Türk Sofrası: Yemek Tanıma için Türk Yemekleri Barındıran Bir Denektaşı Veri Kümesi
Turkish Cuisine: A Benchmark Dataset with Turkish Meals for Food Recognition

Cem Güngör, Fatih Baltacı, Aykut Erdem, Erkut Erdem
Bilgisayar Mühendisligi Bölümü, Hacettepe Üniversitesi, Ankara, Türkiye
{cem.gungor,fatih.baltaci}@hacettepe.edu.tr, {aykut,erkut}@cs.hacettepe.edu.tr

***Özetçe* —Görüntülerde yemek tanıma, bilgisayarlı görüde son yıllarda çalışılmaya başlanan bir problemdir. Yemek tanıma yöntemlerinin eğitiminde ve başarımlarının değerlendirilmesinde kullanılagelen denektaşı veri kümeleri dünyaca bilinen yemeklerin örnek görüntülerini içermektedir. Ancak bu veri kümeleri detaylı incelendiğinde; bunların çok azında Türk yemeklerinin olduğu görülmektedir. Bu çalışmada öncelikle Türk yemeklerini kapsayan bir veri toplama işi gerçekleştirilmiş ve "TürkSofrası-15" adını verdiğimiz, her yemek sınıfında 500 adet görüntü bulunan yeni bir veri kümesi oluşturulmuştur. Buna ek olarak yemek tanıma için Google Inception v3 derin sinir ağ modelini esas alan ve öğrenme transferi tekniği ile eğitilen yeni bir yaklaşım önerilmiştir. Bu amaçla Türk Sofrası veri kümesi literatürde çokça kullanılan Food-101 veri kümesi ile birleştirilmiş ve geliştirilen derin öğrenme tabanlı yaklaşımın başarım analizi 113 yemek sınıfı içeren bu bütünleşik veri kümesi üzerinde gerçekleştirilmiştir. Elde ettiğimiz sonuçlar, Türk yemeklerini tanımanın belirli zorluklar içermesine rağmen belli bir başarı ile gerçekleştirilebileceğini göstermektedir.**

***Anahtar Kelimeler*—derin öğrenme, görüntü tanıma, veri seti toplama**

***Abstract*—Food recognition in still images is a problem that has been recently introduced in computer vision. The benchmark data sets used in training and evaluation of food recognition methods contain sample images of popular foods from the globe. However, when they are examined thoroughly, it can be observed that very few of them are Turkish dishes. In this study, we first carry out a data collection process for Turkish dishes and construct a new dataset named "TurkishFoods-15" containing 500 images in each food class. In addition, we introduce a novel food recognition approach that depends on fine-tuning Google Inception v3 deep neural network model based on transfer learning. For this purpose, our Turkish cuisine dataset was combined with the widely used Food-101 dataset from the literature and the performance analysis of the developed deep learning-based approach is carried out on this combined dataset containing 113 food classes. Our results show that the recognition of Turkish dishes can be achieved with certain success even though it does not have certain difficulties.**

***Keywords*—deep learning, image recognition, collecting dataset**

978-1-5090-6494-6/17/$31.00 ©2017 IEEE

I. GİRİŞ

Derin öğrenme olarak adlandırılan yapay nöral ağlar tabanlı yeni nesil öğrenme yaklaşımları, başta bilgisayarlı görü ve doğal dil işleme olmak üzere birçok alanda önemli ilerlemelerin olmasına yol açmıştır. Teorik gelişmeler yanında bu başarının altında iki önemli etken daha yatmaktadır; bunlar İnternet kullanımının artması ile birlikte büyük veri kümelerinin daha kolay oluşturulabilir hale gelmesi ve hesaplamalarda kullanılan grafik kartlarının başarımlarının artmasına ek olarak fiyatlarının ciddi oranda düşmesidir. Bu bildiride bilgisayarlı görü literatüründe çalışılan bir konu olan yemek tanıma problemi ele alınmaktadır [1]–[6]. Yemek tanıma amacıyla oluşturulmuş veri kümelerine bakıldığında bunların neredeyse hiçbirinde Türk yemeklerinin bulunmadığı görülmektedir. Bu çalışmamızda, bilinen Türk yemeklerinin görüntülerinden oluşan bir veri kümesi ile bu eksiğin giderilmesi amaçlanmıştır. Ayrıca bu veri kümesi kullanılarak derin öğrenme teknikleri ile Türk yemeklerini diğer yemek türleri ile birlikte yüksek doğruluk oranıyla sınıflandırılması adına bir ön çalışma da yer verilmektedir.

