TimeSformer ile Transformer Tabanlı Zamansal-Uzamsal Video Sınıflandırması

1. GİRİŞ

Video sınıflandırma, bilgisayarla görü alanında giderek artan bir öneme sahiptir. Günümüzde, video verilerinin hızla artması ve bu verilerin analiz edilmesi gerekliliği, video sınıflandırma tekniklerinin geliştirilmesini zorunlu kılmaktadır. Video sınıflandırma, bir videonun içeriğini anlamak ve bu içeriği belirli sınıflara atamak amacıyla kullanılır. Bu süreç, hem zamansal hem de uzamsal bilgilerin bir arada işlenmesini gerektirir. Geleneksel yöntemler, bu tür karmaşık verileri işlemekte yetersiz kalırken, derin öğrenme teknikleri, özellikle de evrişimli sinir ağları (CNN) ve transformatör tabanlı modeller, video sınıflandırmada önemli başarılar elde etmiştir.

Son yıllarda, transformatör tabanlı modeller, doğal dil işleme (NLP) alanında elde ettikleri başarıların ardından bilgisayarla görü alanında da kullanılmaya başlanmıştır. Bu modeller, özellikle uzun mesafeli bağımlılıkları modelleme yetenekleri sayesinde, video sınıflandırma gibi hem zamansal hem de uzamsal bilgilerin bir arada işlenmesi gereken görevlerde etkili olmuştur. TimeSformer, bu alanda öne çıkan bir model olup, video sınıflandırma için hem zamansal hem de uzamsal bilgileri aynı anda işleyebilen bir transformatör mimarisidir (Oliveira & de Matos, 2022).

Bu çalışmada, TimeSformer ve diğer transformatör tabanlı modellerin video sınıflandırma performansları incelenmekte ve bu modellerin farklı veri kümeleri üzerindeki genelleme yetenekleri değerlendirilmektedir. Ayrıca, transfer öğrenme ve topluluk öğrenmesi gibi tekniklerin bu modellerin performansını nasıl etkilediği araştırılmaktadır. Bu sayede, video sınıflandırma görevlerinde daha yüksek doğruluk oranlarına ulaşmak ve hesaplama kaynaklarını daha verimli kullanmak amaçlanmaktadır.

2. LİTERATÜR ARAŞTIRMASI

Video sınıflandırma, bilgisayarla görü alanında uzun yıllardır üzerinde çalışılan bir konudur. İlk çalışmalar, geleneksel makine öğrenmesi tekniklerine dayanmaktaydı. Ancak, bu yöntemler, özellikle büyük ölçekli veri kümeleri üzerinde yetersiz kalmaktaydı. Derin öğrenmenin ortaya çıkışıyla birlikte, özellikle evrişimli sinir ağları (CNN) video sınıflandırma görevlerinde yaygın olarak kullanılmaya başlandı. CNN'ler, görüntülerden özellik çıkarma konusunda oldukça başarılıdır ve bu özellikleri kullanarak videoları sınıflandırabilirler. Ancak, CNN'lerin temel sınırlaması, zamansal bilgileri işleme konusundaki yetersizlikleridir (Duvvuri et al., 2023).

Zamansal bilgileri işlemek için, uzun kısa vadeli bellek (LSTM) ve diğer tekrarlayan sinir ağı (RNN) modelleri kullanılmıştır. Bu modeller, videolardaki kareler arasındaki zamansal bağımlılıkları modelleyebilir. Ancak, RNN tabanlı modellerin eğitimi zor ve zaman alıcıdır. Ayrıca, uzun mesafeli bağımlılıkları modelleme konusunda da sınırlıdırlar.(Wang & Wang, 2021)

Transformatör tabanlı modeller, doğal dil işleme (NLP) alanında elde ettikleri başarıların ardından bilgisayarla görü alanında da kullanılmaya başlanmıştır. Transformatörler, özellikle uzun mesafeli bağımlılıkları modelleme yetenekleri sayesinde, video sınıflandırma gibi görevlerde etkili olmuştur. Video Swin Transformer (VST), bu alanda öne çıkan bir model olup, video sınıflandırma için hem zamansal hem de uzamsal bilgileri aynı anda işleyebilen bir transformatör mimarisidir (Oliveira & de Matos, 2022). VST, çeşitli veri kümelerinde doğruluk ve verimlilik açısından son teknoloji (state-of-the-art) sonuçlar elde etmiştir. Özellikle, Kinetics-400 veri kümesi üzerinde önceden eğitilmiş bir model kullanılarak FCVID ve Something-Something gibi büyük ölçekli veri kümeleri üzerinde transfer öğrenme yöntemiyle performansı incelenmiştir. Bu yaklaşım, modeli sıfırdan eğitmeye kıyasla yaklaşık 4 kat daha az bellek gerektirmektedir (Oliveira & de Matos, 2022).

