CNN-LSTM tabanlı Görsel Soru Cevaplama Modeli

BAŞAK BULUZ
DEEP LEARNING TURKIYE

DEEP CON '18

Başak BULUZ

Deep Learning Türkiye
İstanbul Aydın Üniversitesi – Bilgisayar Programcılığı (ING)

Matematik-Bilgisayar Lisans Eğitimi, İstanbul Aydın Üniversitesi
Bilgisayar Mühendisliği (ING) Lisans Eğitimi , İstanbul Aydın Üniversitesi
Bilgisayar Mühendisliği Yüksek lisans Eğitimi , Gebze Teknik Üniversitesi
Bilgisayar Mühendisliği Doktora Eğitimi , Gebze Teknik Üniversitesi

Görsel Soru Cevaplama (VQA) Nedir?

VISUAL QUESTION ANSWERING PROBLEM

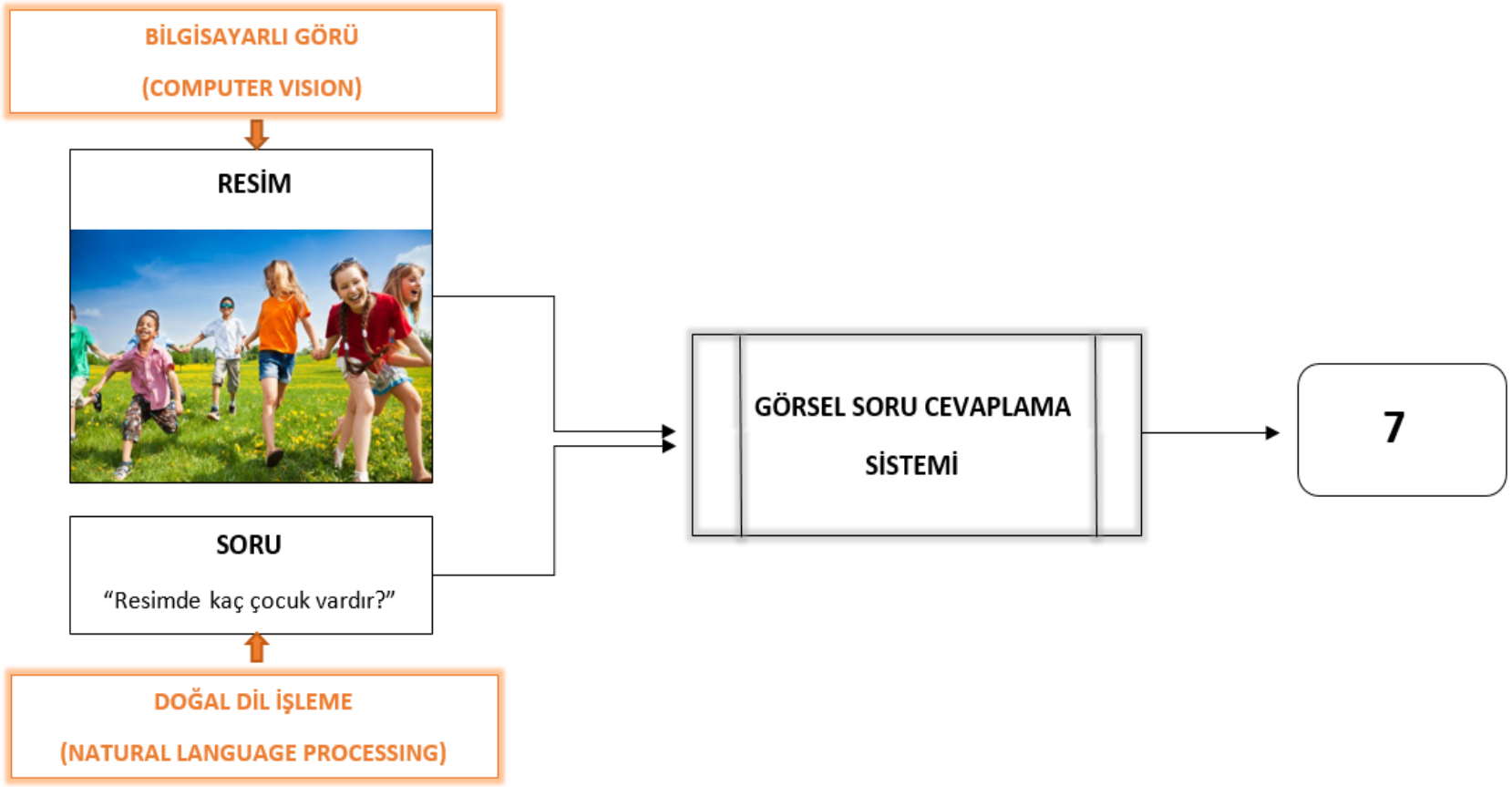
Görsel Soru Cevaplama (VQA) Nedir?

