Anıl Özcan – 235B7022

Dönem Sonu Projesi – Ara Raporu

Action recognition, bir bilgisayar sisteminin görüntüler veya videolar gibi dijital verilerden insan eylemlerini tanımlama ve anlama yeteneğini ifade eder. Perakende bağlamında, eylem tanıma, mağaza hırsızlığı gibi davranışları tespit etmek ve tanımlamak için kullanılabilir. Perakendeciler, bilgisayarla görme ve derin öğrenme tekniklerinden yararlanarak şüpheli faaliyetleri otomatik olarak tespit eden ve işaretleyen sistemler geliştirebilir, böylece güvenliği artırmaya ve kayıpları önlemeye yardımcı olabilir. Bu sistemler video görüntülerini gerçek zamanlı veya geriye dönük olarak analiz ederek kayıp önleme ve mağaza yönetimi için değerli bilgiler sağlayabilir. Perakendede sağlam ve etkili hırsızlık tanıma elde etmek için bilgisayarla görme ve derin öğrenme algoritmalarının bir kombinasyonu kullanılabilir. Geliştirilen proje ele alındığında, action recognition altında sadece shoplifting eylemlerinin tespit edilmesi ve perakendecilik üzerinde hırsızlık tabanlı kayıpların önlenmesine yönelik bir model ortaya koymak amaçlanmaktadır. Raporun devamında konusu kapsamında konusundaki literatür taraması ve esinlenilen makaleye değinilip, mimariden bahsedilmiştir.

"The "InternVideo: General Video Foundation Models via Generative and Discriminative Learning" başlıklı makalede, hem generative hem de discriminative self-supervised video learning kullanarak video içeriğini daha iyi anlamak için tasarlanmış bir temel model olan InternVideo tanıtılmaktadır. Öncelikle hareketsiz görüntülere odaklanan önceki modellerden farklı olarak InternVideo, maskeli video modelleme ve video-dil kontrastlı öğrenmeyi içermekte ve bu da videoyla ilgili çeşitli görevlerdeki performansını artırmaktadır. Makale, video işlemenin hesaplama açısından zorlu olduğunu belirtmekte bu karşılık yetenekli bir video temel modelinin faydalarının bu maliyetlerden daha ağır bastığını savunmaktadır. InternVideo'nun 39 video veri kümesi üzerinde en son teknolojiye sahip performansa ulaştığı ve video eylem tanıma, video-dil hizalama ve açık dünya video uygulamaları gibi görevlerde etkinliğini gösterdiği gösterilmiştir. Daha da önemlisi, Kinetics-400 ve Something-Something V2 ölçütlerinde sırasıyla %91,1 ve %77,2'lik top-1 accuracy oranına ulaşarak video anlama için genel uygulanabilirliğini ortaya koymuştur. InternVideo'nun kodu, daha fazla araştırma ve geliştirme için GitHub'da kullanıma sunulması planlanmıştır. (Wang et al., 2022)

Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa ve diğerleri tarafından hazırlanan "On the Benefits of 3D Pose and Tracking for Human Action Recognition" başlıklı paperda, videolardaki insan eylemlerini tanımak için 3D izleme ve pozları kullanmanın avantajlarını araştırmaktadır. Uzayda sabit bir noktayı analiz etmek yerine, Lagrangian bakış açısını benimseyerek, bireylerin zaman içindeki hareketlerini izlemeye odaklanıp ve bu da bir nehir boyunca bir tekneyi takip etmeye benzetilmiştir. Çalışma, 3D izleme teknolojisini kullanarak video aracılığıyla insanların yörüngelerini işleyen bir yöntem geliştirip - bireyleri kareler arasında tanımlayıp ilişkilendirirken aynı zamanda 3D temsillerine de erişmiektedir. Bu, insanların yörüngesini, kimliğini, 3D pozunu ve konumunu hesaba katarak eylemlerin tanınmasına olanak tanıdığından bahsedilmiştir. Bir transformer mimarisi içeren önerilen model, yalnızca poz bilgilerini kullanan temel modelleri geliştirmek için 3D vücut pozu ve izleme verilerinden yararlanır. Yaklaşım, AVA v2.2 veri setinde state-of-art performansa ulaşarak gelişmiş eylem tanıma yetenekleri göstermekte ve 3D bağlamın dahil edilmesinin insan eylemlerini ve etkileşimlerini daha iyi anlamaya yardımcı olduğunu ortaya koymaktadır. Araştırma sonuçları ve kod çevrimiçi olarak ulaşılabilir durumdadır. (Rajasegaran et al., 2023)