Transfer öğrenme, video sınıflandırma görevlerinde sıkça kullanılan bir tekniktir. Bu teknik, önceden eğitilmiş bir modelin, yeni bir görev üzerinde yeniden eğitilmesini içerir. Transfer öğrenme, özellikle büyük veri kümeleri üzerinde eğitilmiş modellerin, daha küçük veri

kümeleri üzerinde de etkili olmasını sağlar. VST'nin transfer öğrenme uygulanarak yeniden eğitilmeden de farklı alanlardaki videoları sınıflandırabildiği gösterilmiştir. Ancak, bu yalnızca hedef sınıfların, modelin eğitildiği sınıflarla aynı türde olduğu durumlarda geçerlidir. Örneğin, Kinetics-400'den FCVID'e transfer öğrenme yapıldığında yüksek doğruluk oranları elde edilmiştir; çünkü her iki veri kümesi de ağırlıklı olarak nesneleri hedef almaktadır. Ancak, hedef sınıflar eğitildiği sınıflardan farklı türdeyse, transfer öğrenme sonrası doğruluğun düşük olması beklenmektedir. Bunu, çoğunlukla nesneleri içeren Kinetics-400 veri kümesinden, daha çok eylemleri temsil eden Something-Something veri kümesine transfer öğrenme yapıldığında gözlemlenmiştir(Oliveira & de Matos, 2022).

Topluluk öğrenmesi (ensemble learning), birden fazla modelin bir araya getirilerek daha güçlü bir model oluşturulması tekniğidir. Bu teknik, özellikle video sınıflandırma görevlerinde, bireysel modellerin performansını artırmak için kullanılır. Topluluk öğrenmesi, farklı modellerin güçlü yönlerini bir araya getirerek, daha yüksek doğruluk oranları elde edilmesini sağlar. Ancak, topluluk öğrenmesi, bireysel modellere göre daha fazla hesaplama kaynağı gerektirir ve eğitim süresi daha uzundur (Duvvuri et al., 2023). Örneğin, UCF-11 veri kümesi üzerinde yapılan bir çalışmada, özel bir CNN modeli ile topluluk öğrenme modeli karşılaştırılmış ve topluluk öğrenme modelinin daha yüksek doğruluk oranları elde ettiği gösterilmiştir (Duvvuri et al., 2023) .

Derin öğrenme modellerinin video sınıflandırma görevlerindeki başarısı, özellikle büyük veri kümeleri üzerinde eğitilmiş modellerin kullanılmasıyla artmıştır. ResNet18, VGG19 ve AlexNet gibi önceden eğitilmiş modeller, video sınıflandırma görevlerinde yüksek doğruluk oranları elde etmiştir. Özellikle, ResNet18 mimarisi, katman sayısı ve derin ağların etkisiyle en yüksek doğruluk oranını (%98.8) elde etmiştir (Akarsu & Karacalı, 2023). Bu sonuçlar, evrişim katmanlarının sayısı arttıkça, daha üst düzey özellikler elde edildiğini ve bu durumun test doğruluğunu önemli ölçüde artırdığını göstermektedir (Akarsu & Karacalı, 2023).

İnsan iskeleti davranış tanıma, video sınıflandırma görevlerinde giderek daha fazla ilgi görmektedir. Bu yaklaşım, basit bir arka plana sahip olması, aydınlatma faktörlerinden etkilenmemesi ve insan görünümündeki değişikliklere duyarlı olmaması nedeniyle avantajlıdır. Derin öğrenme yöntemi kullanılarak, insan iskeleti davranışlarını tespit etmek ve sınıflandırmak için çerçeve örnekleme ve görüntü örnekleme gibi video ön işleme işlemleri gerçekleştirilmiştir. Ardından, bir LSTM katmanı kullanılarak karelerin ardışık özellikleri çıkarılmış ve tam bağlı katman aracılığıyla sınıflandırma etiketinin olasılığı hesaplanmıştır. Bu

modelin insan iskeleti davranışı sınıflandırmadaki doğruluk oranı neredeyse %100'dür (Wang & Wang, 2021).

Sonuç olarak, video sınıflandırma görevlerinde, transformatör tabanlı modeller ve transfer öğrenme gibi teknikler, yüksek doğruluk oranları elde etmek için etkili yöntemlerdir. Ancak, bu modellerin farklı veri kümeleri üzerindeki genelleme yetenekleri ve hesaplama kaynaklarının verimli kullanımı hala araştırılmaya devam eden konulardır. Bu çalışmada, TimeSformer ve diğer transformatör tabanlı modellerin video sınıflandırma performansları incelenmekte ve bu modellerin farklı veri kümeleri üzerindeki genelleme yetenekleri değerlendirilmektedir. Ayrıca, transfer öğrenme ve topluluk öğrenmesi gibi tekniklerin bu modellerin performansını nasıl etkilediği araştırılmaktadır.

KAYNAKÇA

Akarsu, E., & Karacalı, T. (2023). Video Classification Results with Artificial Intelligence and Machine Learning. *International Journal of Innovative Research and Reviews (INJIRR)*, *7*(1), 22–26. http://www.injirr.com/article/view/194

Duvvuri, K., Kanisettypalli, H., Jaswanth, K., & Murali, K. (2023). Video Classification Using CNN and Ensemble Learning. *2023 9th International Conference on Advanced Computing and Communication Systems, ICACCS 2023*, 66–70. https://doi.org/10.1109/ICACCS57279.2023.10112975

Oliveira, D., & de Matos, D. M. (2022). *Transfer-learning for video classification: Video Swin Transformer on multiple domains*. http://arxiv.org/abs/2210.09969

Wang, J., & Wang, Z. (2021). Research on Video Classification Based on Deep Learning. *Proceedings - 2021 International Conference on Computer Network, Electronic and Automation, ICCNEA 2021*, 111–115. https://doi.org/10.1109/ICCNEA53019.2021.00034