- Bu problemde metinler şeklinde ifade edilen soruların işlenmesi tarafı bir **doğal dil işleme problemi** iken; resimler içerisinden cevapların üretiminde, her bir soru ayrı bir **bilgisayarla görü problemine** işaret ediyor olabilir!

- Resimde kaç çocuk vardır?
  - **Nesne/varlık sayma (Counting)**
- Resimde turuncu kıyafeti olan bir çocuk var mıdır?
  - **Nesne/varlık tanıma (Object Detection)**
- Hava güneşli mi?
  - **Sahne sınıflandırması (Scene classification)**
- Resimdeki çiçekler hangi renk?
  - **Öznitelik sınıflandırması (Attribute classification)**

SORULAR..?

SİSTEMİN GENEL GÖRÜNÜŞÜ



VERİ KÜMELERİ

- **DAQUAR (Dataset for Question Answering on Real-world images) :**
  - 2015 yılında yayınlanmış ilk veri kümesi
  - Toplam 1449 resim ve her bir resme ait yaklaşık 8 farklı soru olmak üzere toplamda12,468 soru içerir.
- **COCO-QA:**
  - 123,287 adet doğal resim ve bu resimlere ait açıklamalar içeren MS COCO veri kümesinden oluşturulmuştur.
  - Bu veri kümesindeki sorular resimlere ait açıklamalardan doğal dil işleme yöntemleri ile otomatik oluşturulmuştur.
  - 78,736'i eğitim ve 38,948'de test için kullanılacak olan soru-cevap içerir.
  - Soruların yaklaşık %70'i resim içerisindeki nesneler ile ilgili ve diğer sorular ise konum, renk ve sayı bilgisi üzerine yoğunlaşmıştır. Her bir sorunun cevabı ise yalnızda tek bir kelimedir.