"InternVideo2: Scaling Video Foundation Models for Multimodal Video Understanding" başlıklı makalede yukarıdaki InternVideo modelinin anlatıldığı makaleye benzer bir şekilde eylem tanıma ve video merkezli diyalog gibi çeşitli video ve ses görevleri için tasarlanmış state-of-art video temel modeli olan InternVideo2'yi tanıtıyor. Bu model, maskelenmiş video belirteci yeniden yapılandırma, modlar arası kontrast öğrenme ve bir sonraki belirteç tahmini dahil olmak üzere multi-self-supervised ve weak-supervised öğrenme yöntemlerini birleştiren aşamalı bir eğitim paradigması kullanıyor. Mekansal-zamansal tutarlılığı ve çok modlu hizalamayı sağlamak için eğitim için büyük ölçekli bir veri kümesinden yararlanarak 60'tan fazla farklı görevde performansı artırmak için hem verilerin hem de modelin boyutunu ölçeklendiriyor. InternVideo2'nin videolardaki uzun dizileri etkili bir şekilde muhakeme ederek ve karmaşık bağlamları kavrayarak diğer modellerden daha iyi performans gösterdiği vurgulanmıştır. (Wang et al., 2024)

Eylem tanıma alanındaki araştırmaların kapsamlı bir incelemesi yapıldıktan sonra, özellikle mağaza hırsızlığı tespitiyle ilgili çalışmalara odaklanılmıştır. Analiz, perakende ortamlarında hırsızlık olaylarını tespit etme ve önleme zorluğunu ele almak için bilgisayarla görme ve derin öğrenme tekniklerini kullanan ve giderek büyüyen bir çalışma grubunu ortaya çıkarmıştır. Birçok çalışma, mağaza hırsızlığı faaliyetlerini otomatik olarak tespit etmek ve işaretlemek için güvenlik kamerası görüntülerinden ve gelişmiş algoritmalardan yararlanmanın etkinliğini göstermiştir. Bu yaklaşımlar, normal alışveriş davranışı ile mağaza hırsızlığı ile ilişkili şüpheli faaliyetler arasında doğru ayrım yapma konusunda umut verici sonuçlar ortaya koymuştur. Genel olarak, perakende sektöründe hırsızlık tanıma konusunun incelenmesi, bilgisayarla görme ve derin öğrenme teknolojilerinin güvenlik önlemlerini artırma ve ticari ortamlardaki kayıpları azaltma potansiyeline ışık tutmuştur.

"Detection of Shoplifting on Video Using a Hybrid Network" başlıklı makale, mağaza hırsızlığı konusunu ve üretilen büyük miktarda veri nedeniyle video gözetimi yoluyla tespit etmenin zorluklarını tartışmaktadır. Bunu ele almak için yazarlar, evrişimsel ve tekrarlayan ağları birleştiren hibrit bir sinir ağına dayalı bir sınıflandırıcı tanıtmışlardır. Konvolüsyonel ağlar, özellikleri çıkarmak için video karelerini analiz ederken, tekrarlayan ağlar (özellikle geçitli tekrarlayan birimler) video dizilerini sınıflandırmak için bu özellikleri zaman içinde işler. UCF-Crime veri kümesini kullanan sınıflandırıcı, video verilerinden mağaza hırsızlığını tespit etmede % 93 gibi yüksek bir doğruluk oranına ulaşmıştır. Sonuçlar, bu yaklaşımın etkinliğini göstermektedir ve araştırma, gözetim sistemlerinde gerçek zamanlı hırsızlık tespiti için hibrit ağın pratik uygulamalarını ilerletmeyi amaçlamaktadır (Kirichenko et al., 2022). UCF-Crime veri seti, diğer anormal platformu sağlar davranış türlerinin yanı sıra mağaza hırsızlığı da dahil olmak üzere çeşitli suç faaliyetlerini gösteren gözetim videolarından oluşan yaygın olarak tanınan bir veri setidir. Araştırmacılara otomatik gözetim ve suç tespiti için algoritmalar geliştirmek ve değerlendirmek için sağlam bir eğitim ve test platformu sağlamaktadır.

