Evrişimsel Sinir Ağlarında Kavram Yorumlama için bir Puanlama Yöntemi
A Scoring Method for Interpretability of Concepts in Convolutional Neural Networks

2022 30th Signal Processing and Communications Applications Conference (SIU) | 978-1-6654-5092-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/SIU55565.2022.9864930

Mustafa Kağan Gürkan, Nafiz Arıca
Bahçeşehir Üniversitesi
İstanbul, Türkiye
mustafakagan.gurkan@bahcesehir.edu.tr,
nafiz.arica@eng.bau.edu.tr

Fatoş Yarman Vural
Orta Doğu Teknik Üniversitesi
Ankara, Türkiye
vural@ceng.metu.edu.tr

***Özetçe* —Bu çalışmada, evrişim katmanlarındaki öznitelik çıkarma işlemine odaklanarak CNN modellerinin yorumlanabilirliğini ölçmek için bir puanlama algoritması öneriyoruz. Önerilen yaklaşım, önceden tanımlanmış bir liste üzerinden kavram analizi ilkesine dayanmaktadır. Ağın her kavrama karşı duyarlılığını ölçen bir harita oluşturulur. Bu harita hazır olduktan sonra çeşitli imgeler girdi olarak uygulanabilir ve gizli düğümleri yüksek oranda aktif olan kavramlarla eşleştirilir. Son olarak, değerlendirme algoritması, son tahmin esnasında bu açıklamaları kullanmak için devreye girer ve insan tarafından anlaşılabilir açıklamalar sağlar.**

Anahtar Kelimeler—Açıklanabilir yapay zeka, evrişimsel sinir ağları, kavram temelli analiz, çalışma anında yorumlanabilirlik

***Abstract*—In this paper, we propose a scoring algorithm for measuring the interpretability of CNN models by focusing on the feature extraction operation at the convolutional layers. The proposed approach is based on the principal of concept analysis, for a predefined list of concepts. A map of the network is created based on its responsiveness against each concept. Once this map is ready, various images can be applied as inputs and they are matched with the concepts whose hidden nodes are highly activated. Finally, the evaluation algorithm kicks in to use these descriptions during the final prediction and provides human-understandable explanations.**

Keywords—Explainable AI, convolutional neural networks, concept-based analysis, runtime interpretation

I. Giriş

Derin ağlar, düşük soyutlama seviyeli girdi ile yüksek soyutlama seviyeli çıktı arasındaki kara kutular olarak kabul edilir. Veri kümelerindeki yanlılık, eğitim ve test küme dağılımlarındaki tutarsızlık ve verilerin istatistiksel yetersizliği gibi çeşitli faktörler, modellerin yanlış sonuç vermesine neden olabilir [1-2]. Bu nedenle, modellerin adalet, güvenilirlik, açıklayıcı doğrulanabilirlik, kullanılabilirlik vb. özelliklerinin ölçülmesine ihtiyaç vardır [3]. Bu gibi durumlarda, yorumlanabilirlik, bu kriterleri ölçmek için bir ara adım olarak kullanılabilir. Verimli veri ön işleme ve model seçimi, model hakkındaki anlayışımızı artırsa da, tam olarak yorumlanabilirlik ancak çalışma zamanı işlemleri hakkında fikir edinerek sağlanabilir. Bu nedenle, eğitimden sonra modelin içsel yapısını açıklama çalışmaları büyük ilgi çekmektedir. Bu alandaki çalışmalar klasik iki ana akıma odaklanmıştır: *özelliklerin ilişkilendirilmesi* [1, 4-5] ve *özelliklerin görselleştirilmesi* [6].

