

**MULTIMODAL MACHINE COMPREHENSION OF
HOW-TO INSTRUCTIONS WITH IMAGES AND TEXT**

**GÖRÜNTÜ VE METİN İÇEREN ÇOK KIPLI NASIL
YAPILIR TALİMATLARININ MAKİNE İLE KAVRANMASI**

SEMİH YAĞCIOĞLU

ASSOC. PROF. DR. MEHMET ERKUT ERDEM

Supervisor

ASSOC. PROF. DR. İBRAHİM AYKUT ERDEM

2nd Supervisor

Submitted to

Graduate School of Science and Engineering of Hacettepe University

as a Partial Fulfillment to the Requirements

for the Award of the Degree of Doctor of Philosophy

in Computer Engineering

January 2023

This work titled “**Multimodal Machine Comprehension of How-to Instructions with Images and Text**” by **Semih Yağcıoğlu** has been approved as a thesis for the Degree of **DOCTOR OF PHILOSOPHY IN COMPUTER ENGINEERING** by the below mentioned Examining Committee Members.

Prof. Dr. Pınar Duygulu Şahin

Head

Assoc. Prof. Dr. Mehmet Erkut Erdem

Supervisor

Prof. Dr. Deniz Yüret

Member

Prof. Dr. İlyas Çicekli

Member

Prof. Dr. Pınar Karagöz

Member

This thesis has been approved as a thesis for the Degree of **DOCTOR OF PHILOSOPHY IN COMPUTER ENGINEERING** by Board of Directors of the Institute for Graduate Studies in Science and Engineering on ... / ... /

Prof. Dr. Salih Bülent ALTEN
Director of the Institute of
Graduate School of Science and Engineering

ETHICS

In this thesis study, prepared in accordance with the spelling rules of Institute of Graduate Studies in Science of Hacettepe University,

I declare that

- all the information and documents have been obtained in the base of the academic rules.
- all audio-visual and written information and results have been presented according to the rules of scientific ethics
- in case of using others works, related studies have been cited in accordance with the scientific standards
- all cited studies have been fully referenced
- I did not do any distortion in the data set
- and any part of this thesis has not been presented as another thesis study at this or any other university.

... / ... /

Semih Yağcıoğlu

YAYINLAMA FİKRİ MÜLKİYET HAKLARI BEYANI

Enstitü tarafından onaylanan lisansüstü tezimin tamamını veya herhangi bir kısmını, basılı (kağıt) ve elektronik formatta arşivleme ve aşağıda verilen koşullarla kullanıma açma iznini Hacettepe üniversitesine verdiğimi bildiririm. Bu izinle Üniversiteye verilen kullanım hakları dışındaki tüm fikri mülkiyet haklarım bende kalacak, tezimin tamamının ya da bir bölümünün gelecekteki çalışmalarda (makale, kitap, lisans ve patent vb.) kullanım hakları bana ait olacaktır.

Tezin kendi orijinal çalışmam olduğunu, başkalarının haklarını ihlal etmediğimi ve tezimin tek yetkili sahibi olduğumu beyan ve taahhüt ederim. Tezimde yer alan telif hakkı bulunan ve sahiplerinden yazılı izin alınarak kullanması zorunlu metinlerin yazılı izin alarak kullandığımı ve istenildiğinde suretlerini Üniversiteye teslim etmeyi taahhüt ederim.

Yükseköğretim Kurulu tarafından yayınlanan **“Lisansüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına İlişkin Yönerge”** kapsamında tezimin aşağıda belirtilen koşullar haricince YÖK Ulusal Tez Merkezi / H. Ü. Kütüphaneleri Açık Erişim Sisteminde erişime açılır.

- ☐ Enstitü yönetim kurulu kararı ile tezimin erişime açılması mezuniyet tarihimden itibaren 2 yıl ertelenmiştir.
- ☐ Enstitü yönetim kurulu gerekçeli kararı ile tezimin erişime açılması mezuniyet tarihimden itibaren ay ertelenmiştir.
- ☐ Tezim ile ilgili gizlilik kararı verilmiştir.