VERİ KÜMELERİ

- **VQA:**

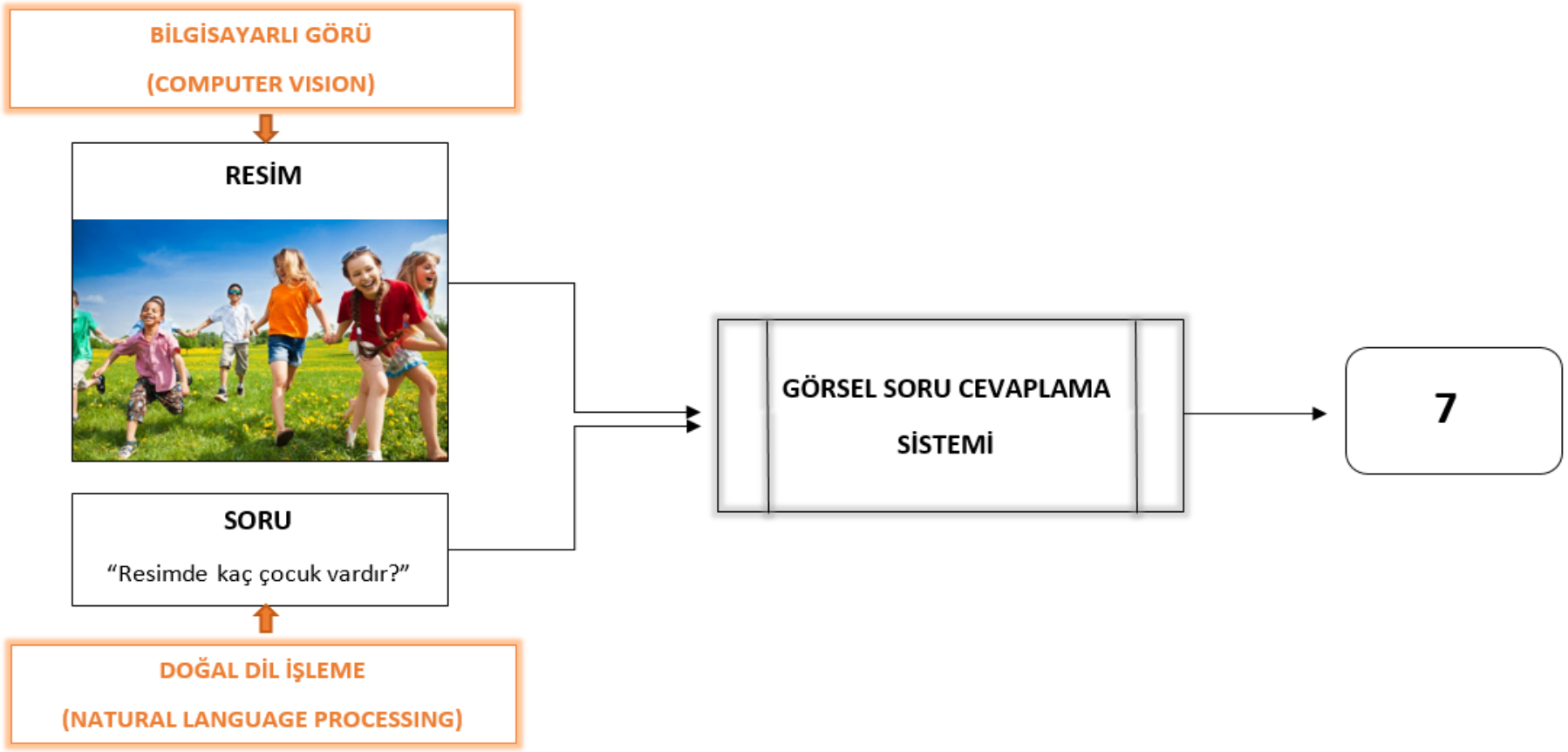
COCO VQA	SYNTH VQA	VQA
• MS COCO • Doğal Resimler	• 50.000 resim • Soyut resimler	• Her görüntüye ait 3 soru, 10 olası cevap • Açık uçlu sorular • Çoktan seçmeli sorular

VERİ KÜMELERİ

- **FM-IQA**:
  - Tam ismi "The Freestyle Multilingual Image Question Answering" olan bu veri kümesi COCO görüntü veri kümesini temel alarak oluşturulmuş, soru ve cevapları insanlar tarafından oluşturulmuş bir diğer görsel soru cevaplama veri kümesidir.
  - Aslında Çince olarak toplanan bu veri kümesinin, İngilizce'ye çevrilmiş haline de erişilebilmek mümkün.
  - Cevapların tam cümle olmasına da izin verilmiş olması ve iki dili destekliyor olması ile diğer veri kümelerinden ayrılmıştır.

- **VISUAL GENOME**
  - 2017 yılında yapılan 'Visual Genome: Connecting language and vision using crowdsourced dense image annotations' akademik çalışması ile yayınlanan veri kümesinde YFCC100M ve COCO görüntü veri kümelerinin birleştirilmesiyle elde edilmiş 108,249 görüntü, bu görüntülere ait 1.7 milyon soru-cevap çifti yer almaktadır.
  - Bu veri kümesinde yalnızca 'ne, nerede, nasıl, ne zaman, kim' soru kelimelerini içeren sorular bulunmaktadır, evet/hayır soruları yoktur.

Genel Yaklaşım

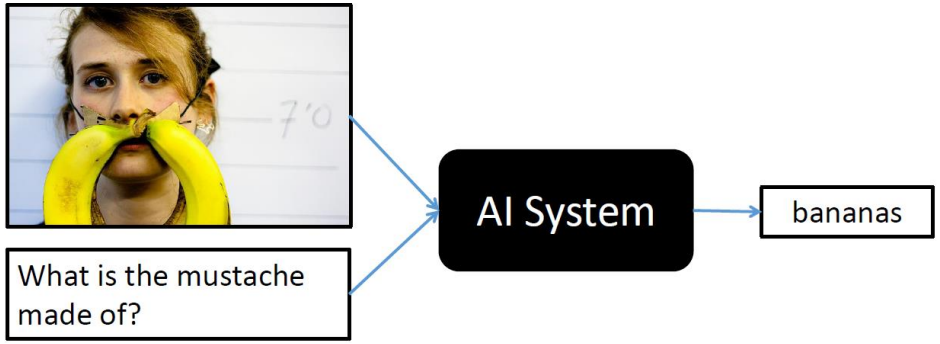


Genel Yaklaşım

- Yaklaşımların hepsi temelde;

