Parça Tabanlı Eğitimin Evrişimli Yapay Sinir Ağları ile Nesne Konumlandırma Üzerindeki Etkisi
Effect of Patch Based Training on Object Localization with Convolutional Neural Networks

Semih Orhan, Yalın Baştanlar
Bilgisayar Mühendisliği Bölümü
İzmir Yüksek Teknoloji Enstitüsü
Urla, İzmir, Türkiye
semihorhan@yandex.com, yalinbastanlar@iyte.edu.tr

***Özetçe* —Evrişimli yapay sinir ağları (EYSA) bir derin öğrenme yöntemi olarak son yıllarda imge sınıflandırma ve nesne konumlandırma başta olmak üzere bilgisayarlı görü problemlerinin çözümü konusunda birçok başarı elde etmiştir. Bunun için farklı yapıda ve derinlikte birçok model geliştirilmiştir. Bu çalışmamızda, leoparların resimler içerisindeki konumları evrişimli yapay sinir ağları ile bulunmuştur. Konum bulmak için, literaturde sıkça kullanılan: içinde nesne olan ve olmayan imgelerle model eğitme ve bizim önerdiğimiz: imgeden nesneye ait bölgelerden ve nesneye ait olmayan bölgelerden alınan parçalar ile model eğitme yöntemi karşılaştırılmıştır. Resimlerden parçalar alınarak eğitilen modelin, tüm imge ile eğitilen modelere göre üstün başarı gösterdiği gözlenmiştir.**

Anahtar Kelimeler—Derin yapay sinir ağları, evrişimli yapay sinir ağları, nesne tanıma, nesne konumlandırma.

***Abstract*—In recent years, Convolutional Neural Networks (CNNs) have shown great performance not only in image classification and image recognition tasks but also several tasks of computer vision. A lot of models which have different number of layers and depths, have been proposed . In this work, locations of leopards are tried to be identified by deep neural networks. To accomplish this task, two different methods are applied. First of them is training neural network using with entire images, second of them is training neural networks using with image patches which are cropped from full size of images. Patch training model has shown better performance than full size of image trained model.**

Keywords—Deep neural networks, convolutional neural networks, object recognition, object localization.

I. GİRİŞ

Bilgisayarla görü alanında nesneleri tanımak için geliştirilen algoritmalar, imgeler içinde yer alan nesneleri doğru bir şekilde etiketlemeye ve o nesnelerin etrafını saracak bir kutu çizmeye çalışırlar. Günümüze kadar bu amaç için birçok algoritma geliştirilmiştir. Bunlardan bazıları: anahtar nokta torbası (Bag of keypoints [1]), yönlü gradyan histogramları (Histogram of oriented gradients [2]) ve esnek parçalı modelleme (Deformable Part Models [3]) yaklaşımları kullanan çalışmalardır. Günümüzde ise evrişimli yapay sinir ağları (EYSA) popülerlik kazanmıştır.

Yapay sinir ağları matematiksel olarak ilk defa 1943 yılında modellenmiş olmasına rağmen [4], sonraki yıllarda hakkında yapılan olumsuz eleştiriler [5], bu alanda çalışan araştırmacıları uzaklaştırmış, yapılan yatırımları azaltmıştır. 2012 yılında, AlexNet [6], ILSVRC [7] yarışmasının imge sınıflandırma kategorisinde üstün bir başarı etmiş ve birinci sırayı alan ilk EYSA olmuştur. Sonraki yıllarda EYSA'lar hep üstün gelmiştir. Bu başarım, araştırmacıların dikkatini tekrardan bu alana çekmiştir. 2015 yılında ResNet [8], kaybolan gradyan problemine çözüm getirerek imge sınıflandırma ve nesne tanıma kategorilerinde birinci olmuştur.