... / ... /

Semih Yağcıoğlu

ABSTRACT

MULTIMODAL MACHINE COMPREHENSION OF HOW-TO INSTRUCTIONS WITH IMAGES AND TEXT

Semih Yağcıoğlu

Doctor of Philosophy , Computer Engineering

Supervisor: Assoc. Prof. Dr. Mehmet Erkut Erdem

2nd Supervisor: Assoc. Prof. Dr. İbrahim Aykut Erdem

January 2023, 166 pages

In the blink of an eye, we understand what we are looking at. Most of our brain is organized to process the visual information we receive; thus, replicating human intelligence requires a complete understanding of human vision. But, is understanding vision enough to understand human intelligence? Probably not. Besides our visual perception skills, language is an essential and unique ability and a natural way of communication for humans.

For thousands of years, humankind has been telling stories and giving instructions through spoken language. One of the earliest written forms of language is instructions, specifically food recipes. These instructions not only help us understand what the people of that time ate but also teach us how they used to live their lives. Instructions have been around us for centuries, be it in the form of recipes, or how-to guides, written on stone tablets or books, or else published on the web. How-to instructions with images and text are perfect candidates for understanding human intelligence, and understanding them is an important, intriguing research problem to solve. Modern how-to guides of any form almost always contain multimodal information, such as images, videos, and text. Instructions are key to

understanding and replicating a process, and how-to guides are great sources of instruction, as we can replicate the same process by just following the guide. Furthermore, how-to instructions often involve a joint understanding of multiple modalities of information *e.g.* images and text. However, they are also very challenging to understand as they often contain multimodal information such as images and text, consist of multiple objects and entities as well as require a procedural understanding of actions and interactions between such entities often referred in from one modality into another. How-to guides such as cooking recipes, typically consist of multiple steps involving various objects and entities, most of which interact with each other through different actions. Considering an action as a combination of a verb and an object or entity, being able to generalize to unseen compositions of these action compounds pose a great challenge. In this regard, understanding how visually grounded textual instructions might help models’ systematic generalization abilities remains an important research problem.

In this thesis, we examine multimodal machine comprehension of how-to instructions with images and text, review related literature, and point out current challenges. We also propose methods to address some of these challenges and ways to improve upon existing approaches. The main contributions of this thesis can be summarized as follows. We investigate machine comprehension and reasoning problems and review the previous literature to lay the grounds for understanding multimodal how-to instructions. We survey compositional generalization literature, highlight current research challenges, and discuss its relation to understanding multimodal how-to instructions. We introduce a multimodal benchmark how-to instructions dataset comprised of cooking recipes with images and text. We propose novel methods for understanding multimodal procedures. Finally, we present a challenging multimodal compositional generalization setup and propose methods to benchmark and show multimodality’s contribution to significantly improve the current state of the art in understanding multimodal how-to instructions and conclude with future research directions and discuss open challenges.

Keywords: machine comprehension and reasoning, how-to instructions, multimodality, procedural understanding, compositionality

ÖZET

GÖRÜNTÜ VE METİN İÇEREN ÇOK KİPLİ NASIL YAPILIR TALİMATLARININ MAKİNE İLE KAVRANMASI

Semih Yağcıoğlu

Doktora, Bilgisayar Mühendisliği

Danışman: Doç. Dr. Mehmet Erkut Erdem

Eş Danışman: Doç. Dr. İbrahim Aykut Erdem

Ocak 2023, 166 sayfa

Göz açıp kapayıncaya kadar, neye baktığımızı anlıyoruz. Beynimizin büyük bir kısmı, aldığımız görsel bilgileri işlemek için organize edilmiştir; bu nedenle, insan zekasını taklit etmek, görmenin tam olarak anlaşılmasını gerektirir. Ancak görmeyi anlamak, insan zekasını anlamak için yeterli midir? Muhtemelen değil. Dil, görsel algı becerimizin yanı sıra, insanlar için vazgeçilmez ve eşsiz bir yetenek ve doğal bir iletişim biçimidir.