1. **görüntü öznitelikleri elde edilmesi**
2. **soru özniteliklerinin elde edilmesi**
3. **özniteliklerin bir cevap üretmek için birleştirilmesi**

adımlarını sağlamaya yönelik bir yaklaşım benimser.



Genel Yaklaşım

- Görüntüye ait özniteliklerin çıkarılmasında;
  - VGGNet,
  - ResNet,
  - GoogleNet
- Soru özniteliklerinin çıkartılmasında;
  - kelime/Sözcük Çantası (bag-of-words),
  - uzun Kısa Vadeli Hafıza Ağları (LSTM) ,
  - geçitlenmiş özyinelemeli birimler (gated recurrent units -GRU)
  - skipthought vektörler
- Bu iki temel işlem tamamlandıktan sonra cevabın üretilmesi safhasında ise genel yaklaşım bu zor problemi **bir sınıflandırma problemine dönüştürmek** yönündedir.

Genel Yaklaşım

- Temel problem sınıflandırma problemine çevrilerek çözülmemiş, bazılarında cevabın üretilmesi (answer generation) yönünde farklı yaklaşımlarda denenen çalışmalar:
  - Ask Your Neurons: A Neural-based Approach to Answering Questions about Images
  - Are you talking to a machine? Dataset and methods for multilingual image question answering
  - Compositional memory for visual question answering
  - What value do explicit high level concepts have in vision to language problems?
  - Fvqa: fact-based visual question answering

Resimde kaç çocuk vardır?

GÖRÜNTÜ ÖZNİTELİKLERİNİN ÇIKARILMASI (IMAGE FEATURE EXTRACTION)

- Önceden eğitilmiş (pre-trained) CNN'in sondan bir önceki katmanının çıktısı
- Önceden eğitilmiş (pre-trained) CNN tarafından oluşturulmuş öznitellik haritasındaki yerel öznitelikler
- ....

SORU ÖZNİTELİKLERİNİN ÇIKARILMASI (QUESTION FEATURE EXTRACTION)

- Kelime/Sözcük Çantası (bag-of-words)
- LSTM / GRU dil modelleri
- Doğal dil ayrıştırıcıları (Natural language parser)
- ....

ALGORİTMA

- Bitiştirme (Concatenation)
- Terimsel çarpım / toplam (Elementwise product/sum)
- Bilineer Ortaklama (Bilinear Pooling)
- Dikkat modelleri (Attentive models)
- Bayesçi modeller (Bayesian models)
- ...

SINIFLANDIRICI

6
5
7
8
Evet
güneşli
:
:
:
Koşmak

ÖZNİTELİKLERİN BİRARAYA GETİRİLMESİ

- Bu zorlu görevde ki en çok soru işareti doğuran adım, görüntü ve soru özniteliklerinin bir araya getirilerek cevabın üretilmesidir.

  - Temel Modeller
  - Bayesçi ve Soruya Duyarlı Modeller (Bayesian and Question-Aware Models)
  - Dikkat Temelli Modeller (Attention Based Models)
  - …