Nesne tanıma için geliştirilmiş birçok EYSA modeli bulunmaktadır. Bunlardan bazıları: OverFeat [9], Faster R-CNN [10], Oquab et al. [11]'dır. OverFeat modelinde, EYSA önce sınıflandırıcı olarak eğitilmiş, daha sonra eşiksiz en büyük işlev (softmax) katmanı silinip yerine bağlanım (regression) katmanı eklenerek nesnenin konumu tahmin edilmeye çalışılmıştır. Faster R-CNN [10] modelinde ise, nesnenin olası konumları son evrişim katmanı üzerinden kayan pencere yaklaşımı ile otomatik olarak önerilmiş ve önerilen konumlar içinde nesne aranmıştır. Bu şekilde önemli bir hız artışı sağlanmıştır. Oquab et al. [12], EYSA'ları imge sınıflandırması için kullanılmış olsa bile nesne konumları hakkında bilgi verdiğini söylemiştir. Aynı yazarlar daha sonraki bir çalışmada [11], ağ yapısını değiştirmiş, yapay sinir ağlarını eğitmek için sıkça kullanılan gözetimli (supervised) eğitim yönteminin yerine, yarı-gözetimli (weakly-supervised) bir eğitim yöntemi izleyerek nesnelerin konumlarını bulmaya çalışmıştır. Yarı-gözetimli yöntemin izlenmesindeki amaç, gözetimli eğitim yönteminin aksine, her bir nesneyi kapsayan kutu çizimine gerek olmaması, bu sayede veri kümesi hazırlamak için gereken insan iş gücünü büyük ölçüde azalmasıdır.

Çalışmamızda, EYSA ile imge içerisindeki nesnenin yerini bulmak için iki farklı eğitim yöntemi izlenmiştir. İlk yöntem Overfeat'e [9] benzer bir yaklaşım olup, model nesneyi içeren ve içermeyen imgeler ile eğitilmiştir ve ardından test imgelerinde boyutu değişen ve kayan pencereler yöntemi ile nesne aranmıştır. Bizim önerdiğimiz yöntemde ise, her iki sınıf için imgenin hedef nesneye ait olan ve olmayan yerlerinden

978-1-5090-6494-6/17/$31.00

parçalar (yamalar) alınarak eğitim işlemi yapılmıştır. Yapay sinir ağları, ResNet [8] modeli kullanılarak eğitilmiştir. Başarım test imgelerindeki leopar konumlarının doğru bir şekilde bulunması olup, önerilen eğitim yönteminin, tüm imge ile eğitim yöntemine göre daha başarılı olduğu gözlenmiştir.

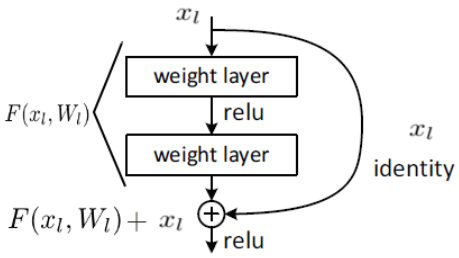
Bildirimizin diğer kısımları şu şekilde organize edilmektedir: Bölüm 2'de derin artık öğrenme yöntemi anlatılmakta, Bölüm 3'de veri kümesinin oluşturulması hakkında bilgi verilmekte, 4. Bölüm'de, nesne konumlarını bulmak için uygulanan yöntemler anlatılmakta, sonuçlar kısmı ise 5. Bölüm'de yer almaktadır.

II. Derin Artık Öğrenme

EYSA modellerinin başarım oranı kullanılan modelin sınıflandırma kabiliyetine bağlıdır. Bunun için çalışmamızda, en güncel modellerden ResNet [8] modeli seçilmiştir. ResNet modelinde, katman sayısı arttığında başarımın düşmesi şeklinde kendini gösteren bozulma (degradation) problemi için çözüm üretilmiş, derin artık öğrenme (deep residual learning) yöntemi önerilmiştir [8]. Derin artık öğrenme yöntemi birçok artık bloktan oluşur. Artık bloklar değişen aralıklarla ve sıklıklarla oluşturulabilir. Genel denklem aşağıda görülmektedir [13]:

$$y_l = F(x_l, W_l) + x_l \tag{1}$$

Denklem (1)'de yer alan x_l girdi değeri, y ise çıktı değeridir. Bir sonraki katman, $x_{l+1} = f(y_l)$ ile elde edilir. f ise aktivasyon fonksiyonudur. $F(x_l, W_l)$ birden fazla evrişim katmanına sahip artık blokları temsil eder. x_l aktivasyon fonksiyonuna girdi olarak verilmeden artık öğrenme bloğunun çıktı değeri ile toplanır (Şekil 1 [8]). $F(x_l, W_l) + x_l$ işlemi, eleman eleman toplama işlemidir.



