

Görü ve Dil için Derin Öğrenme

Aykut Erdem, Erkut Erdem, Nazlı İkizler Cinbiş



HACETTEPE UNIVERSITY
COMPUTER VISION LAB



Sunum planı



- Bütünleşik dil ve görme problemlerinin kısa bir tarihçesi
- Görüntü altyazılıama (*image captioning*)



- Sinirsel sözcük gömmeleri (*neural word embeddings*)
- Sıralama (*ranking*) ve aktarma (*transfer*) tabanlı yaklaşımalar



- Oluşturma (*generation*) tabanlı yaklaşımalar
- Diğer uygulamalar

Tarihçe ve görüntü altyazılama problemi



Önce, biraz tarihçe..

- **Bilgisayarlı görü** berimsel yaklaşımlardan yararlanarak görüntüleri ve videoları anlamayı amaçlar.
- **Doğal dil işleme** berimsel yaklaşımalar kullanarak metinleri analiz etmeyi veya yeni metinler yaratmayı amaçlar.
- Her iki disiplin de **yapay zeka** ve **yapay öğrenme** üzerine kurulu benzer yöntemler kullanır.
 - Örneğin, temsil amaçlı kullanılagelen kelime torbası (*bag of words*) yöntemi
- Son bir kaç yıldır ise her iki disiplin birbirlerine hiç olmadığı kadar yakınlaşmış durumdadır.

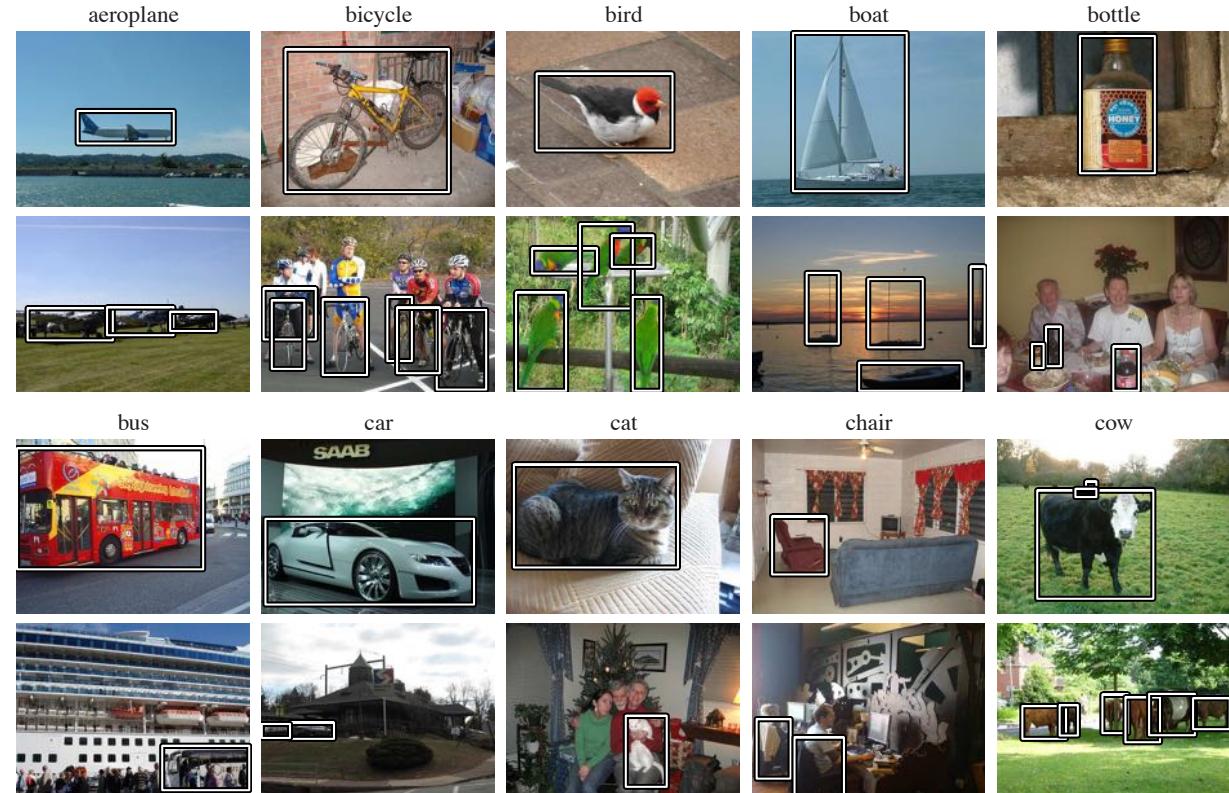
Önce, biraz tarihçe..

- Erken dönem çalışmalar genellikle verilen bir görüntüyü sözcüklerle eşleştirmektedirler.
- Örnek problemler:
 - **Kişi, nesne veya sahne tanıma** (*recognition*) [Everingham vd. 2010],...]
 - **Görüntü etiketleme** (*image tagging*) [Mori vd. 1999, Duygulu vd. 2002,...]
 - **Anlamsal bölütleme** (*semantic segmentation*), [Shotton vd. 2008, Tighe ve Lazebnik 2010,...]
 - **Görsel nitelikler** (*visual attributes*) [Farhadi vd. 2009,...]
- Bu çalışmaların tümü tek veya çok etiketli sınıflandırma problemleridir.