Yöntem

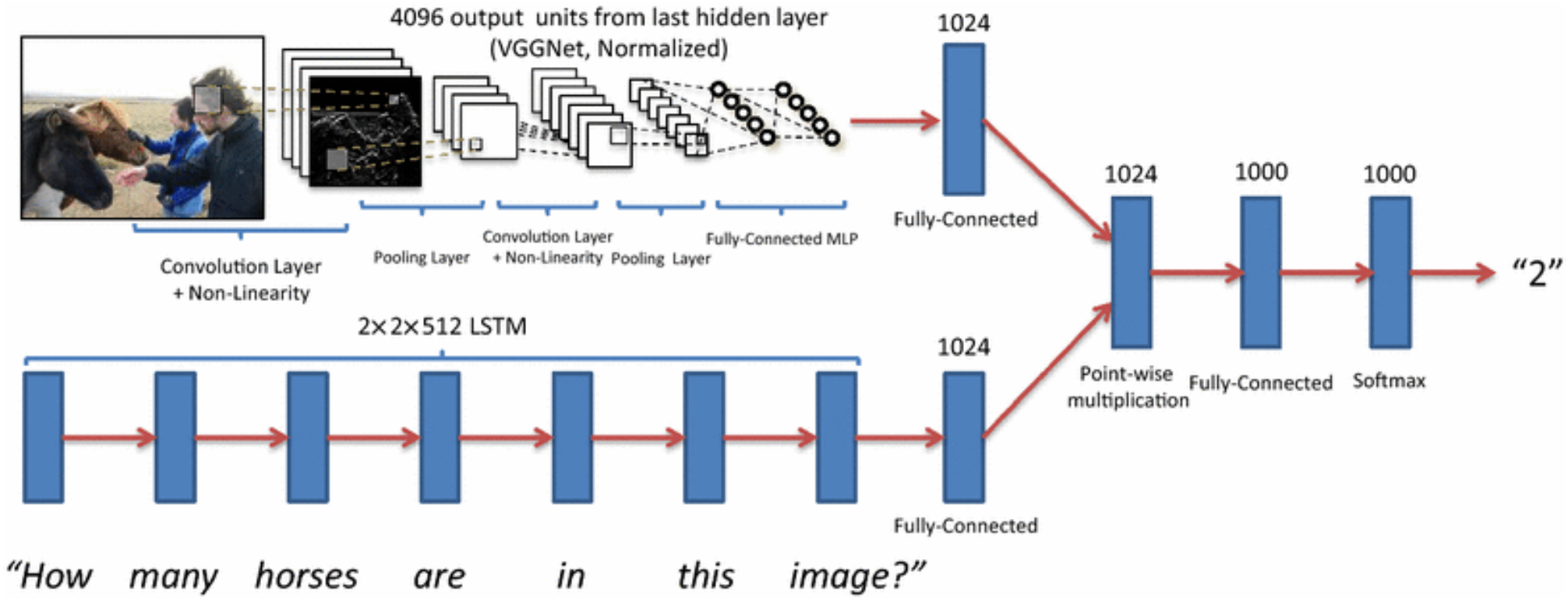
- **VERİ KÜMESİ (VQA)**

COCO VQA	SYNTH VQA	VQA
• MS COCO • Doğal Resimler	• 50.000 resim • Soyut resimler	• Her görüntüye ait 3 soru, 10 olası cevap • Açık uçlu sorular • Çoktan seçmeli sorular

Yöntem

3-4 (15.3%)	**5-8** (39.7%)	**9-12** (28.4%)	**13-17** (11.2%)	**18+** (5.5%)
Is that a bird in the sky?	How many pizzas are shown?	Where was this picture taken?	Is he likely to get mugged if he walked down a dark alleyway like this?	What type of architecture is this?
What color is the shoe?	What are the sheep eating?	What ceremony does the cake commemorate?	Is this a vegetarian meal?	Is this a Flemish bricklaying pattern?
How many zebras are there?	What color is his hair?	Are these boats too tall to fit under the bridge?	What type of beverage is in the glass?	How many calories are in this pizza?
Is there food on the table?	What sport is being played?	What is the name of the white shape under the batter?	Can you name the performer in the purple costume?	What government document is needed to partake in this activity?
Is this man wearing shoes?	Name one ingredient in the skillet.	Is this at the stadium?	Besides these humans, what other animals eat here?	What is the make and model of this vehicle?

YÖNTEM


4096 output units from last hidden layer
(VGGNet, Normalized)
1024
Convolution Layer
+ Non-Linearity
Pooling Layer
Convolution Layer
+ Non-Linearity
Pooling Layer
Fully-Connected MLP
Fully-Connected
1024
1000
1000
"2"
2×2×512 LSTM
1024
Point-wise
multiplication
Fully-Connected
Softmax
Fully-Connected
"How many horses are in this image?"

DEEP CON '18

TEŞEKKÜRLER ☺

basakbuluz@aydin.edu.tr