Derin öğrenmenin bilgisayarlı görü açısından beraberinde getirdiği önemli avantajlardan birisi evrişimsel sinir ağları (*convolutional neural networks*) ile ImageNet [7] gibi çok büyük veri kümeleri üzerinde görüntü sınıflandırma amacıyla öğrenilen özniteliklerin diğer problemlerin etkin bir şekilde çözümünde kolaylıkla uyarlanabiliyor oluşudur. "Öğrenme transferi" (*transfer learning*) diye adlandırılan bu yaklaşımda normalde büyük ölçekte veri gerektiren, öğrenimi uzun süre ve güçlü donanım gerektiren problemler daha az işlem gücü ve kısa zamanda çözülebilmektedir. Bu çalışmamız kapsamında oluşturulan veri kümesi üzerinde test etmek istediğimiz noktalardan birisini de bu oluşturmaktadır. Çalışma sonucunda, diğer veri kümeleri ile de birleştirilmesi ve çeşitli veri büyütme teknikleri kullanılarak 113.000 yemek görüntüsünden oluşan bir veri kümesi elde edilmiş ve bu veri kümesi üzerinde yemek sınıflandırma yapan bir derin ağ modeli eğitilmiştir. Şekil 1'de bu model ile edilen örnek bir sınıflandırma sonucu gösterilmektedir.

II. İLGİLİ ÇALIŞMALAR

Görüntülerden yemek tanıma üzerine yapılan ilk çalışmalar elle oluşturulan görsel öznitelikler kullanmış ve yemek sınıflandırmayı bu öznitelikler üzerinden gerçekleştirmiştir [1], [2],

Yaprak sarma: 0,564
Pekin ördeği: 0,384
Biftek: 0,014
Baklava: 0,011

Şekil 1: Örnek bir sınıflandırma sonucu. İlk dört olası tahmin gösterilmiştir.

[4]. Örneğin, Yang vd. [1], yerel özniteliklere dayalı olarak yemek malzemelerinin ikili ilişkilerini kodlayan bir temsil önermiş ve sınıflandırma için destek vektör makinesi (*support vector machine (SVM)*) kullanmıştır. Bir diğer çalışmada, Bossard vd. [2] rastgele orman (*random forest*) yöntemi ile ayırt edici görüntü parçalarını belirlemeye dayalı bir sınıflandırma yaklaşımı ortaya koymuşlardır. Beijbom vd. [4] tanıma başarısını arttırmak için görüntülerin konum bilgisini kullanarak ilgili restoranların yemek menülerini işlemlere dahil etmeyi düşünmüştür.

Derin evrişimsel sinir ağların görüntü sınıflandırma ve nesne tanımadaki başarısıyla beraber bu tür derin öğrenme yaklaşımları günümüzde artık yemek tanıma için de başarıyla uygulanmaktadır. Yakın tarihli bir çalışmada, Myers vd. [5] derin ağlar ile görüntülerden yemek tanımanın ötesinde ilgili yiyeceklerin kalorisinin de tahmin edileceğini ortaya koymuştur. Bir diğer çalışmada [6], yazarlar ön eğitilmiş bir ağ modeli ile sosyal medya üzerinden paylaşılan görüntülerin bir yemeğe ait olup olmadığını belirlemeye çalışmışlardır.