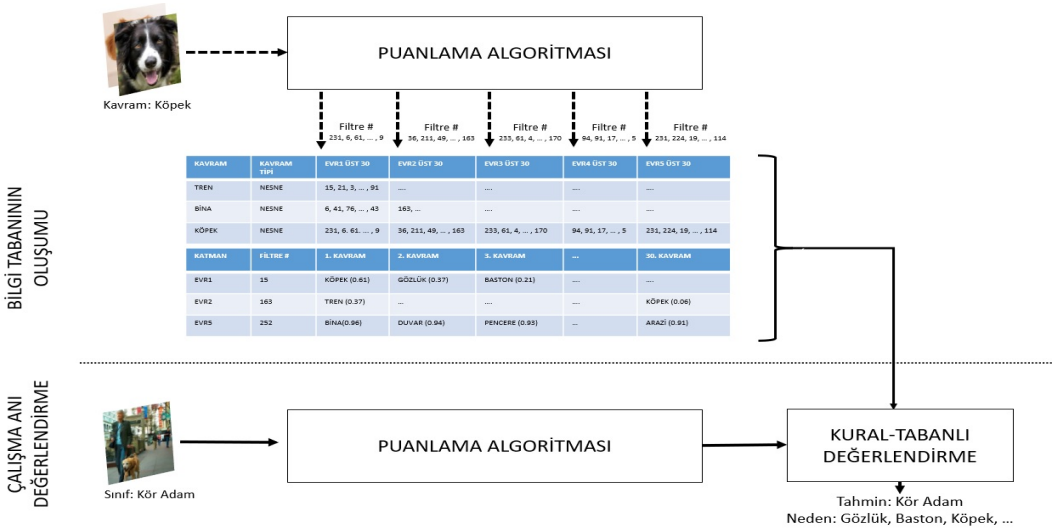
978-1-6654-5092-8/22/$31.00 ©2022 IEEE

Son zamanlarda, *kavram temelli analize* odaklanan üçüncü bir akım ortaya çıkmaktadır. Kararın nedenleri sorulduğunda, insanların açıklamaları çoğu zaman eylemler, nesneler, renk/şekil vb. gibi kavramlar üzerinden tanımlamalar olur. Kavram temelli teknikler, bireysel özelliklere veya piksellere önem vermek yerine ağın kavram öğrenme kabiliyetinin analizini hedefler. Bu yorumlanabilir kavramlar, birimler üzerine atanmış temsiller [7] ya da birden çok birime, hatta katmanlara yayılmış ilişkisel bir aktivasyon [2] olarak ortaya çıkabilir.

II. Tasarım Mimarisi ve Motivasyon

Kavram temelli analiz üzerine mevcut araştırmalar, bir modelin kavramları anlayabilme yeteneğine odaklansalar da, her filtreyi tek bir kavramla eşleştirmek [7] veya her kavram için bütün bir katman genelinde tanımlama yapmak [2] gibi bazı ciddi sınırlamalar bulunmaktadır. İlk durumda, filtre bir çok kavramı verimli bir şekilde algılayabilse dahi filtre başına bir kavram sınırlaması nedeniyle bu kavramlardan bazılarını tamamen göz ardı eder. Bütün katman tabanlı ikinci yöntemde ise, filtrelerin küçük bir kısmı çok yüksek aktivasyon değerlerine sahip olabilir, ancak geri kalanlar karışık yanıtlara sahip olduğunda, nihai sonuç en iyi ihtimalle ortalama olur. Mevcut yöntemlerin bir diğer dezavantajı, ölçümlerin yapılma şeklidir. Ana hedefleri, tahmin için bir sebep sağlamak yerine modelin davranışını anlamaktır. Bu nedenle, aktivasyon değerlerini eşikledikten sonra kesiştirilmiş bölgelerin bileşimi, IoU, (Intersection over union) metodunu uygularlar. Ne yazık ki, bu yaklaşımın iki önemli dezavantajı vardır: i) yüksek eşikler nedeniyle, etkinleştirilen parçaların çoğu tamamen göz ardı edilir, ii) IoU hesaplamalarının doğası gereği, girdi imgesinin bölütlere ayrılmış bir sürümüne ihtiyaç duyulması çalışma zamanı yorumlama istekleri için ciddi bir sorun teşkil etmektedir.