İncelenen son makale, projenin esinlenileceği makaledir "Detecting abnormal behavior in megastore for intelligent surveillance through 3D deep convolutional model" başlıklı makalede video gözetimi yoluyla mega mağazalardaki ve dükkanlardaki anormal davranışların tespiti için gelişmiş üç boyutlu konvolüsyonel sinir ağlarının kullanılması üzerine bir çalışma sunmaktadır. Bu çalışmada önerilen ağ, video içeriğini verimli bir şekilde analiz etmek ve uzamsal-zamansal özellikleri çıkarmak için 18 konvolüsyonel işlem kullanarak 15 katman derinliğinde tasarlanmıştır. Bu özellikler daha sonra öğrenme sürecinde kullanılmakta ve dizileri etiketleyen bir softmax katmanıyla sonuçlanmaktadır. Yazarlar, beş tür davranışı temsil eden video kliplerden oluşan bir

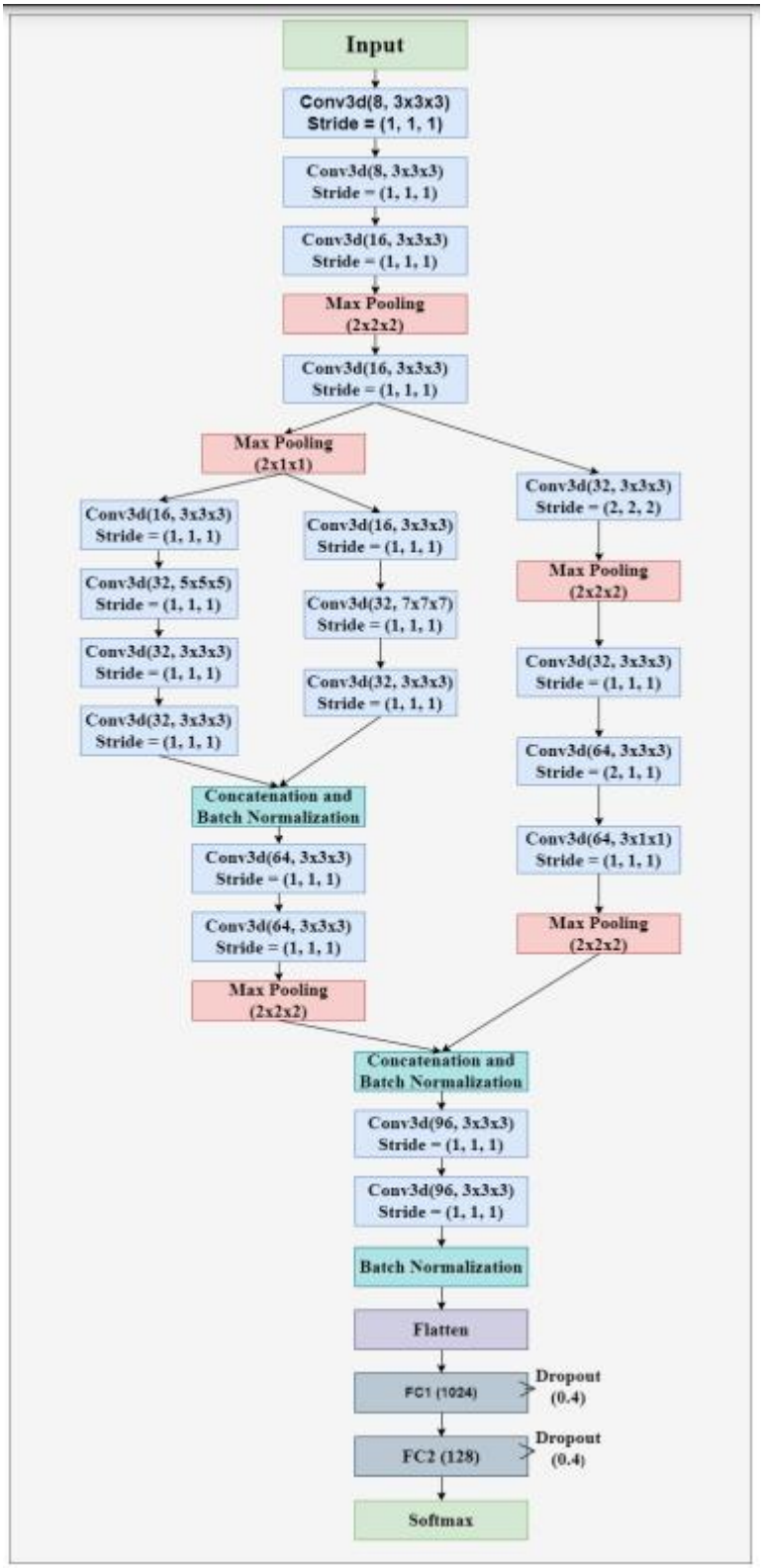

Figure 1 Önerilen Model Mimarisi

veri kümesi oluşturdu: normal, hırsızlık, içme, yemek yeme ve zarar verme. Sistem, insan eylemlerini analiz ederek bu anormal davranışları tanımlamak ve bu tür davranışlar tespit edilirse

uyarılar oluşturmak için tasarlanmıştır. Sentezlenen veri kümesi üzerinde yapılan kapsamlı deneyler %90,90'a varan bir doğruluk oranıyla sonuçlanmıştır. Araştırma, insan gözetiminden kaynaklanan yüksek yanlış sınıflandırma oranları gibi mevcut gözetim sistemlerinin karşılaştığı zorlukları, güvenliği artırmak ve perakendeciler için kayıplara neden olan olayları azaltmak için otomatik bir çözümle ele almayı amaçlamaktadır.

