Veri Temizliği ve Etiketleme:

Bu çalışma, görüntü tabanlı yapay zekâ modellerinde (özellikle sınıflandırma ve segmentasyon gibi bilgisayarlı görü görevlerinde) kullanılacak veri setinin hazırlanması sürecini kapsamaktadır. Bir modelin başarısı yalnızca kullanılan algoritmaya değil, büyük ölçüde eğitildiği verinin kalitesine bağlıdır. Bu nedenle veri seti oluşturulurken yalnızca görsellerin toplanması yeterli değildir; verilerin temizlenmesi, düzenlenmesi, doğru biçimde etiketlenmesi ve varsa maskelerin kalitesinin kontrol edilmesi gerekir. Bu bölümde veri temizliği ve etiketleme kavramları açıklanmış, ardından izlenecek süreç adım adım özetlenmiştir.

Veri temizliği (data cleaning), veri seti içindeki hatalı veya modelin öğrenmesini olumsuz etkileyebilecek örneklerin tespit edilip ayıklanması işlemidir. Görsel veri setlerinde en sık karşılaşılan problemler çözünürlük farklılıkları, pozlama sorunları ve tekrar eden görüntülerdir.

Çözünürlük, görselin piksel boyutunu ifade eder ve modelin detayları öğrenebilmesi için yeterli düzeyde olmalıdır. Çok düşük çözünürlüklü görsellerde nesne detayları kaybolduğu için modelin doğru öğrenmesi zorlaşır. Öte yandan aşırı yüksek çözünürlükler işlem maliyetini artırabilir ve standartlaştırmayı zorlaştırabilir. Bu nedenle veri setinde çözünürlüklerin belirli bir aralıkta tutulması ve gerekirse yeniden boyutlandırma uygulanması önerilir.

Pozlama (exposure) ise görselin ışık dengesini ifade eder. Aşırı karanlık veya aşırı parlak görüntülerde kıyafet dokusu, renk ayrıntıları ve sınırlar kaybolabilir. Bu durum etiketleme sürecini zorlaştırdığı gibi modelin öğrenmesini de olumsuz etkiler. Ayrıca bulanıklık, hareket kaynaklı bozulmalar ve düşük kalite (noise, sıkıştırma artefaktları) gibi sorunlar da veri setinin genel kalitesini düşürür. Veri temizliği aşamasında bu tür görseller tespit edilerek veri setinden çıkarılır veya ayrı bir kategori altında işaretlenir.

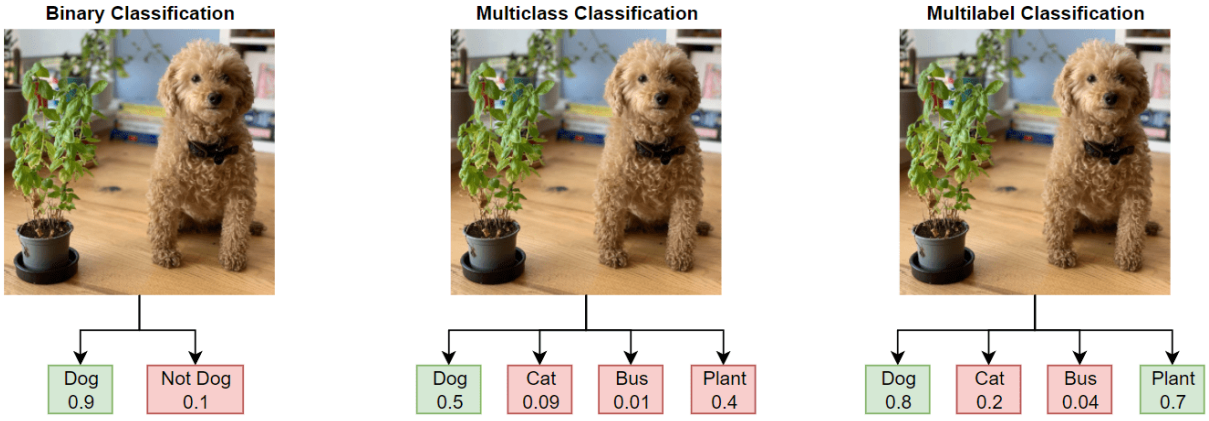
Görsel veri setlerinde dikkat edilmesi gereken bir diğer konu **tekrar veriler (duplicate/near-duplicate)** problemidir. Aynı görselin veya çok benzer görsellerin veri setinde fazla sayıda bulunması modelin genelleme kabiliyetini azaltabilir ve ezberlemeye yol açabilir. Bu nedenle veri setinde birebir tekrar eden görüntüler ve aynı fotoğrafın farklı kırpma/yeniden boyutlandırma gibi varyasyonları kontrol edilerek azaltılmalıdır. Böylece veri seti daha dengeli ve temsil gücü yüksek hale gelir.