Bu makalede sunulan kavram temelli yaklaşım, tahminin arkasındaki mantığı yorumlarken bu sorunları ele almak için tasarlanmıştır (Şekil 1). Herhangi bir CNN modeli ile



Şekil 1: Yorumlanabilirlik Yönteminin Genel Mimarisi

değişiklik gerektirmeden çalışabilir. Tek gereksinim, modelin kavramlara karşı davranışını öğrenmek için bu kavramları temsil eden bölgelere ayrıştırılmış bir dizi imgenin bir kerelik işlenmesidir. Önerilen yöntem, IoU hesaplamalarının yukarıda bahsedilen eksikliklerinden dolayı, her bir *(kavram, filtre)* çiftinin ilişkilendirilebilmesi adına yeni bir puanlama algoritması tanımlamaktadır. İlk aşamada, her bir filtre ve kavram için azami puanlara bakılarak modelin bir ***Bilgi Tabanı (BT)*** oluşturulur. Girdi imgesine modelin cevabı gizli düğümleri yüksek derecede aktif olan kavramlarla eşleştirilir. Son adımda, en iyi eşleşen kavramları belirlemek için kural tabanlı değerlendirme algoritması devreye girer ve yazılı açıklamalar sağlanır.

A. Puanlama Algoritması

Yöntemin ilk adımı, ağın belirli bir katmanındaki her bir filtre ile her kavram arasındaki ilişkiyi belirlemektir. Önerilen puanlama algoritması (Şekil 2) her kavram için her filtrenin algılama kesinliği ve algılama kuvveti olmak üzere iki ana faktöre odaklanır. Bir filtrenin kesinliği, söz konusu kavramı, girdi resmin içinde diğer kavramlar arasında tanımlayabilme yeteneği olarak kabul edilir. Puanlama algoritmasının bu kısmı sadece BT oluşturulurken kullanılır. Kesinliğin ölçülebilmesi için kavramın tam konumu gereklidir. Bu nedenle, I giriş imgesi ile aynı boyutta, NxM, bölütlenmiş bir imge kullanılır. Bölütlenmiş imge SI şu şekilde tanımlanabilir;

$$SI = \bigcup R_k, \quad k = 1, ..., n \tag{1}$$

Bu denklemde R_k, k kavramını temsil eden imge bölgesini ifade eder. SI bölütlenmiş imgesi, her bir kavramı oluşturan imge bölgesi piksellerine aynı R, G, B değerleri atanan, $R_k = \bigcup p_k$, n adet imge bölgesinin birleşimi ile oluşturulur.

Bir sonraki adım, filtrenin girdi imgesine verdiği yanıtın kesinliğini belirlemektir. Bu amaçla imge ağ üzerinden işlenir ve filtrenin aktivasyon değerleri alınır. Ardından, özellik haritası, SI boyutuyla eşleşmesi için $N \times M$'e yükseltilir ve aktivasyon değerleri normalleştirilir. Normalleştirilmiş piksel aktivasyon değeri, $\hat{P}(x, y)$ aşağıda verilmiştir;

$$\hat{P}(x, y) = \frac{P(x, y)}{\sum_{x=0}^{N} \sum_{y=0}^{M} P(x, y)}. \tag{2}$$

Ardından, SI bölütlenmiş imgesini oluşturan her k kavramı için R_k imge bölgesi dahilindeki piksellerin normalleştirilmiş aktivasyon değerleri kullanılarak ***İsabet Oranı (IO)*** olarak adlandırılan kesinlik puanları tanımlanır. Bu puanların her biri [0, 1] aralığında ve toplamları 1 olacak şekilde oluşur.