İnsanoğlu binlerce yıldır konuşma diliyle hikayeler anlatmakta ve talimatlar vermektedir. Dilin en eski yazılı biçimlerinden biri talimatlar olup özellikle yemek tarifleri talimatların arasında ön plana çıkmaktadır. Bu talimatlar bizlere sadece o zamanın insanların ne yediğini anlamamıza yardımcı olmakla kalmaz, aynı zamanda hayatlarını nasıl yaşadıklarını da öğretir. İster tarifler şeklinde, ister nasıl yapılır yönergeleri şeklinde, ister taş tabletlere veya kitaplara yazılmış olsun, ister web’de yayınlanmış olsun, talimatlar yüzyıllardır etrafımızdadırlar. Görüntü ve metin içeren nasıl yapılır yönergeleri, insan zekasını anlamak için mükemmel adaylar olmakla birlikte, bunları anlamak, çözülmesi gereken önemli, merak uyandıran bir araştırma problemidir. Herhangi bir biçimdeki modern nasıl yapılır

yönergeleri, neredeyse her zaman resimler, videolar ve metin gibi çok kipli bilgiler içerir. Talimatlar, bir süreci anlamanın ve çoğaltmanın anahtarlarıdır ve nasıl yapılır yönergeleri, yalnızca yönergeleri izleyerek aynı süreci tekrarlayabileceğimiz için harika talimat kaynaklarıdır. Ayrıca, nasıl yapılır yönergeleri genellikle, örneğin görüntü ve metin gibi birden çok kipi ortak bir şekilde anlaşılmasını içerir. Bununla birlikte, bu talimatlar, genellikle görüntüler ve metin gibi çok kipli bilgiler içerdiklerinden, birden çok nesne ve varlıktan oluştuklarından ve genellikle bir kipten diğerine atıfta bulunulan bu tür varlıklar arasındaki eylemler ve etkileşimlerin yordamsal olarak anlaşılmasını gerektirdiğinden, anlaşılması da oldukça zordur. Yemek tarifleri gibi nasıl yapılır yönergeleri, tipik olarak, çoğu farklı eylemler yoluyla birbiriyle etkileşime giren çeşitli nesneleri ve varlıkları içeren birden çok adımdan oluşur. Bir eylemi, bir fiil ile bir nesne veya varlığın birleşimi olarak ele aldığımızda, bu eylemi oluşturan parçaların daha önce gözlemlenmemiş bileşimlerine genelleme yapabilmek büyük bir zorluk teşkil etmektedir. Bu bağlamda, görsel temelli metinsel yönergelerin, makine öğrenmesi modellerinin sistematik genelleme becerilerine nasıl yardımcı olabileceğini anlamak önemli bir araştırma sorunu olmaya devam etmektedir.

Son yıllarda yapay zeka araştırmalarına giderek artan bir ilgi oluşmuştur. Özellikle, büyük ölçekli veri kümelerinin önerilmesi, araştırmacıları daha karmaşık modeller geliştirmeye motive etmiştir. Bilgisayarlı görü ve doğal dil işleme alanlarında, örneğin görüntülerin tasviredilmesi ve görsel soruları yanıtlama gibi görevler bir çok araştırmacının ilgisini çekmiş ve bu problemler üzerinde pek çok çalışma yapılmıştır. Modellerin kalitesi zaman içinde istikrarlı bir şekilde iyileşse de, bu görevlerin doğası gereği, hem görüntü hem de metin alanlarında sorunların ortaklaşa çözülmesini gerektiren çeşitli zorluklar ortaya çıkmıştır.