Şekil 1: Derin artık öğrenme bloğu.

Artık öğrenme bloğu, modelin karmaşıklığını arttırmaz. Eleman eleman toplama işleminin yapılabilmesi için x_l ve $F(x_l, W_l)$'nin in boyutlarının eşit olması gerekmektedir. Eğer x_l ve $F(x_l, W_l)$ in boyutları eşit değil ise, düşük boyut denklem (2)'deki formül ile arttırılır [8]:

$$y_l = F(x_l, W_l) + W_s x_l \tag{2}$$

Artık öğrenme bloğu, $F(x_l, W_l)$, istenilen sayıda katmanda oluşabilir. Genellikle, iki ya da üç katman ile oluşturulur. Artık öğrenme bloğunun, bir katmana sahip olduğu modellerde herhangi bir avantajı olmadığı gözlemlenmiştir.

Şekil 2: Parça veri kümesinden alınmış örnek resimler.

Şekil 3: Tüm imge veri kümesinden alınmış örnek resimler.

III. Veri Kümesinin Oluşturulması

Resimlerdeki leopar konumlarının bulunması için iki farklı sınıf oluşturulmuştur. Bu sınıflar: leopar ve arkaplan sınıflarıdır. Eğitim yöntemlerinin konum bulma üzerindeki etkisini ölçebilmek için ise iki farklı eğitim kümesi oluşturulmuştur. Bu yöntemlerden birincisi parça tabanlı eğitim (bizim önerdiğimiz) yöntemi, ikinci yöntem ise literatürde sıkça kullanılan tüm imge ile eğitim yöntemidir.

i) Parça (desen) tabanlı eğitim kümesi hazırlanırken: leoparların resimler içinde kapsadığı alanlardan 64x64 boyutunda parçalar alınmış ve pozitif eğitim kümesi olarak kullanılmıştır. İmgeler içerisinde leopar içermeyen analardan alınan 64x64'lük parça örnekleri ise negatif eğitim kümesi olarak kullanılmıştır.

ii) Tüm imge ile eğitim kümesi hazırlanırken: leopar içeren resimler 64x64 boyutuna ölçeklenmiş, arka plan resimleri hazırlanırken ise parça tabanlı eğitim kümesinin hazırlanmasında olduğu gibi, resimlerin leopar içermeyen kısımlarından 64x64 boyutunda parçalar alınarak hazırlanmıştır. Burada, öncesinde leopar içermeyen tüm resimleri negatif küme olarak kullanma yaklaşımı denenmiş, eğitilen model imge sınıflandırma için çok iyi bir performans göstermiş (%99.2) olmasına rağmen nesne konumlandırma için çok kötü bir performans vermiştir. Bunun nedeninin tüm imgede daha çok detay olması, imgeden alınan parçalarda ise daha az detay barındırması olarak değerlendirilmiştir. Netice olarak, parçalardan oluşan negatif küme, bu yaklaşım için daha iyi sonuç verdiği için tercih edilmiştir.

İki veri kümesi arasındaki tek fark, birinci veri kümesinde leopar sınıfı: pozitif örneklerden alınan parçalardan oluşurken, ikinci veri kümesinde leopar sınıfı: leopar içeren tüm imgelerden oluşmaktadır. Her iki veri kümesinden alınan örnek resimler sırasıyla Şekil 2 ve Şekil 3'te görülmektedir. Her bir sınıf, eğitim için yaklaşık olarak 800 imge içermektedir. Modeller 78 adet leopar imgesi ile test edilmiştir.

IV. Nesne Konumlandırma

A. Sıcaklık Haritası

Nesne konumumunun bulunması için kayan pencere yaklaşımı izlenmiştir. Parça tabanlı veri kümesi ile eğitilen modelde, 64x64 boyutunda ve her adımda 16 piksel kayan pencereler yapay sinir ağına verilmiş, sonuç değerlerinin sıcaklık haritası çizilmiştir. Tüm imge ile eğitilen model için ise tüm leopar bedeni imge içerisinde arandığından ve leoparların boyutları değiştiğinden klasik kayan pencereler yaklaşımı uygulanmış, en küçüğü 64x64 olmak üzere boyutları büyüyen ve her adımda 16 piksel kayan pencereler yapay sinir ağına verilmiş, ağ sonuçlarına göre sıcaklık haritası çizilmiştir. Sıcaklık haritasında 16x16'lık bir kutucuk için elde edilebilecek en yüksek değer o kutucuğu içeren her pencereden leopar sonucu alınmasıdır.