$$IO_k = \sum_{\forall (x,y) \in R_k} \hat{P}(x, y), \tag{3}$$

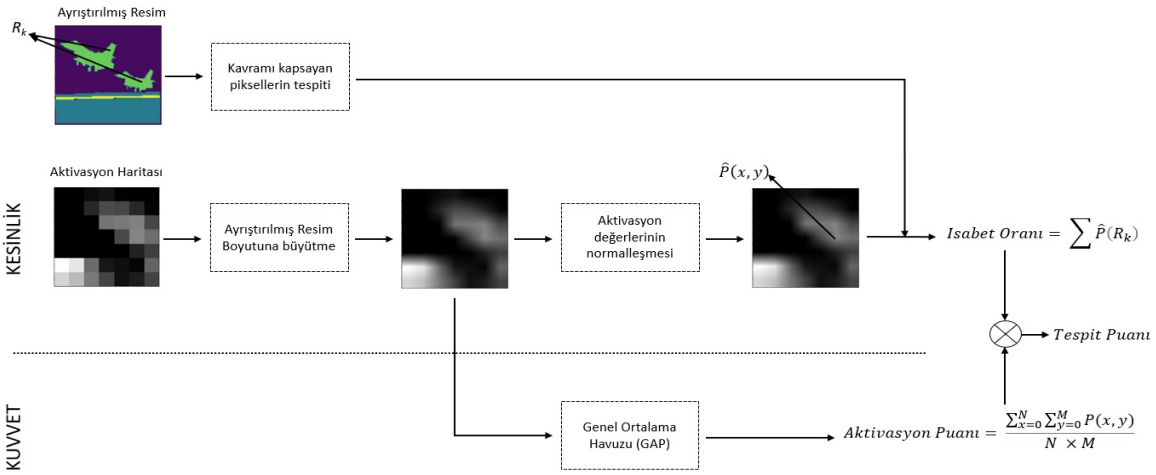
Puanlama algoritmasını oluşturan ikinci ana faktör olan algılama kuvveti, yüksek aktivasyon değerlerinin ağın kararları üzerinde daha fazla etkiye sahip olduğu temeline dayanmaktadır. Bu kuvvet ölçümlemesi için, [5]'de önerilen Genel Ortalama Havuzu (Global Average Pooling - GAP) yöntemi kullanılarak her filtreye ***Aktivasyon Puanı (AP)*** denen bir ağırlık değeri atanır. Bu ağırlık, büyütülmüş aktivasyon haritasındaki bütün piksellerin değerlerinin ortalaması alınarak hesaplanır.

$$Aktivasyon\,Puani = \frac{\sum_{x=0}^{N} \sum_{y=0}^{M} P(x, y)}{N \times M}. \tag{4}$$

Her bir *(kavram, filtre)* ikilisi için son puan olan ***Tespit Puanı (TP)*** ise $TP = IO \times AP.$ şeklinde hesaplanır.

B. Bilgi Tabanı

Önerilen yorumlanabilirlik yönteminin omurgası BT'nin hazırlanması aşamasıdır. BT'nin oluşturulması için, modelin ayrıştırılmış resimlerden oluşan bir veri kümesi üzerinden çalıştırılması gerekmektedir. Her kavram için karşılık gelen görseller, puanlama algoritması aracılığıyla birer birer işlenir. Sonunda, her bir k kavramı ve f filtresi için ortalama IO_{kf} ve AP_{kf} değerleri elde edilir. Bu değerler $\#filtre \times \#kavram$ boyutunda iki veri tablosu oluşturmak için kullanılır. Bu tablolardan ilki, filtrelerin kavram başına algılama keskinliğini



Şekil 2: Puanlama Algoritması

gösteren IO puanlarından oluşur. İkinci tablo ise ortalama TP değerlerini depolar ve aşağıdaki iki sözlüğün ve BT'nin oluşturulmasında kullanılır (Şekil 1).

- **BT1 - Kavram başına Filtreler**: Her kavram için en yüksek tespit puanına sahip Y adet filtre.
- **BT2 - Filtre başına Kavramlar**: Her filtre için en yüksek tespit puanına sahip Z adet kavram.

burada Y ve Z değerleri ayarlanabilir hiper parametrelerdir.