Yukarıda sıralanan çalışmalara ek olarak Kawano ve Yanai [3], varolan yemek sınıflandırıcılarını farklı dünya mutfaklarından yemekleri tanımak için hangi şekilde uyarlanabileceğini araştırmış ve bu amaçla SVM tabanlı bir alan uyarlama (*domain adaptation*) yaklaşımı önermiştir. Bir başka ilgili çalışmada, Malmaud vd. [8] yemek tarifleri ile yemek yapma videoları arasındaki eşlemeleri hesaplayan bir yaklaşım önermiştir.

III. VERİ KÜMESİ

Verilen bir yemek görüntüsünden yemek tanıma işlemini gerçekleştirecek derin ağ modelinin eğitilmesi için literatürde mevcut olan veri kümelerinden yararlanılmış ve bu veri kümeleri ünlü Türk yemeklerinden örnekler içeren kendi oluşturduğumuz veri kümesi kullanılarak zenginleştirilmiştir (bkz. Şekil 2). Aşağıda bu hazır veri kümelerine ve kendimizin topladığı ve Türk Sofrası adını verdiğimiz veri kümesine özel bilgiler özetlenmektedir. Literatürde mevcut veri kümelerinin çalışmamız kapsamında kullanılmasının ana nedeni geliştirdiğimiz derin öğrenme tabanlı ve öğrenme transferi stratejisi kullanan yaklaşımımızın daha büyük boyutlu bir veri kümesi üzerinde test etme isteğidir.

A. *Türk Sofrası*

Türk Sofrası adını verdiğimiz veri kümesi için ünlü Türk yemeklerinden örnek görüntüler içeren bir veri kümesi toplanmıştır[1]. Bu veri kümesi için ilgili görüntüler genellikle Google

[1]https://vision.cs.hacettepe.edu.tr/data.php

Şekil 2: Önerilen bütünleşik veri kümesi. Her yemek sınıfı için temsili bir görüntü gösterilmektedir.

TABLO I: Yemek veri kümeleri ile ilgili istatistiksel bilgiler.

Veri kümesi	Yemek sınıf sayısı	Her sınıftaki görüntü sayısı	Kaynak
Food-101	97	1000	foodspotting.com
Türk Sofrası	15	~500	Google Görseller
MMSPG	1	1000	*belirtilmemiş*

Görsellerden alınmıştır. Bu görüntüleri elde etmek için baştan belirlediğimiz sorgulamalar gerçekleştirilmiş ve sorgulamalardan dönen tüm görüntülerin indirebilmiştir. İndirilme işlemi bittikten sonra bu görüntülerin üzerinden elle geçerek alakasız olan görüntüler belirlenip silinmiştir. Bu işlem sonucunda 15 farklı Türk yemeği için, her sınıftan yaklaşık 500 görüntü elde edilmiştir.

B. *Food-101 [2]*

Yemek tanıma için önerilen ilk veri kümelerinden biri olan bu veri kümesi Bossard vd. tarafından oluşturulmuş olup toplam 101 yemek sınıfından örnek görüntüler içermektedir [2]. Her sınıfta toplamda 1.000 tane görüntü bulunmaktadır ve bu görüntüler foodspotting.com adresinden toplanmıştır.

C. *MMSPG Food Image Dataset*

Bu veri kümesi yemek sınıflandırma amacıyla önerilmiş olan Food-5K [6] ve Food-11 [9] veri kümelerinin bütünleştirilmesiyle oluşturulmuştur. Food-5K'da toplamda 5.000 adet (2.500 yemek ve 2.500 yemek olmayan) görüntü bulunmaktadır. Bu nedenle bu veri kümesi literatürde sadece verilen bir görüntünün yemek görüntüsü olup olmadığı tespitinde kullanılmaktadır. Food-11 ise 16.643 görüntüden oluşmaktadır ve bu görüntüler 11 ana yemek kategorisine bölünmüştür. Bunlar: Ekmek, süt ürünleri, tatlı, yumurta, kızarmış yiyecek, et, makarna, pirinç, su ürünleri, çorba, sebze/meyve'dir.