B. Nesnenin Sınırlarının Çizilmesi

Sıcaklık haritası üzerinde nesne kutularının çizilmesi için, sıcaklık haritası çıkarılan imgeler değişen eşik değerlerine göre ikili (binary) imgeye çevirilmiştir. Görültüleri gidermek için: açma (opening), kopan parçaları birleştirmek için ise kapama (closing) morfolojik işlemleri uygulanmıştır. Ardından, nesne konumlarını bulmak için bağlı bileşen algoritması kullanılmıştır. Örnek bir imgenin sıcaklık haritası, ikili imgeye çevrilmiş hali, açma ve kapama işleminin sonucu, tahmin edilen nesne konumu Şekil 4'te yer almaktadır. Parça tabanlı eğitim yöntemi ile tüm imge eğitim yöntemi arasındaki farkı görselleştirmek için, tahmin edilen nesne konumlarından bazılarına Şekil 5'te yer verilmiştir.

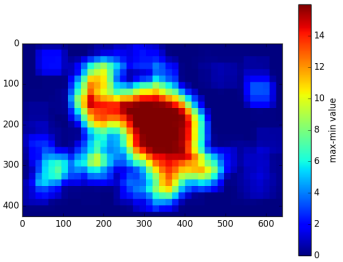
Sıcaklık haritasından elde edilen tahmin penceresi ile elle etiketlenen kapsayan kutunun örtüşme oranı denklem (3)'de verilen formül ile hesaplanır. Örtüşme oranı eşik değerin üzerindeyse, tahmin doğru olarak kabul edilir.

$$\textit{Örtüşme Oranı} = \frac{Kutu_{\text{tahmin}} \cap Kutu_{\text{etiket}}}{Kutu_{\text{tahmin}} \cup Kutu_{\text{etiket}}} \qquad (3)$$

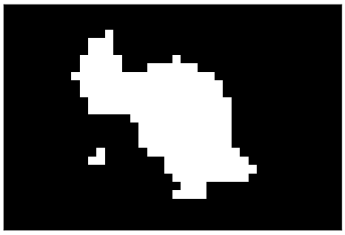
Parça tabanlı modelin test seti üzerindeki doğru tahmin oranları Tablo I'de, tüm imge ile eğitilen modelin doğru tahmin oranları Tablo II'de yer almaktadır. Sonuçları daha iyi kavrayabilmek için farklı örtüşme oranları ve farklı eşik değerleri ile karşılaştırma yapılmıştır. Kullanılan eşik değerleri sıcaklık haritasından alınabilecek en yüksek değer 1 olacak şekilde normalize edilmiş, 0.4-0.8 arası eşik değerlerinin sonuçları raporlanmıştır. Tablo I ve Tablo II'deki sonuçlara göre: parça tabanlı eğitim yönteminin, tüm imgeler ile eğitim yöntemine göre daha iyi bir performans gösterdiği gözlenmiştir. Şekil 5'teki örneklerden de görülebileceği üzere başarımdaki artışın nedeni, önerilen yöntemin çoğu durumda nesnenin sınırlarını girinti çıkıntıları da dahil iyi bir şekilde tespit etmesidir. Sonuçlar, konum bilgisi bulunmak istenen imge, özel bir desene sahipse, parça (desen) tabanlı eğitim yönteminin başarıyı arttıracağı tezini doğrulamaktadır.