Nesne/sahne tanıma (*object/scene recognition*)

girdi: görüntü

çıktı: görüntünün
ait olduğu
görsel sınıf



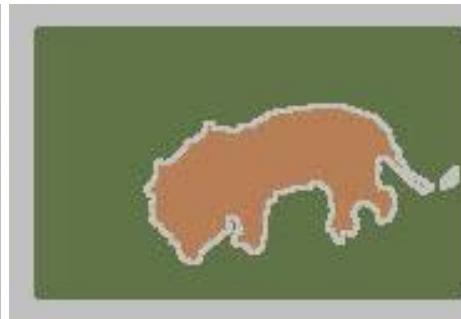
Görüntü etiketleme (*image tagging*)

girdi: görüntü

çıktı: ilişkili görsel anahtar sözcükler



sea sky sun waves



cat forest grass tiger



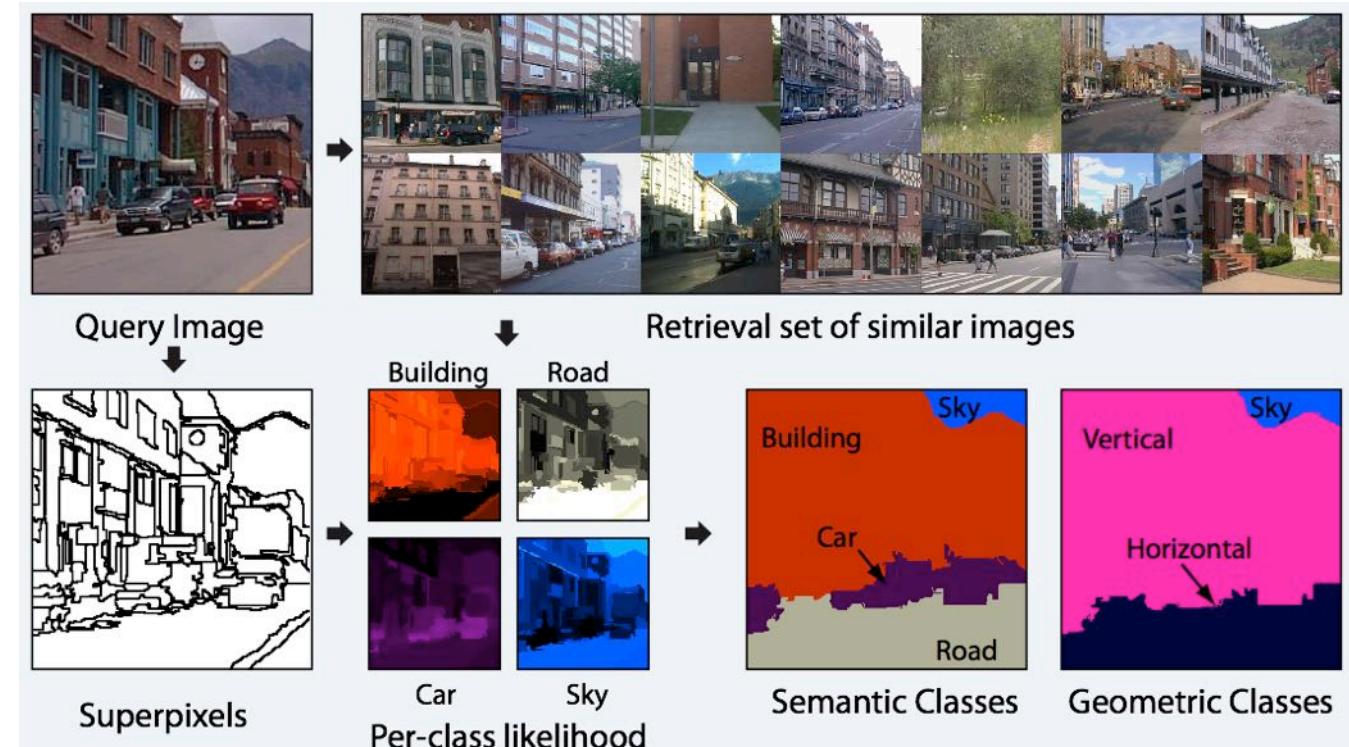
jet plane sky

Fig. 1. Examples from the Corel data set. We have associated keywords and segments for each image, but we don't know which word corresponds to which segment. The number of words and segments can be different; even when they are same, we may have more than one segment for a single word, or more than one word for a single blob.