C. Çalışma Anı Değerlendirme

BT oluşumu tamamlandıktan sonra modelin tahminini çalışma zamanında yorumlamak mümkündür. İlk işlem puanlama algoritmasının çalıştırılması ve her *(kavram, filtre)* ikilisi için bir TP değerinin çıkarılmasıdır (Şekil 1). Bu aşamada girdi resmin ayrıştırılmış eşleniği olmadığından, hesaplamalar sırasında IO katsayıları depolanmış veri tablosundan çekilir. Sonunda, $\#filtre \times \#kavram$ boyutlu bir puanlama matrisi elde edilir. İkinci kısımda, TP değerleri kullanılarak kural tabanlı değerlendirme yöntemi uygulanarak, tahmin edilen sınıf ve tespit edilen kavramlardan oluşan yazılı bir çıktı oluşturulur. Bu kısım için küçük farkları olan beş yöntem önerilmiştir.

İlk yöntem yalnızca çalışma anından alınan değerleri kullanır. Bütün TP değerleri içinde en yüksek olanın hangi filtreyi kullandığını araştırır ve bu filtredeki en yüksek puanları veren Z adet kavramın ağırlıklarını 1 artırır. Bu işlemler toplam 20 defa tekrarlandığında [8] elde edilen ağırlıklar kavramların ne kadar yüksek yüzde ile güvenilir olduklarını verir. İkinci yöntem TP matrisinden en yüksek puanlı filtreyi arar, ancak en iyi Z kavramı belirlerken çalışma anı sonuçları yerine $BT2$ sözlüğünü kullanır. Bu kavramlara eşit ağırlıkta puan verir. Fakat gerçekte, bir filtrenin algılama kapasitesi baştan sona kadar aynı değildir ve tüm kavramlar için tek tip bir ağırlık artışı yanıltıcı olabilir. Bu nedenle, üçüncü yöntem en iyi Z kavramlarını sıralar ve ağırlıkları sıralarına göre artırır.

İlk üç yöntem, her filtre için en iyi kavramları depolayan $BT2$ sözlüğüne odaklanmaktadır. Fakat, $BT1$ sözlüğündeki her kavram için en yüksek puanlı filtreler bilgisi de aynı oranda yararlı olabilir. Nitekim, dördüncü yöntem çapraz doğrulama kullanır. En yüksek TP çalışma zamanı puanına sahip filtreye bağlı en iyi Z kavramı her zamanki gibi $BT2$'den alınır. Fakat bu sefer, $BT1$'den de kavramlara atanan en iyi Y filtre incelenir ve eşleşme olur ise kavramın ağırlığı bir artar. Beşinci ve son yöntem de, çapraz doğrulama ve tekil olmayan ağırlık artırma seçeneklerini kullandığımız üçüncü ve dördüncü yöntemlerin birleşimidir.

III. Deneyler

BT oluşturmak için çeşitli kavramları temsil eden hem normal hem de ayrılmış imgelerden oluşan Broden veri seti [2, 7] kullanılmıştır. Renkler, nesneler, nesne parçaları, dokular ve malzemeler gibi farklı kategorilerden çeşitli kavramları temsil eden 63 binden fazla bölütlenmiş imgeden oluşur. Büyük kavramların etkinleştirildikleri imgelerde çok düşük ölçeklerde temsil edilen kavramlar önemsiz kılınabilip yanlış sonuçlara yol açabilir. Bu nedenle, alanı 50 pikselden küçük olan kavramlar (bölütlenmiş imgeler 112 x 112 boyutundadır) göz ardı edilmiştir. Aynı şekilde, daha sağlam bir değerlendirme yapılabilmesi adına her kavram için en az 100 örnekleme olması gerekmektedir. Bu kısıtlamların sonunda, 32 binden fazla imge ve bilgi tabanı dağarcığını temsilen 200 uygun kavramdan oluşan bir veri seti elde edilmiştir.