V. Değerlendirme

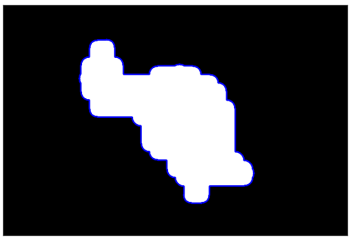
Bu çalışmada iki farklı eğitim yöntemi ile leoparların konumları bulunmaya çalışılmıştır. Birincisi yöntemde, resimlerden alınan parçalar ile model eğitilmiş, ikinci yöntemde ise tüm imge ile eğitim yapılmıştır. Parçalar ile eğitilen modelin,



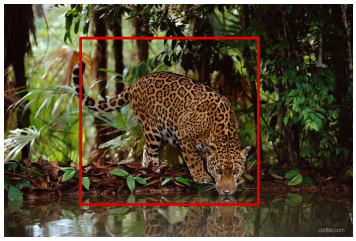
(a)



(b)



(c)



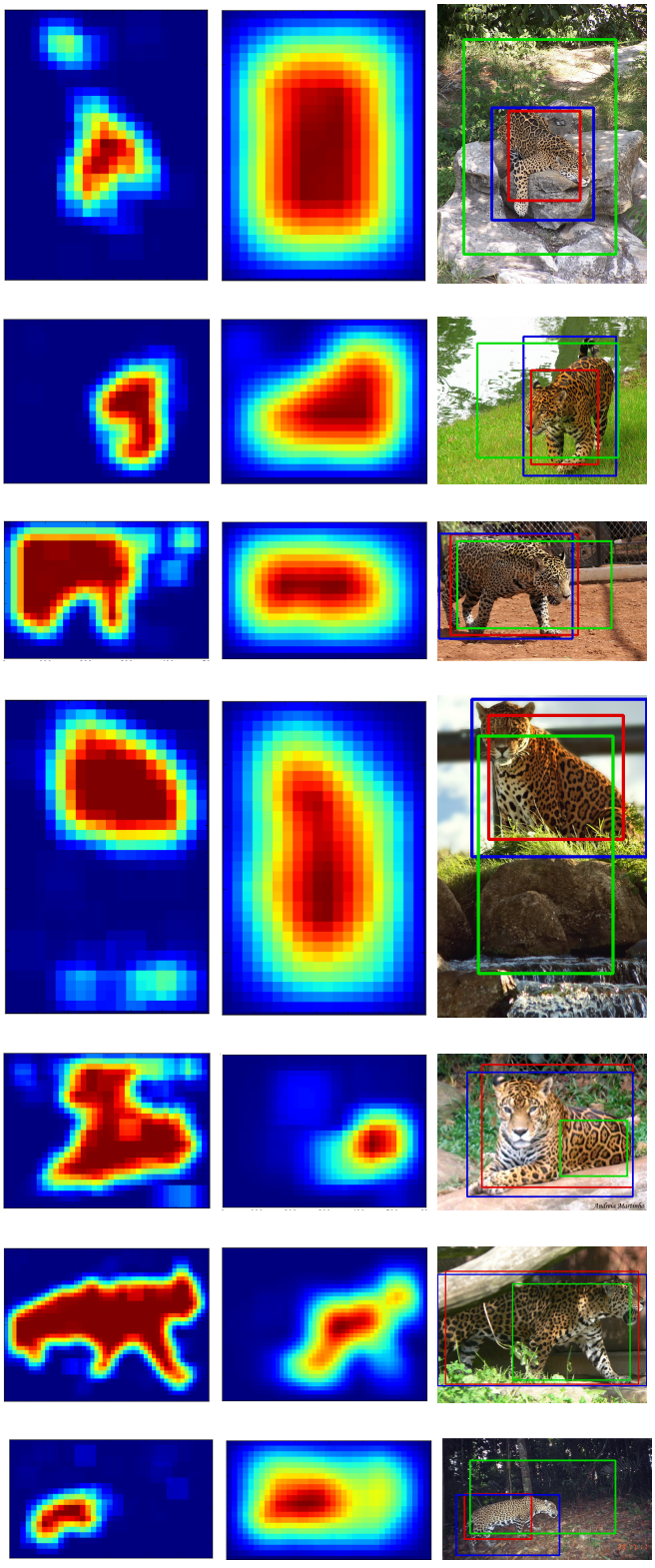
(d)

Şekil 4: İmgenin sıcaklık haritası (a), sıcaklık haritasının ikili imgeye çevrilmiş hali (b), açma (opening) ve kapama (closing) işleminin sonucu (c), tahmin edilen nesne kutusu (d).

TABLO I: Parçalar ile eğitilmiş modelin başarım tablosu.