Anlamsal bölütleme (semantic segmentation)

girdi: görüntü

çıktı: 1. bölütleme haritası, ve
2. her bölüm için ait olduğu görsel sınıfının etiketi



Görsel nitelikler (visual attributes)

girdi: görüntü

çıktı: görüntüyü niteleyen sıfatlar



'is 3D Boxy'
 'is Vert Cylinder'
 'has Window'
 'has Row Wind'
 X'has Headlight'



'has Hand'
 'has Arm'
 X'has Screen'



'has Head'
 'has Hair'
 'has Face'



'has Head'
 'has Torso'
 'has Arm'



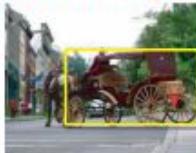
'has Head'
 'has Ear'
 'has Snout'



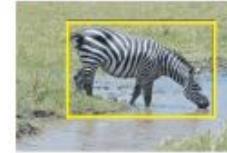
'has Head'
 'has Ear'
 'has Snout'



X'has Furniture Back'
 X'has Horn'
 X'has Screen'
 'has Plastic'
 'is Shiny'



'is 3D Boxy'
 'has Wheel'
 'has Window'
 'is Round'
 ''has Torso'



'has Tail'
 'has Snout'
 'has Leg'
 X'has Text'
 X'has Plastic'



'has Head'
 'has Ear'
 'has Snout'
 'has Leg'
 'has Cloth'



'is Horizontal Cylinder'
 X'has Beak'
 X'has Wing'
 X'has Side mirror'
 'has Metal'



'has Head'
 'has Snout'
 'has Horn'
 'has Torso'
 X'has Arm'

Görüntü-sözcük eşleştirmelerinden fazlasına neden ihtiyaç duyuyoruz?

- Sözcük bazlı yöntemler kısıtlı bir sözcük hazinesine (görsel sınıflar, anahtar sözcükler, vs.) sahiptirler.
- Eşlenen kelimeler veya kelime listeleri, görüntüde mevcut olan nesne ve varlıklar arasındaki ilişkileri ifade etmede yetersiz kalırlar:
 - Uzamsal ilişkiler (“masanın üzerindeki bardak”)
 - Eylemler (“sürahiyeden bardağa su dolduran bir kişi”)

Son yıllardaki gelişmeler

- Gündelik hayatı karşılaştığımız problemlerin bir çoğu **çok kipli** (*multi-modal*) bir yapıdadır.
- Gazete haberleri, kullanım kılavuzları, sosyal ağlardaki paylaşımlar, vs. **metinsel ve görsel bilgiyi birlikte** içermektedir.
- Literatürde bilgisayarlı görü ve doğal dil işleme yöntemlerinin bir arada kullanıldığı **yeni problemler** ve **yeni algoritmalar** önerilmeye başlanmış durumdadır:
 - **Görüntü altyazılıama** (*image captioning*), video altyazılıama, görsel soru yanıtlama (*visual question answering*), deyişe istinaden tanıma (), vs.

Bu görüntüyü nasıl açıklarsınız?



Bu görüntüyü nasıl açıklarsınız?



A child eating a cookie at a lemonade stand.

A Lemonade stand is manned by a blonde child with a cookie.

A small child at a lemonade ve cookie stand on a city corner.

A small child eating a cookie while standing next to a lemonade stand.

Young child behind lemonade stand eating a cookie.

Görüntü altyazılıama



A child eating a cookie at a lemonade stand.

A Lemonade stand is manned by a blonde child with a cookie.

A small child at a lemonade ve cookie stand on a city corner.

A small child eating a cookie while standing next to a lemonade stand.

Young child behind lemonade stand eating a cookie.

Girdi: görüntü **çıktı:** doğal bir cümle

- **Görüntü altyazılıama** verilen bir görüntüyü otomatik olarak doğal cümlelerle açıklamayı amaçlamaktadır.
- Görüntüde mevcut olan nesneler, onların nitelikleri, gerçekleşen eylemler ve sahne içindeki ilişkileri özetlenmektedir.

Görüntü altyazılıama

Automatic Description Generation from Images: A Survey of Models, Datasets, and Evaluation Measures

Raffaella Bernardi
University of Trento, Italy

Ruket Cakici
Middle East Technical University, Turkey

Desmond Elliott
University of Amsterdam, Netherlands

Aykut Erdem
Erkut Erdem
Nazli Ikizler-Cinbis
Hacettepe University, Turkey

Frank Keller
University of Edinburgh, UK

Adrian Muscat
University of Malta, Malta

Barbara Plank
University of Copenhagen, Denmark

BERNARDI@DISI.UNITN.IT

RUKEN@CENG.METU.EDU.TR

D.ELLIOTT@UVA.NL

AYKUT@CS.HACETTEPE.EDU.TR
ERKUT@CS.HACETTEPE.EDU.TR
NAZLI@CS.HACETTEPE.EDU.TR

KELLER@INF.ED.AC.UK

ADRIAN.MUSCAT@UM.EDU.MT

BPLANK@cst.dk

Abstract

Automatic description generation from natural images is a challenging problem that has recently received a large amount of interest from the computer vision and natural language processing communities. In this survey, we classify the existing approaches based on how they conceptualize this problem, viz., models that cast description as either generation problem or as a retrieval problem over a visual or multimodal representational space. We provide a detailed review of existing models, highlighting their advantages and disadvantages. Moreover, we give an overview of the benchmark image datasets and the evaluation measures that have been developed to assess the quality of machine-generated image descriptions. Finally we extrapolate future directions in the area of automatic image description generation.