Bir sonraki adım, çalışmak için bir model seçilmesidir. Bu çalışmada, Places365 veri kümesine [7] karşı eğitilmiş ResNet18 modeli analiz için tercih edilmiştir. Tüm uygun kavramlar için Broden veri kümesindeki tüm örnekler uygulandıktan ve ortalama değerler alındıktan sonra, 512 x 200 boyutunda IO ve TP veri tabloları elde edilmiştir (burada 512, ResNet18'in son evrişim katmanındaki filtre sayısıdır). Buna müteakip, hem Y hem de Z değerleri 30 olarak ayarlanarak BT oluşturulmuştur. Başka bir deyişle, her filtre için en çok etkinleştirilen 30 kavram ve her kavram için en çok etkinleştirilen 30 filtre tanımlanmıştır. Son aşama, çalışma anı yorumlarını tetiklemektir. Şekil 3'te verilen iki resim, test örnekleri olarak seçilmiş ve önerilen beş değerlendirme yönteminin performansı, modelin hangi kavramlara güvendiğine dayanan raporlarla ölçülmüştür.

İlk basit yaklaşımda, odak modelin genel olarak en iyiyi belirlemek için eğitildiği kavramlara verilmesinden dolayı çok

TABLO I: Test Örneklerinin Güven Puanları: (a) Hava Sahası (b) Yaya Geçidi

Kavram	2. Yöntem		3. Yöntem		4. Yöntem		5. Yöntem	
	Güven	Sıralı Liste	Güven	Sıralı Liste	Güven	Sıralı Liste	Güven	Sıralı Liste
Uçak	100%	1.	87.17%	1.	65%	1.	63.8%	1.
Gökyüzü	85%	2.	44.83%	10.	30%	3.	26.33%	4.
Uçak	75%	10.	51.83%	6.	50%	2.	44.83%	2.

(a)

	Güven	Sıralı Liste	Güven	Sıralı Liste	Güven	Sıralı Liste	Güven	Sıralı Liste
Otobüs	100%	1.	90.33%	1.	40%	4.	40%	4.
Yol	95%	3.	66.83%	4.	45%	3.	41%	2.
Çimen	60%	17.	28.67%	19.	15%	25.	8.5%	28.
Gökyüzü	55%	20.	25%	21.	20%	10.	15.5%	12.
Ağaç	65%	13.	24%	22.	20%	10.	15.33%	13.
Araba	80%	6.	50.67%	6.	50%	1.	41.5%	1.
Kamyon	35%	34.	9.67%	45.	30%	6.	21.83%	8.
Pist	35%	34.	19.67%	31.	20%	10.	16.17%	11.
Otobüs	75%	8.	41.5%	10.	50%	1.	40.5%	3.

(b)

yüksek güven puanlarıyla birçok yanlış pozitif rapor edilmiştir. Bu kötü sonuçlar, çalışma anı etkinleştirme değerlerinin bir modelin kararını yorumlamada tek başına yeterli olmadığını ortaya koymuş ve daha iyi sonuçlar için BT kullanımını haklı çıkarmıştır. BT temelli kontrolü kullanan ikinci yöntem ile raporlarda listelenen kavramların sayısı artmış, ancak aynı zamanda yanlış pozitiflerin güven düzeyi de önemli ölçüde azalmıştır. Üçüncü yöntem, güven seviyelerini daha da düşürüp, yanlış pozitiflerin çoğunu ihmal edilebilir hale getirmiştir.

