*, 1135–1143.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., … Adam, H. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. *ArXiv*.

Koonce, B. (2021). MobileNet v1. *Convolutional Neural Networks with Swift for Tensorflow*, 87–97. https://doi.org/10.1007/978-1-4842-6168-2_8

Krizhevsky, A. (2009). *CIFAR10 Dataset*. Retrieved from https://www.cs.toronto.edu/

IJPAS
ijpas@munzur.edu.tr
ISSN: 2149-0910

Research article/Araştırma makalesi
DOI:10.29132/ijpas.995579

kriz/cifar.html

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Proceedings of the 25th International Conference on Neural Information Processing Systems*, *1*, 1097–1105. https://doi.org/10.1145/3065386

Li, D., Zhou, A., & Yao, A. (2019). HBONet: Harmonious bottleneck on two orthogonal dimensions. *Proceedings of the IEEE International Conference on Computer Vision*, *2019-Octob*, 3315–3324. https://doi.org/10.1109/ICCV.2019.00341

Loshchilov, I., & Hutter, F. (2017). SGDR: Stochastic gradient descent with warm restarts. *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, 1–16.

Ma, N., Zhang, X., Zheng, H. T., & Sun, J. (2018). Shufflenet V2: Practical guidelines for efficient cnn architecture design. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11218 LNCS*, 122–138. https://doi.org/10.1007/978-3-030-01264-9_8

Rinner, B., & Wolf, W. (2008). An introduction to distributed smart cameras. *Proceedings of the IEEE*, *96*(10), 1565–1575. https://doi.org/10.1109/JPROC.2008.928742

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018a). MobileNetV2: Inverted residuals and linear bottlenecks. *ArXiv*, 4510–4520.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018b). MobileNetV2: Inverted Residuals and Linear Bottlenecks. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 4510–4520. https://doi.org/10.1109/CVPR.2018.00474

Shawahna, A., Sait, S. M., & El-Maleh, A. (2019). FPGA-Based accelerators of deep learning networks for learning and classification: A review. *IEEE Access*, *7*, 7823–7859. https://doi.org/10.1109/ACCESS.2018.2890150

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 1–14.

Weibin Liu, Chao Zhang, & Baozong Yuan. (2002). *AVR theory, techniques and application.* (69775003), 1163–1166. https://doi.org/10.1109/icosp.2000.891751

Zhang, X., Zhou, X., Lin, M., & Sun, J. (2018). ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 6848–6856. https://doi.org/10.1109/CVPR.2018.00716