1. Introduction

Over the past two decades, the fields of natural language processing (NLP) and computer vision (CV) have seen great advances in their respective goals of analyzing and generating text, and of understanding images and videos. While both fields share a similar set of methods rooted in artificial intelligence and machine learning, they have historically developed separately, and their scientific communities have typically interacted very little.

Recent years, however, have seen an upsurge of interest in problems that require a combination of linguistic and visual information. A lot of everyday tasks are of this nature, e.g., interpreting a photo in the context of a newspaper article, following instructions in conjunction with a diagram or a map, understanding slides while listening to a lecture. In

Automatic Description Generation from Images: A Survey of Models, Datasets, ve Evaluation Measures

R. Bernardi, R. Cakici, D. Elliott, **A. Erdem, E. Erdem, N. Ikizler-Cinbis, F. Keller, A. Muscat, B. Plank**

*Journal of Artificial Intelligence Research, Vol 55,
pages 409-442, February 2016*

Görüntü altyazılıama

- **Pratik uygulamalar**
 - Anlamsal görüntü veya video arama, görme engelli kişilere yönelik uygulamalar, vb.
- **Bilimsel merak**
 - Görme ve dil yetisini birlikte anlamak için bir test alanı
 - Görüntüleri ve cümleleri anlamak farklı düzeyden belirsizliklerle başa çıkabilmeyi gerektirir.
 - Yeni araştırma problemleri (görsel soru yanıtlama, deyişe istinaden tanıma, vs.)
- Görsel ve metinsel verinin bütünsel bir yapıda işlenmesini gerektirir.

Görüntü açıklamalarından kastedilen nedir?

- Bir görüntüyü açıklamak farklı açılardan ele alınabilir (Hodosh vd. 2013):

1. Görsel olmayan açıklamalar

- Görüntü veya videolarda olmayan ve arka plana ait açıklamalar (örn., görüntünün çekildiği yer ve zaman bilgisi)

2. Algısal açıklamalar

- Görüntülerin alt düzey özelliklerine ait açıklamalar (örn., görüntüdeki baskın renk tonu)

3. Kavramsal açıklamalar

- Görsel ve kavramsal bilgi içeren açıklamalar.
- Görüntüde mevcut olan varlıklar, onların özelliklerini, aralarındaki ilişkileri ve gerçekleşen eylemleri anlatan açıklamalar.

Görüntü açıklamalarından kastedilen nedir?



The dog is wearing a red sombrero.

Çok İyi: Fotoğraftaki tüm ana varlıklar, tam olarak ve uygun bir şekilde açıklanmış.

Dog wearing a red hat.

İyi: Bu şekildeki tamamlanmamış cümleler kabul edilebilir.

The white dog is wearing a pink collar.

Orta: Köpeği açıklamış ama şapkadan bahsetmemiş.

The red hat is adorned with gold sequins.

Kötü: Şapkayı detaylı şekilde açıklamış ancak köpektен bahsetmemiş.

The dog.

Çok Kötü: Yeterli detay yok. Herhangi bir fotoğraftaki, herhangi bir köpek olabilir.

Görüntü altyazılıma yaklaşımları



A child eating a cookie at a lemonade stand.

- **Görüntü altyazılıma** verilen bir görüntüyü otomatik olarak doğal cümlelerle açıklamayı amaçlamaktadır.
- Mevcut yaklaşımlar üç ana grup altında toplanabilir:
 1. Sıralama tabanlı,
 2. Aktarma tabanlı,
 3. Oluşturma tabanlı

Oluşturma tabanlı yaklaşımlar

- Görüntünün içeriğini tanıma ve kestirmeye çalışırlar:
 - Eylem, sahne ve nesne bilgisi
 - Nesne ve sahne nitelikleri
 - Nesneler arasındaki uzamsal ilişkiler
- Çıkarılan görsel bilgiyi bir doğal dil yaratma (natural language generation) yöntemi kullanarak cümle veya cümlelere dökerler.

Farhadi vd., 2010, Feng ve Lapata, 2010, Kulkarni vd., 2011, Li vd., 2011, Nwogu vd., 2011, Yang vd., 2011, Gupta vd., 2012, Kuznetsova vd., 2012, Mitchell vd., 2012, Ushiku vd., 2012, Kulkarni vd., 2013, Elliott ve Keller, 2013, Yatskar vd., 2014