Bu doğrultuda ön plana çıkan ve araştırmacıların ilgisini çeken konuların başında kavrama ve muhakeme problemleri dikkat çekmektedir. Geniş bir bağlamda bakacak olursak kavrama, bir şeyin ne anlama geldiğini idrak etme yeteneğini, muhakeme ise bilinen gerçeklerden sonuç çıkarmayı ifade etmektedir. Kavrama ve muhakeme yetenekleri üst düzey bilişsel beceriler olmakla birlikte, sadece makineler için değil insanlar için de oldukça zorlayıcı

problemler olarak kabul görmektedir. Her halükarda, makinelerin insan düzeyinde kavrama ve akıl yürütmeye ulaşması için halihazırda oldukça önemli eksikler bulunmaktadır.

Bu tez, hem görseller, hem de metinler için makine kavraması ve makine muhakemesi problemlerini incelemekte ve çok kipli muhakame ve kavrama problemlerini kapsamlı bir şekilde ele almaktadır. Ayrıca görseller ve metinleri anlamak ve akıl yürütmek için literatürde öne çıkan veri kümelerini incelemekle birlikte, daha önce önerilmiş görevleri genel bir bakışla ele almaktadır. Bu tezde, özellikle, makinelere okumayı öğretmeyi ve ardından bu konuda sorular sorarak okuduklarını anlamalarını sorgulamayı amaçlayan makine okuması ve anlaması görevlerini incelemekte, aynı zamanda görsel akıl yürütme alanında muhakeme ve çıkarım üzerine odaklanmakta, ayrıca çok kipli nasıl yapılır talimatlarının makine muhakemesi ve makine kavranması ile idrak edilmesi üzerinde çalışmaktayız. Bu problemlerin çözümü, tek kip ile dahi düşünüldüğünde oldukça zorlu problemler teşkil etmektedir. Örneğin, makine okuduğunu anlama testleri, makinelerin verilen metni ne kadar iyi anladıklarını, soruları ne kadar doğru yanıtladıklarını ölçerek değerlendirmeyi amaçlar. Görsel akıl yürütme, sahne, aktörler, varsa eylemler, bağlam, nesneler ve bunların bir biri ile olan etkileşimleri gibi görsel verilerin derinlemesine anlaşılmasını gerektirir. Görsel verileri anlama konusunda onlarca yıllık araştırma yapılmıştır. Görsel muhakeme ve anlama tipik olarak insanlarla, yani doğal dil yoluyla etkileşimi gerektirdiğinden, bu görev pratik olarak görüntüler veya videolar ve metin gibi çok kipli verileri aynı anda ele almayı gerektirmektedir. Tipik bir görsel muhakeme görevinde, bir makinenin görsel verileri anlaması, bu görsel ile ilgili sorulan soruyu kavraması ve son olarak görsel veriler ve soru bağlamında doğru bir cevap vermesi beklenir. Bu bağlamda, görüntü ve metinlerden oluşan nasıl yapılır talimatlarının makine kavranması ve muhakemesi oldukça önemli ve bir o kadar zor problemler olarak öne çıkmaktadır.

İnsanlar önceden bildikleri kavramların yeni bileşimleri ile ilk kez karşılaşmalar bile kolayca anlayabilmekte, bildikleri kavram ve nesneleri zahmetsizce bir araya getirerek yeni bileşimler oluşturabilmektedirler. Bu bağlamda, bileşimsel genelleme son yıllarda araştırmacıların ilgisini çekmekle birlikte, bu problem, çok kipli nasıl yapılır talimatlarının kavranması ve muhakemesi için oldukça önem arz etmektedir. Bu tezde, dilbilimsel

bileşimler üzerine mevcut çalışmaların kapsamlı bir incelemesini sunmakta, mevcut görevleri ve veri kümelerini sınıflandırarak tartışmaktayız. Ayrıca, sinir ağı mimarileri ve bileşimsel genelleme için önerilen öğrenme stratejilerini incelemekte, ve mevcut görevleri, veri kümelerini, yöntem ve öğrenme stratejilerini tartışarak bu alanda önerilmiş çalışmaları kapsamlı bir şekilde ele almaktayız. Bununla birlikte, sistematik genelleme alanındaki mevcut kısıtlara vurgu yaparak ve gelecekteki muhtemel araştırma istikametlerini tartışıyor ve çok kipliliğin sistematik bileşime katkısını inceliyoruz.