Deneylerimizde yukarıda sıralan üç farklı yemek veri kümesi bütünleştirilerek kullanılmıştır. MMSPG Food Image Dataset veri kümesinde mevcut olan yemek sınıfları yemek isimlerinden çok yemek kategorileri olduğundan bu veri kümesinden sadece 1 adet yemek sınıf alınmıştır. Food-101 veri kümesinden de Türk yemek kültürüne yakın olduğunu tespit ettiğimiz 97 yemek sınıfı alınmıştır. Bu sınıflara Türk Sofrası veri kümesi dahil edildiğinde; sonuç olarak toplamda 113 sınıflık bir veri kümesi elde edilmiştir. Bunlardan 16 tanesi en meşhur Türk yemeklerinden olan biber dolması, börek, baklava, çiğ köfte, enginar, hamsi, hünkar beğendi, içli köfte, ıspanak, kebap, kısır, kuru fasulye, lokum, mantı, simit ve yaprak sarmadır. Bu yemeklerden 15 tanesi hazırlanan Türk

Sofrası veri kümesine, 1 tanesi (baklava sınıfı) Food-101 veri kümesine aittir. Tüm dağılım Tablo 1'de görülebilir.

IV. YÖNTEM

A. *Eğitim*

Derin öğrenmede kullanılan ağlar, belli bir hiyerarşi içinde tanımlı çok katmanlı yapıdadır. İlk katmanlarda verinin daha basit öznitelikleri öğrenilirken, ileri katmanlarda bu basit öznitelikler birleştirilerek daha karmaşık ve anlamsal bilgiyi kodlayan öznitelikler öğrenilmiş olmaktadır. Örneğin; yüz tanıma probleminde ilk katmanlarda görüntüdeki kenarlar öğrenilirken; ileriki katmanlarda kenarların birleşimlerinden oluşan yüz parçaları (göz, burun vb.) ve daha da ilerideki katmanlarda bunların birleşiminden oluşan yüz öğrenilmiş olmaktadır.

Derin öğrenme güçlü bir öğrenme yaklaşımı olmasına rağmen sıfırdan derin bir ağ modelini eğitmek çok fazla işlem gücü ve zaman gerektirmektedir. Ancak öğrenme transferi *(transfer learning)* uygulanarak bu problemin üstesinden gelmek mümkündür. Öğrenme transferinde; önceden eğitilmiş bir derin öğrenme modeli alınmakta ve bu modelin ilk katmanların kodladıkları basit öznitelik çıkarıcıların (kenarlar, şekiller vb.) yeterince öğrenildiği ve bu özniteliklerin çoğunun farklı veri kümeleri için ortak olduğu kabul edilerek sadece son katmanlarda ele alınan problem için yeni bir veri kümesiyle eğitim gerçekleştirilmektedir. Bu sayede zamandan ve işlem gücünden büyük oranda tasarruf edilmiş olmaktadır.

Öğrenme transferini uygulamak amacıyla deneylerde Google tarafından geliştirilmiş ve eğitilmiş olan Inception v3 [10] derin evrişimsel sinir ağı modeli kullanılmıştır. Bu model ImageNet [7] ile 1.000 sınıftan oluşan 1,2 milyon görüntü ile eğitilmiştir. Bu modelin seçilmesinin nedeni bu modelin ILSVRC2012 (Image-Net Large Scale Visual Recognition Challenge 2012)'de en düşük hata oranına sahip olmasıdır. Çalışmada önceki katmanlarının ağırlıkları sabit tutularak, sadece son katmana ince ayar *(fine-tuning)* yapılmıştır.

Eğitim esnasında veri kümesinde her sınıf için veri arttırma (*data augmentation*) yöntemleri kullanılarak örnek sayısı arttırılmıştır. Özellikle belirtmek gerekirse; görüntünün aynasını alma (*image mirroring*), parlaklığını artırıp azaltma, gürültü ekleme gibi görüntü işleme teknikleri (bkz. Şekil 3) kullanarak her sınıfta toplamda 1000 adet görüntü olacak şekilde veri kümesi büyütülmüştür. Böylece eğitimde toplamda 113.000 görüntü kullanılmıştır.