Geliştirilmesi planan model yukarıdaki makaleden esinlenildiği için büyük oranda bu mimariyi taklit edecektir. Kazanımlar doğrultusunda model içerisindeki herhangi bir katman- parallel katman çıkartılıp eklenebilir. Farklı ön işleme adımları uygulanabilir. Temsil alınan mimari soldaki gibidir.

References

Ansari, M A., Singh, D K., & Singh, V P. (2023, June 1). Detecting abnormal behavior in megastore for intelligent surveillance through 3D deep convolutional model. De Gruyter, 74(3), 140-153. https://doi.org/10.2478/jee-2023-0020

Kirichenko, L., Радівілова, T., Sydorenko, B., & Yakovlev, S. (2022, November 6). Detection of Shoplifting on Video Using a Hybrid Network. https://doi.org/10.3390/computation10110199

Rajasegaran, J., Pavlakos, G., Kanazawa, A., Feichtenhofer, C., & Malik, J. (2023, April 3). On the Benefits of 3D Pose and Tracking for Human Action Recognition. https://doi.org/10.48550/arxiv.2304.01199

Wang, Y., Li, K., Li, X., Yu, J., He, Y., Guo, C., Pei, B., Zheng, R., Xu, J., Wang, Z., Shi, Y., Jiang, T., Li, S., Zhang, H., Huang, Y., Qiao, Y., Wang, Y., & Wang, L. (2024, March 22). InternVideo2: Scaling Video Foundation Models for Multimodal Video Understanding. https://doi.org/10.48550/arxiv.2403.15377

Wang, Y., Li, K., Li, Y., He, Y., Huang, B., Zhao, Z., Zhang, H., Xu, J., Liu, Y., Wang, Z., Xing, S., Chen, G., Pan, J., Yu, J., Wang, Y., Wang, L., & Qiao, Y. (2022, December 6). InternVideo: General Video Foundation Models via Generative and Discriminative Learning. https://doi.org/10.48550/arxiv.2212.03191