Veri seti temizlendikten sonra etiketleme (labeling) aşamasına geçilir. **Etiketleme**, görsellerin model tarafından öğrenilebilmesi için anlamlı sınıf bilgileriyle işaretlenmesi sürecidir.

- Örneğin kıyafet veri setlerinde görseller "hoodie", "coat", "dress" gibi sınıflarla etiketlenebilir. Bu etiketler, modelin eğitim sırasında hangi görselin hangi sınıfa ait olduğunu öğrenmesini sağlar. Etiketleme yapılırken en önemli nokta, etiketlerin tutarlı ve standart olmasıdır. Aynı sınıf için farklı yazımların kullanılması (örneğin "hoodie", "hoodi", "kapüşonlu") veri setinde karışıklığa neden olabilir. Bu nedenle etiketleme öncesinde sınıf isimleri belirlenmeli ve bir etiket sözlüğü oluşturulmalıdır.

Etiketleme sürecinde **multi-label** yaklaşımı da kullanılabilir. Multi-label etiketleme, bir görselin tek bir etikete değil birden fazla etikete sahip olabilmesidir.

- Kıyafet örneğinde bir görsel yalnızca "hoodie" olarak değil, aynı zamanda "black", "oversize", "winter" gibi ek özelliklerle de etiketlenebilir. Bu yaklaşım, veri setinin daha zengin hale gelmesini sağlar ve modelin daha ayrıntılı öğrenmesine olanak tanır. Özellikle kıyafet öneri sistemleri, özellik çıkarımı ve detaylı sınıflandırma gibi uygulamalarda multi-label etiketleme önemli bir avantaj sağlar.



Bazı projelerde görsellere başlıklandırma (captioning) uygulanır. **Görsel başlıklandırma**, her görselin kısa bir açıklama cümlesiyle ifade edilmesidir.

- Örneğin "A woman wearing a black hoodie in a studio background" gibi bir cümle, görselin içerdiği temel bilgileri metin olarak temsil eder. Bu yöntem özellikle metin-görsel ilişkisi kuran generatif modellerde ve veri setinin dokümantasyonunda fayda sağlar.

Segmentasyon kullanılan veri setlerinde ayrıca **maske (segmentation mask**) kavramı bulunur. **Segmentasyon maskesi,** görseldeki hedef nesnenin piksel düzeyinde işaretlenmiş halidir. Bu maske sayesinde model yalnızca nesnenin ne olduğunu değil, görüntüde tam olarak hangi bölgede bulunduğunu da öğrenir. Segmentasyon maskelerinin kalitesi model performansını doğrudan etkiler. Kalitesiz maskeler genellikle sınır taşmaları, nesnenin bir kısmının eksik işaretlenmesi, maske içinde kopukluklar veya gereksiz küçük parçalar (gürültü) şeklinde kendini gösterir. Bu nedenle maske kalitesi, görsel üzerinde maske bindirme (overlay) yöntemiyle kontrol edilerek değerlendirilmelidir.

Etiketleme ve maske üretiminde kullanılan araçlar, süreci hızlandırmak ve standartlaştırmak açısından önemlidir. **Label Studio ve CVAT** bu amaçla kullanılan yaygın araçlardır.

Label Studio genellikle sınıflandırma, çoklu etiketleme ve görsel başlıklandırma gibi görevlerde pratik bir kullanım sunar.

CVAT ise özellikle bounding box ve segmentasyon gibi detaylı anotasyonlarda daha güçlü araçlara sahiptir. Bu tür araçlar sayesinde etiketleme süreçleri daha düzenli yürütülür, hata oranı azaltılır ve veri seti üretimi daha verimli hale gelir.

Sonuç olarak veri temizliği ve etiketleme aşamaları, model eğitiminden önce gerçekleştirilmesi gereken kritik hazırlık süreçleridir. Etiketleme sürecinde tutarlılık sağlanması, multi-label yapıların doğru kurgulanması ve caption kullanımının değerlendirilmesi veri setinin öğrenilebilirliğini yükseltir. Segmentasyon maskelerinin kalite kontrolü ise piksel düzeyinde öğrenme yapan modellerde doğrudan performans artışı sağlar.