B. *Uygulama Detayları*

Literatürde kullanılagelen farklı derin öğrenme kütüphaneleri bulunmaktadır. Bu çalışmada Google Brain Team tarafından geliştirilmiş açık kaynaklı bir kütüphane olan Tensorflow [11] kullanılmıştır. Eğitim ve test işlemleri Nvidia GTX960M GPU kartı üzerinde gerçekleştirilmiştir.

Derin öğrenmede her katmanın çıktısı kendisinden sonraki katmanın girdisi şeklindedir. Model eğitilirken sadece son katman eğitildiği için öncelikle görüntüler modelde son katmana kadar işlenerek son katmanın girdilerini oluşturan filtre çıktıları bulunmakta ve bunlar global görüntü özniteliği olarak saklanmaktadır. Her görüntü için bu veri bulunduktan sonra bunlar kullanılarak son katmandaki ağırlıklar eğitim süresince öğrenilmektedir. Veri kümesindeki görüntülerin son katmana

Şekil 3: Kullanılan veri büyütme teknikleri.

TABLO II: Sayısal sonuçlar.

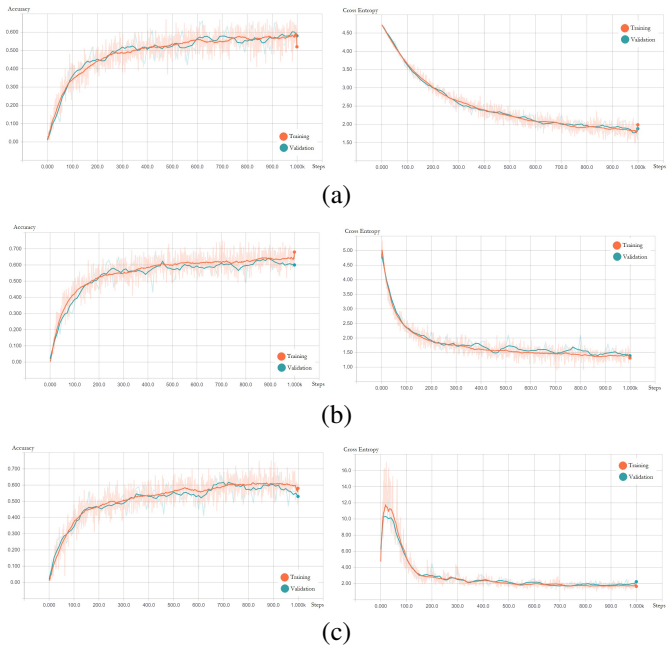
Öğrenme oranı	Test doğruluk oranı
0,05	%57,6
0,30	%62,7
0,70	%55,7

kadar işlenerek bu özniteliklerin çıkartılması ve saklanması 2 saat 24 dakika sürmüştür. 1.000 adım kullanılarak gerçekleştirilen ön eğitim, dosya tarama işlemi bittikten sonra 3 dakika 34 saniye sürmektedir. 10.000 adımla olan son eğitim işlemi ise dosya tarama işlemi hariç tutularak 25 dakika 24 saniye sürmüştür. Tüm testlerde veri kümesinin %10'luk bir kısmı test için, %10'luk bir kısmı doğrulama *(validation)* için rastgele düzende ayrılmıştır ve veri kümesindeki tüm görüntüler 299×299 piksel boyutuna getirilmiştir.