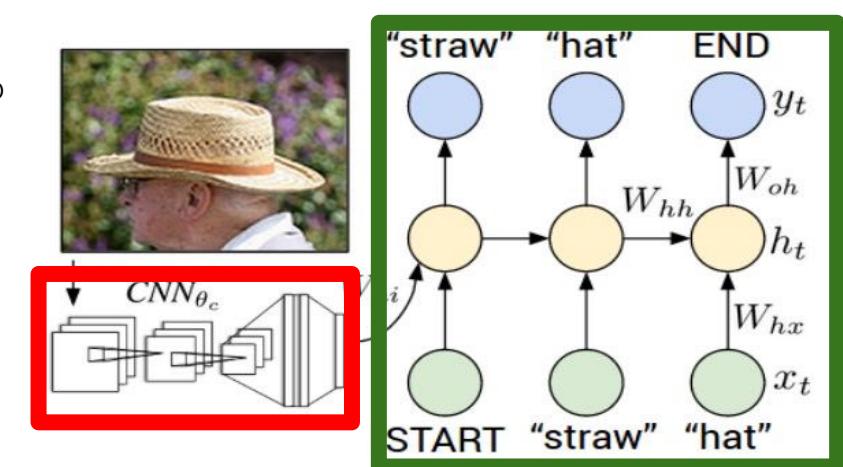


"This picture shows one person, one grass, one chair, ve one potted plant. The person is near the green grass, ve in the chair. The green grass is by the chair, ve near the potted plant." [Kulkarni vd., 2011]

Oluşturma tabanlı yaklaşımlar

- Son yıllarda bu amaçla derin öğrenme yaklaşımlarından yararlanılmaktadır.
- Explain Images with Multimodal Recurrent Neural Networks, Mao vd., 2015
- Deep Visual-Semantic Alignments for Generating Image Descriptions, Karpathy ve Fei-Fei, 2015
- Show ve Tell: A Neural Image Caption Generator, Vinyals vd., 2015
- Long-term Recurrent Convolutional Networks for Visual Recognition ve Description, Donahue vd., 2015
- Learning a Recurrent Visual Representation for Image Caption Generation, Chen ve Zitnick, 2015
- What Value Do Explicit High Level Concepts Have in Vision to Language Problems? Wu vd., 2016
- ...

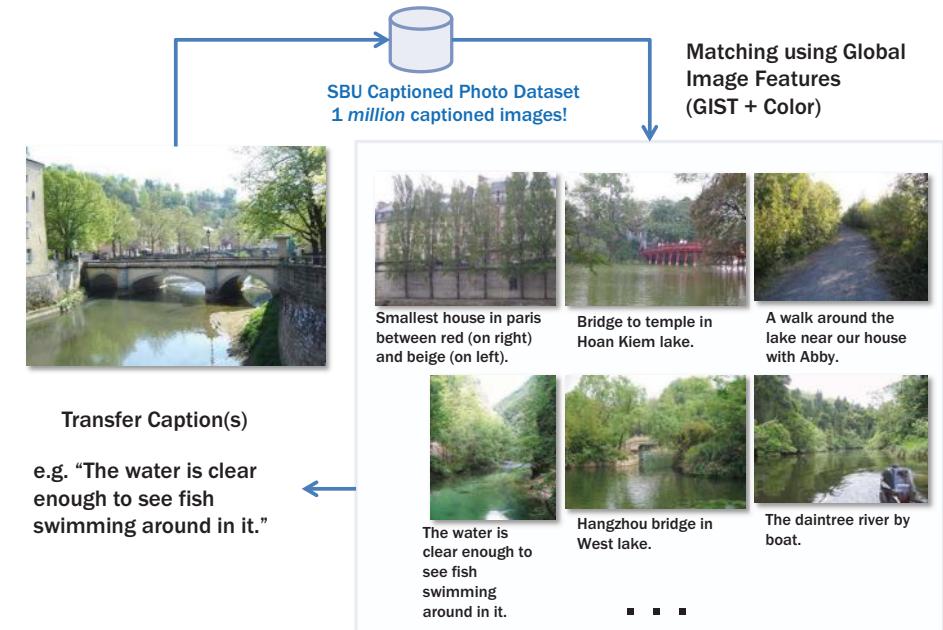
Recurrent Neural Network



Convolutional Neural Network

Aktarma tabanlı yaklaşımalar

- Verilen bir görüntü için büyük bir görüntükümesi içinde ona benzer görüntüler aranır.
- Benzer görüntüler içinden en ilgili olan görüntünün açıklaması girdi görüntüsünü açıklamak için kullanılır.

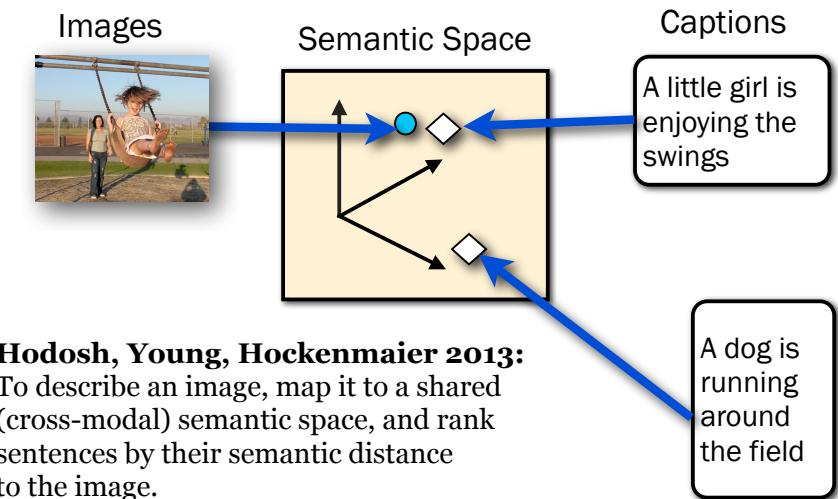


Ordonez vd., 2011, Patterson vd., 2014, Mason ve Charniak, 2014, Devlin vd., 2015, Yagcioglu vd., 2015

[Ordonez vd., 2011]

Sıralama tabanlı yaklaşımlar

- GörSEL ve metinsel veri için ortak bir anlamsal uzay öğrenirler.
- Verilen bir görüntü için bu ortak uzayda çapraz biçimli (cross-modal) getirim işlemi gerçekleştirirler.
- Canonical Correlation Analysis (CCA), Kernelized CCA, Predictive Dual-view Hashing (PDH) gibi yapay öğrenme teknikleri kullanılmaktadır.