Örtüşme Oranları	Eşik Değerleri				
	0.4	0.5	0.6	0.7	0.8
0.5	0.97	0.94	0.94	0.85	0.74
0.6	0.92	0.87	0.83	0.70	0.47
0.7	0.76	0.66	0.56	0.41	0.18

tüm imge ile eğitilen modele göre üstün bir performans gösterdiği gözlenmiştir. Konum bilgisi bulunmak istenen nesne eğer belirgin bir desene sahip ise, önerdiğimiz parça tabanlı eğitim yaklaşımının nesne konumlandırmada avantaj getireceği ortaya çıkmıştır. Gelecekte, veri kümemizi genişleterek, belirgin desene sahip farklı nesneler üzerinde çok sınıflı bir konumlandırma çalışması yapmayı düşünmekteyiz.



Şekil 5: Birinci sütundaki imgeler: parça tabanlı eğitim yöntemi ile eğitilmiş modelin sıcaklık haritalarını belirtir, ikinci sütundaki imgeler: tüm imge ile eğitilmiş modelin sıcak haritalarıdır, üçüncü sütunda yer alan mavi kutular: nesnelerin gerçek kapsan kutuları, kırmızı kutular: parça tabanlı eğitim yöntemi ile eğitilmiş modelin tahmin ettiği kutular, yeşil ise tüm ile eğitilmiş modelin tahmin ettiği kutulardır.

TABLO II: TÜM İMGE İLE EĞİTİLMİŞ MODELİN BAŞARIM TABLOSU. DÖRT FARKLI BÜYÜKLÜKTEKİ PENCERE İLE NESNE ARANMIŞTIR (64X64, 96X96, 128X128, 160X160).

Örtüşme Oranları	Eşik Değerleri				
	0.4	0.5	0.6	0.7	0.8
0.5	0.83	0.77	0.58	0.41	0.19
0.6	0.60	0.51	0.37	0.23	0.03
0.7	0.23	0.29	0.18	0.06	0.0

BİLGİLENDİRME

Bu çalışma TÜBİTAK ARDEB 115E918 numaralı proje kapsamında desteklenmiştir.

KAYNAKLAR

[1] Csurka, G., Dance, C., Fan, L., Willamowski, J. and Bray, C., May., "Visual categorization with bags of keypoints". In Workshop on statistical learning in computer vision, ECCV (Vol. 1, No. 1-22, pp. 1-2), 2004.

[2] Dalal, N. and Triggs, B., June, "Histograms of oriented gradients for human detection". In Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on (Vol. 1, pp. 886-893). IEEE, 2005.

[3] Felzenszwalb, P., McAllester, D. and Ramanan, D., "A discriminatively trained, multiscale, deformable part model". IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008.

[4] McCulloch, W.S. and Pitts, W., "A logical calculus of the ideas immanent in nervous activity". The bulletin of mathematical biophysics, 5(4), pp.115-133, 1943.

[5] Minsky, M. and Papert, S., "Perceptrons: An introduction to Computational Geometry", M.I.T. Press, Cambridge, Mass., 1969.

[6] Krizhevsky, A., Sutskever, I. and Hinton, G.E., "Imagenet classification with deep convolutional neural networks". In Advances in neural information processing systems (pp. 1097-1105), 2012.

[7] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. and Berg, A.C., "Imagenet large scale visual recognition challenge". International Journal of Computer Vision, 115(3), pp.211-252, 2015.

[8] He, K., Zhang, X., Ren, S. and Sun, J., "Deep residual learning for image recognition". In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 770-778), 2016.

[9] Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R. and LeCun, Y., "Overfeat: Integrated recognition, localization and detection using convolutional networks". arXiv preprint arXiv:1312.6229, 2013.

[10] Ren, S., He, K., Girshick, R. and Sun, J., "Faster r-cnn: Towards real-time object detection with region proposal networks". In Advances in neural information processing systems (pp. 91-99), 2015.

[11] Oquab, M., Bottou, L., Laptev, I. and Sivic, J., "Is object localization for free?-weakly-supervised learning with convolutional neural networks". In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 685-694), 2015.

[12] Oquab, M., Bottou, L., Laptev, I. and Sivic, J., "Learning and transferring mid-level image representations using convolutional neural networks". In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1717-1724), 2014.

[13] He, K., Zhang, X., Ren, S. and Sun, J. October, "Identity mappings in deep residual networks". In European Conference on Computer Vision (pp. 630-645). Springer International Publishing, 2016.