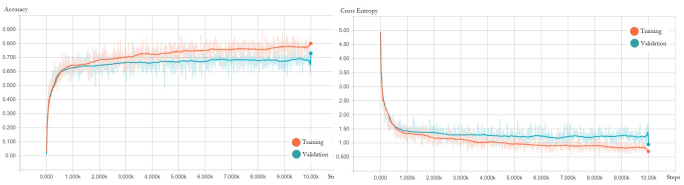
V. DENEYSEL SONUÇLAR

Model farklı parametreler ile eğitilerek test edilmiştir. İlk olarak en iyi öğrenme oranını bulabilmek amacıyla model farklı öğrenme oranları *(learning rate)* ile bir öneğitim gerçekleştirilmiştir. Eğitimin kısa tutulması amacıyla adım sayısı *(step size)* olarak 1.000 kullanılmıştır. Parti büyüklüğü *(batch)* ise 100 seçilmiştir. Farklı öğrenme oranlarının doğruluk ve cross entropy hata grafikleri Şekil 4'te verilmiştir. Bu öğrenme oranlarıyla elde edilen doğruluk oranları da Tablo 2'de gösterilmektedir. Bunlardan görülebileceği üzere 0.1'lik bir öğrenme oranı ile %57,6'lik bir test doğruluk oranı yakalanmıştır. Öğrenme oranı 0,3'e yükseltildiğinde ise eğitim hızı artmış ve test doğruluk oranı %62,7'ye yükselmiştir. Öğrenme oranın 0,7'ye yükseltildiğinde ise test doğruluk oranı %62,7'den %55,7'e düşmüştür.

Bu sonuçlardan yola çıkarak son deneyde öğrenme oranı 0,3 olarak alınmış ve parti büyüklüğü (100) aynı tutularak adım sayısı 10.000 alınarak bir model eğitilmiştir. Bu model ile elde edilen test doğruluk oranı %68,2 olarak hesaplanmıştır. Doğruluk ve çapraz entropi hata grafikleri Şekil 5'te görülebilir. Şekil 6'da bir kaç örnek yemek görüntüsü içim geliştirilen yöntemle elde edilen ilgili tahminler yer almaktadır. Sağ üstte yer alan görüntüdeki hamsi yemeği düzgün tespit edilebilmiştir. Sol üst görüntü için en olası yemek sınıfı çiğ köfte çıkmışken, olasılık bakımından ikinci sırada ise salata bulunmuştur. Bunun sebebi, görüntüde hem çiğ köftenin hem de salatanın bulunmasıdır. Sonuçlarda sağ altta yer alan görüntüdeki simit doğru tahmin edilmesine rağmen, sol alttaki görüntüde kuru fasulyenin doğru tahmin edilemediği gözlenmiştir. Birinci sırada bir İtalyan yemeği olan gnochi, ikinci sırada ise kuru fasulye bulunmaktadır.



Şekil 4: Farklı öğrenme oranları için doğruluk ve çapraz entropi hata grafikleri. (a) η =0,05 (b) η =0,30 (c) η =0,70.



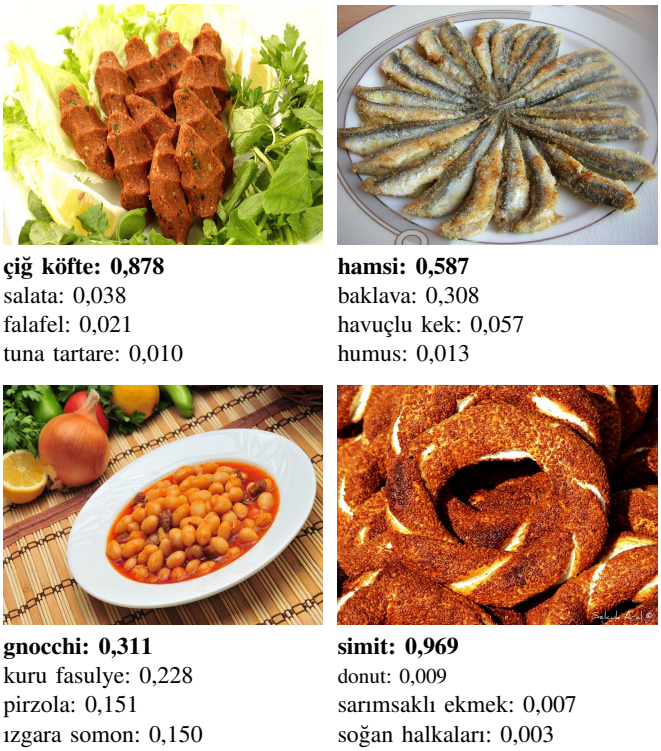
Şekil 5: Adım sayısı=10.000 ve $\eta = 0,30$ için (a) doğruluk grafiği, (b) çapraz entropi hata grafiği.