Hodosh vd., 2013a, 2013b, Rastegari vd., 2013, Socher vd., 2014, Gong vd., 2014, Karpathy vd., 2014, Verma vd., 2014

[Hodosh vd., 2013]

Veri kümeleri



Man sits in a rusted car buried in the sand on Waltarere beach

- **Internet ağından serbestçe toplanmış** görüntüler ve altyazılar
 - BBC News (Feng ve Lapata, 2010), SBU Captioned Photos (Ordonez vd., 2011), YFCC100m (Thomee, 2016)
 - **Avantajları:** Geniş hacimli, doğal cümleler
 - **Dezavantajları:** Kavramsal açıklamalardan uzak olmaları

Veri kümeleri



- A green ve gray plane is taking off from the runway.
- A green ve white cargo plane taking of from an airport.
- A green ve white jet taking off.
- A Jade Cargo jet in green ve white taking off from an airport.
- White ve green commercial airliner taking off of runway.

- Internet ağından toplanmış ve **buna ek olarak açıklamaların elle oluşturulduğu** görüntüler
 - IAPR TC-12 (Grubinger vd., 2006), Pascal Sentences (Rashtchain vd., 2010), Flickr 8k Dataset (Hodosh vd., 2013), Flickr 30k Dataset (Yatskar vd., 2014), MS COCO (Lin vd., 2014)
 - **Avantajları:** Kavramsal açıklamalar içermeleri
 - **Dezavantajları:** Görece küçük hacimli, doğallıktan uzak olabilen cümleler

SBU Veri Kümesi



Man sits in a rusted car buried in the sand on Waltarere beach



Little girl ve her dog in northern Thailand. They both seemed interested in what we were doing



Interior design of modern white ve brown living room furniture against white wall with a lamp hanging



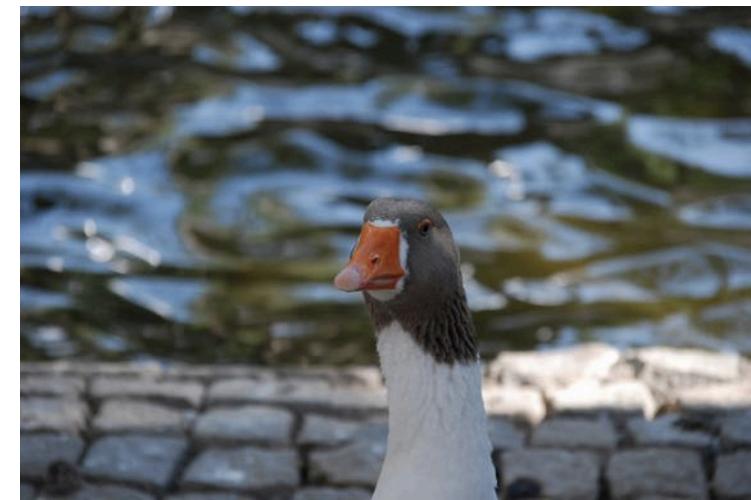
Emma in her hat looking super cute

- Flickr üzerinde kullanıcıların paylaştığı; farklı niteliklerde nesneler ve eylemler içeren, farklı sahneler barındıran yaklaşık 1,000,000 görüntü
- Her bir görüntü için ilgili kullanıcı tarafından yazılmış ortalamada 10 kelime içeren bir açıklama

Yahoo Flickr Creative Commons 100 Million



zojoji temple, minato-ku



Duck @ Kugulu Park. Watch these teeth!!

- Flickr üzerinde kullanıcıların paylaştığı; 100,000,000 görüntü
- Her bir görüntü için (varsı) ilgili kullanıcı tarafından yazılmış bir açıklama

UIUC Pascal Sentences



- A green ve gray plane is taking off from the runway.
- A green ve white cargo plane taking of from an airport.
- A green ve white jet taking off.
- A Jade Cargo jet in green ve white taking off from an airport.
- White ve green commercial airliner taking off of runway.