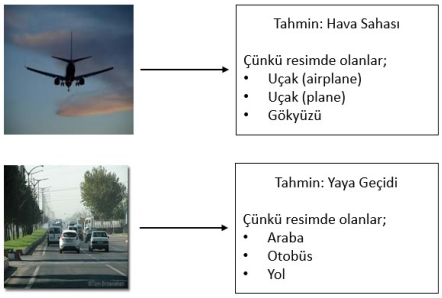
İkinci ve üçüncü yöntemler, güven puanlarını düşürerek kavramlar arasındaki önem derecesini ayırt edilebilir kılmıştır. Ancak, yüksek sayıda kavramın (yüze yakın) rapor edilmesine engel olamamıştır. Dördüncü yöntemin çapraz doğrulama yeteneği bu durumda yardımcı olabilir. Sonuçlarda bu tahmini desteklemekte, bu yaklaşımla yanlış pozitiflerin üçte birinden fazlasının ortadan kaldırıldığını göstermektedir. Paralel olarak, güven puanları da düşmüş, ancak doğru kavramlar kesinlik listesinde daha yüksek sıralarda rapor edilir hale gelmiştir. Beşinci yöntem ise güven seviyelerini daha da düşürmekte ancak bir çok durumda sıralı listeleri kötüleştirmektedir.

Tablo I, örnek görüntülerdeki kavramları ve her bir değerlendirme yöntemi tarafından rapor edilen güven puanlarını listelemektedir (2 adet uçak ve 2 adet otobüs girişinin nedeni, veri setinde bus-autobus ve airplane-plane şeklinde ayrımlar olmasındandır). Çok kötü performans nedeniyle ilk yöntem göz ardı edilmiş, sonuçlar tabloya yansıtılmamıştır. Sonuçlar dördüncü yöntemin diğerlerine göre daha iyi performans gösterdiğini ve gerçek kavramların çoğunu doğru bir güven sırası ile tanımlayabildiğini göstermektedir. Dördüncü yöntem baz alınarak çıkarılan ilk 3 kavrama dayalı yorumlanabilirlik raporu da Şekil 3'te verilmiştir.

IV. Sonuç

Bu çalışmada, CNN modellerinin açıklanabilirliği üzerine yeni bir yaklaşım önerilmiştir. Bir sahneyi bir dizi nesneyle tanımlamanın insani yorumlama yöntemi benimsenmiş ve model, kavram temelli analiz üzerine inşa edilmiştir. Elde edilen ilk sonuçlarda, imgedeki kavramların çoğu başarılı bir şekilde tanımlandığından ve model tahminlerinin neden yapıldığına dair bazı gerekçeler sunabildiğinden dolayı umut vericidir.

Gelecek çalışmalarda BT'nin çeşitli alanlarda performansını değerlendirmek için bu yaklaşımın daha geniş bir görüntü



Şekil 3: Çalışma Anı Yorumlama Raporu (İlk 3 kavram)

kümesine uygulanması planlanmaktadır. Ayrıca, BT sözlük boyutunu ve açıklama kapsamını artırmak için daha fazla nesne, nesnelerin bir kısmı veya diğer kavram türleri ile güncellenecektir. Son olarak, bu yaklaşım farklı CNN modellerine de uygulanarak modellerin açıklanabilirliklerinin karşılaştırması yapılacaktır.

Kaynaklar

[1] R.R. Selvaraju, M. Cogswall, A. Das, R. Vedantam, D. Parikh, D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," 2019, arXiv:1610.02391.

[2] B. Kim et al., "Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)," 2018, arXiv: 1711.11279.

[3] F. Doshi-Velez, B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning," 2017, arXiv:1702.08608.

[4] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek, "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation," PloS ONE, vol. 10, no. 7, Jul. 2015.

[5] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva and A. Torralba, "Learning Deep Features for Discriminative Localization," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, pp. 2921-2929, 2016

[6] D. Ulyanov, A. Vedaldi, V. Lempitsky, "Deep Image Prior," 2018, arXiv:1711.10925

[7] D. Bau, B. Zhou, A. Khosla, A. Oliva, A. Torralba, "Network Dissection: Quantifying Interpretability of Deep Visual Representations," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6541-6549, 2017.

[8] D. Bau, J.-Y. Zhu, H. Strobelt, A. Lapedriza, B. Zhou, A. Torralba, "Understanding the role of individual units in a deep neural network," Proceedings of the National Academy of Sciences of the United States of America (PNAS), Dec. 2020.