Tahminlerin ne kadar güvenilir olduğu rapor edilen olasılık değerlerine bakılarak anlaşılabilmektedir.

Test için ayırmış olduğumuz veri kümesiyle yaptığımız denemelerde elde ettiğimiz sonuçlar bize bazı sınıfların diğerlerinden daha iyi ayrıldığını göstermiştir. Örneğin, öğrenilen model simit görüntüsünü bazen donut olarak algılayabilmektedir. Bu görece beklenen bir sonuçtur çünkü ikisinin de şekli birbirine benzemektedir. Bunu aşmak için öğrenme transferi kullanarak sadece son katmanı eğitmek yerine, daha önceki katmanların da eğitilmesinin gerektiği düşünülmektedir. Fakat daha önceki katmanlardan eğitmek daha önce de belirtildiği üzere daha fazla bellek, işlem gücü ve zaman isteyecektir.

VI. SONUÇ

Bu çalışmamızın sonucunda görüntülerde Türk yemeklerinin tanınması için TürkSofrası-15 isimli onbeş ünlü Türk yemeklerinin görüntülerinden oluşan bir veri kümesi oluşturulmuştur. Bu veri kümesi literatürde halihazırda mevcut olan yemek veri kümeleri ile bütünleştirilmiş ve bu geniş veri üzerinde derin evrişimsel sinir ağlara dayalı bir yemek tanıma sistemi geliştirilmiştir. Veri kümesi daha da büyütülerek daha büyük kapsamlı çalışmalarda kullanılabilir. Türk yemekleriyle ilgili kalori hesaplama, yemek tarifi bulma gibi problemler için de önerilen veri kümesinden faydanılabileceği düşünülmektedir.



Şekil 6: Örnek sınıflandırma sonuçları. Her bir görüntü için ilk dört olası tahmin gösterilmiştir.

KAYNAKLAR

[1] S. L. Yang, M. Chen, D. Pomerleau, and R. Sukthankar, "Food recognition using statistics of pairwise local features," in *CVPR*, 2010.

[2] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101 – mining discriminative components with random forests," in *European Conference on Computer Vision*, 2014.

[3] Y. Kawano and K. Yanai, "Automatic expansion of a food image dataset leveraging existing categories with domain adaptation," in *Proc. of ECCV Workshop on Transferring and Adapting Source Knowledge in Computer Vision (TASK-CV)*, 2014.

[4] D. M. S. S. Oscar Beijbom, Neel Joshi and S. Khullar, "Menu-match: Restaurant-specific food logging from images," in *WACV*, 2015.

[5] A. Myers, N. Johnston, V. Rathod, A. Korattikara, A. Gorban, N. Silberman, S. Guadarrama, and et al., "Im2calories: towards an automated mobile vision food diary," in *ICCV*, 2015.

[6] A. Singla, L. Yuan, and T. Ebrahimi, "Food/non-food image classification and food categorization using pre-trained googlenet model," in *Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management*, pp. 3–11, ACM, 2016.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.

[8] J. Malmaud, J. Huang, V. Rathod, N. Johnston, A. Rabinovich, and K. Murphy, "What's cookin'? interpreting cooking videos using text, speech and vision," in *NAACL*, 2015.

[9] "Food-11 dataset." grebvm2.epfl.ch/lin/food/Food-11.zip. Accessed: 2017-02-13.

[10] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," *CoRR*, vol. abs/1512.00567, 2015.

[11] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. Corrado, A. Davis, J. Dean, and et al., "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," 2015.