- A woman at a dinner table writing on her notebook.
- A woman at the dining table with wine, beer, ve lemons.
- A woman by the table preparing drinks.
- A woman sits with her head down at a table that has alcohol beverages and accessories on it.
- Woman at table busy with something

- 20 farklı nesne sınıfı içeren PASCAL VOC 2008 veri kümelerinden toplanan 1,000 görüntü
- Her görüntü için Amazon Mechanical Turk ile toplanmış 5 değişik açıklama

Flickr8k Veri kümesi



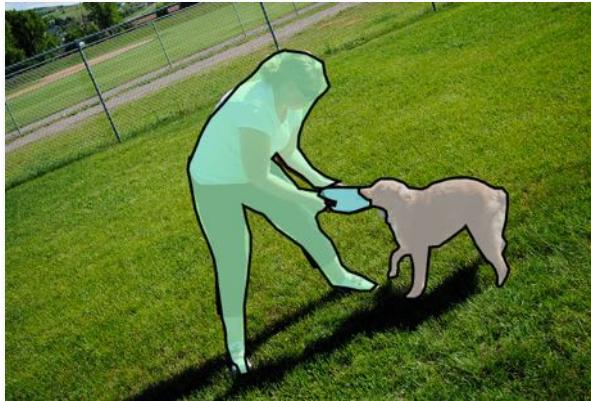
- A black, white, ve brown dog is playing with a green ve pink ball on sand.
 - A dog is in the sand with a pink ve green tennis ball in his mouth.
 - A little dog plays with a colorful ball in the sand.
 - A small terrier plays with a tennis ball on the beach.
 - The white dog, with dark patches, is playing in the sand with a tennis ball in its mouth.
-
- Flickr'dan toplanan insan, köpek, vs. içeren 8,091 görüntü.
 - Her görüntü için kitle kaynak (*crowdsourcing*) yaklaşım kullanılarak toplanmış 5 değişik açıklama

Flickr30k Veri Kümesi



- Gray haired man in black suit ve yellow tie working in a financial environment.
 - A graying man in a suit is perplexed at a business meeting.
 - A businessman in a yellow tie gives a frustrated look.
 - A man in a yellow tie is rubbing the back of his neck.
 - A man with a yellow tie looks concerned.
-
- Flickr'dan toplanan insan, köpek, vs. içeren 31,783 görüntü.
 - Her görüntü için kitle kaynak yaklaşım kullanılarak toplanmış 5 değişik açıklama

MS COCO Veri Kümesi



- a woman playing tug of war with a dog over a white frisbee.
- a woman wrestles a frisbee being held by a dog.
- a dog is biting a frisbee his owner is trying to get
- woman ve dog with frisbee on grass near fence.
- a dog is playing with their owner ve a frisbee.

- Veri kümesi toplamda 300,000 görüntüden oluşmaktadır.
- Her bir görüntüye ait yine kitle kaynak yaklaşım kullanarak toplanmış 5 değişik açıklama
- Görüntülerde bulunan 80 kategoriden nesnelere ait çıkartılan bülütler

Değerlendirme ölçüleri

- Literatürde en sık izlenen yöntem üretilen açıklamaları insan uzmanlarca **nesnel değerlendirmeye** tutmaktadır:
 - Gramer açısından ne kadar doğru?
 - İçerik bakımından ne kadar doğru?
 - Akıcılık (doğallık) bakımından ne dersiniz?
- **Avantajları:** Doğruluk
- **Dezavantajları:** Tekrar edilebilirlik, maliyet, güvenilirlik

Değerlendirme Ölçüleri

- **Sayısal analiz:** Üretilen açıklama ile kesin referans açıklama(lar) ne kadar tutarlı? Bunu otomatik olarak hesaplayabilir miyiz?
- **Avantajları:** Geniş hacimli deneyler
- **Dezavantajları:** İnsan hükümleriyle zayıf uyumluluk
- **Metin özetleme ve bilgisayarlı çeviri** yöntemleri için geliştirilmiş ölçüler
 - BLEU (Papineni vd. 2002), Rouge-1 (Lin ve Hovy 2008), ve METEOR (Banerjee ve Lavie 2005)
- **Görüntü açıklama** için özel olarak geliştirilmiş ölçüler
 - CIDEr (Vedantam vd., 2015) ve SPICE (Anderson vd., 2016)
- **Getirim** (retrieval) için kullanılan ölçülerı
 - Kesinlik/hassasiyet (precision/recall), asıl açıklamanın getirme sırası

MS-COCO Challenge 2015

	CIDEr-D	Meteor	ROUGE-L	BLEU-1	BLEU-2	BLEU-3	BLEU-4
ATT ^[2]	0.958	0.335	0.682	0.9	0.815	0.709	0.599
Google ^[7]	0.946	0.346	0.682	0.895	0.802	0.694	0.587
MSR Captivator ^[16]	0.937	0.339	0.68	0.907	0.819	0.71	0.601
m-RNN ^[10]	0.935	0.325	0.666	0.89	0.798	0.687	0.575
Berkeley LRCN ^[3]	0.934	0.335	0.678	0.895	0.804	0.695	0.585
MSR ^[15]	0.925	0.331	0.662	0.88	0.789	0.678	0.567
ACVT ^[1]	0.924	0.329	0.672	0.892	0.803	0.694	0.582
Nearest Neighbor ^[17]	0.916	0.318	0.648	0.872	0.77	0.655	0.542
Human ^[9]	0.91	0.335	0.626	0.88	0.744	0.603	0.471
Tsinghua Bigeye ^[22]	0.908	0.332	0.663	0.881	0.783	0.67	0.558
m-RNN (Baidu/ UCLA) ^[11]	0.896	0.32	0.668	0.89	0.801	0.69	0.578
Montreal/Toronto ^[14]	0.893	0.322	0.654	0.881	0.779	0.658	0.537

- En başarılı yöntemler derin öğrenme kullanan oluşturma tabanlı modeller
- Sonuçlara göre bilgisayarlar (algoritmalar) insanlardan daha başarılı!
- Peki, gerçekte öyle mi?

MS-COCO Challenge 2015

Results of the [2015 Captioning Challenge](#). Finalized 06/2015. See other tabs for up-to-date results.

	M1	M2	M3	M4	M5
Human ^[5]	0.638	0.675	4.836	3.428	0.352
Google ^[4]	0.273	0.317	4.107	2.742	0.233
MSR ^[11]	0.268	0.322	4.137	2.662	0.234
Montreal/Toronto ^[10]	0.262	0.272	3.932	2.832	0.197
MSR Captivator ^[12]	0.25	0.301	4.149	2.565	0.233
Berkeley LRCN ^[2]	0.246	0.268	3.924	2.786	0.204
m-RNN ^[6]	0.223	0.252	3.897	2.595	0.202
Nearest Neighbor ^[13]	0.216	0.255	3.801	2.716	0.196
PicSOM ^[15]	0.202	0.25	3.965	2.552	0.182
Brno University ^[3]	0.194	0.213	3.079	3.482	0.154
m-RNN (Baidu/ UCLA) ^[7]	0.19	0.241	3.831	2.548	0.195
MIL ^[8]	0.168	0.197	3.349	2.915	0.159

- Mevcut otomatik ölçülerin değerlendirme başarısı其实 kısıtlı ve insan uzmanlarının uzağında.
- Probleme özel yeni otomatik değerlendirme ölçülerine ihtiyacımız var:
 - CIDEr (Vedantam vd., 2015),
 - SPICE (Anderson vd., 2016)

Buraya kadar...

- Görüntü altyazılımanın temelleri
 - Yaklaşımalar
 - Veri kümeleri
 - Değerlendirme ölçütleri



A child eating a cookie at a lemonade stand.

Burada durmak zorunda değiliz...

- Yeni bir araştırma alanı olarak bütünlük görme ve dil problemleri

Görsel soru yanıtlama (*visual question answering*)



How many clouds are in the sky?

None



How many giraffes sitting?

Three



What is behind the photographer?

A bus



What color leaves are on the tree behind the elephant on the left of the photo?

Red

Burada durmak zorunda değiliz...

- Yeni bir araştırma alanı olarak bütünlük görme ve dil problemleri

Deyiše istinaden tanıma ve bölütleme (*identification and segmentation with referring expressions*)



visual object category: person
referring expression:
“the girl with red tie”

[Hu vd., 2016]



The man in black.



The man in red.



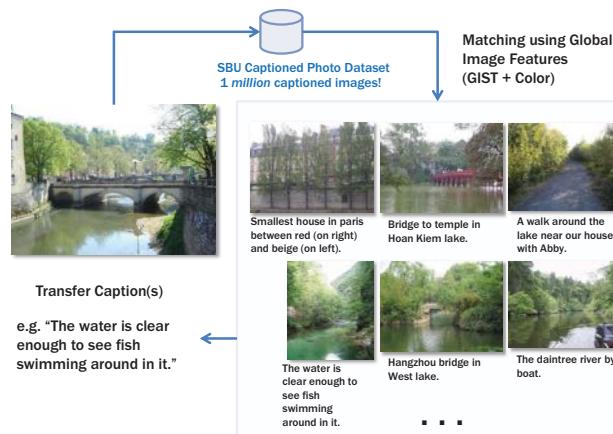
[Mao vd., 2016]

The skis.

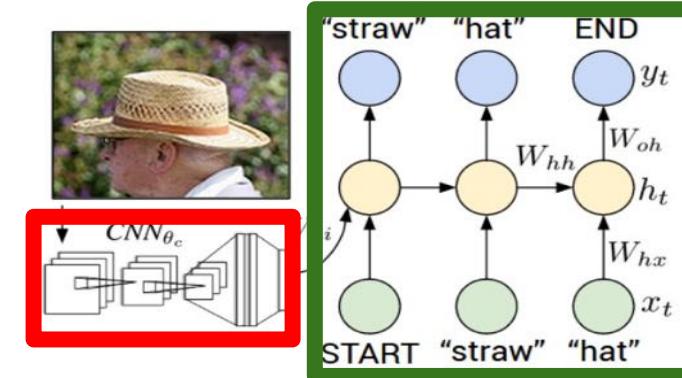


Bundan sonra biraz derinlere dalacağız...

- Sıralama ve aktarma tabanlı yaklaşımalar
- Sinirsel sözcük gömmeleri
- Oluşturma tabanlı yaklaşımalar
- Diğer uygulamalar



Recurrent Neural Network



Convolutional Neural Network