Yemek tariflerini anlamak ve akıl yürütmek, makinelerin yordamsal metinleri yorumlamasını sağlamaya yönelik önemli bir araştırma alanıdır. Bu doğrultuda, bu tezde ayrıca, yemek tariflerinin çok kipli olarak anlaşılması için yeni bir veri kümesi olan RecipeQA veri kümesini sunmaktayız. Sunduğumuz bu veri kümesi, kendi içerisinde başlıklar, açıklamalar ve hizalanmış görüntüler gibi birden çok kipe sahip eğitici yemek tariflerini içermektedir. RecipeQA veri kümesi üzerinde hem tek kipli hem de çok kipli modeller önermekle birlikte, metinsel boşluk doldurma, görsel boşluk doldurma, görsel sıralama, görsel uyum gibi çeşitli muhakeme ve kavrama görevleri üzerinde farklı modeller ile deneyler gerçekleştirmekteyiz. Elde ettiğimiz sonuçlar, RecipeQA veri kümesinin zorlu bir test ortamı ve makine anlama sistemlerini değerlendirmek için ideal bir kıyaslama veri kümesi olarak hizmet edeceğini göstermektedir.

Yemek tariflerinin yordamsal olarak anlaşılması, nesnelerin kavranmasını, durum değişikliklerinin izlenmesini ve zamansal ve nedensel ilişkilerin anlaşılmasını gerektirdiğinden oldukça zorlu bir görevdir. RecipeQA veri kümesini tanıttıktan sonra, farklı bir problem olarak yordamsal ortak akıl bilgisini anlama problemini araştırmaktayız. Özellikle, RecipeQA veri kümesinden yararlanan çok kipli yordamsal bilgiyi anlamak için Yordamsal Muhakeme Ağlarını (PRN) öneriyor, buna ek olarak, metinsel kiplere tamamlayıcı bir anlamsal sinyal sağlamak için görsel kiplerden nasıl faydalanılabileceği sorusunu araştırıyoruz. Önermekte olduğumuz bu model, metin talimatlarını okurken birbiriyle ilişkili varlık durumlarını dinamik olarak güncellemeyi öğrenmektedir. Ayrıca, daha önce önerdiğimiz RecipeQA veri kümesindeki görsel muhakeme görevleri üzerine bir analiz sunmakta, yordamsal çok kipli nasıl yapılır talimatlarının anlaşılması için

sunduğumuz bu yaklaşımımızın, daha önce elde ettiğimiz sonuçların doğruluğunu büyük bir farkla geliştirdiğini görmekteyiz.

Sinir ağı modelleri, pek çok farklı görevde etkileyici bir şekilde iyi performans gösterir, ancak genellikle daha önce gözlemlenmemiş kavramların bileşimsel olarak genelleştirmesinde başarısız olmaktadır. Bu doğrultuda, günlük ev görevlerinden oluşan bir veri kümesini kullanarak görsel ve metinsel bilgiye dayalı talimatlardan oluşan bir bileşimsel ve sistematik veri seti olan EK-100-SYS veri setini önermekte, bu verini kullanarak, sistematik genelleme problemini kapsamlı bir şekilde ele almaktayız. Ayrıca, bu çalışma kapsamında, bilinen kavramlardan yeni bileşimler içeren bir eylemi tahmin etmeyi amaçlayan bir görev ve eylem sınıfını tahminleme görevlerinde birkaç tek kipli ve çok kipli modeller de sunmaktayız. Elde ettiğimiz bulgular, görsel ve işitsel sinyallerden yararlanan modellerin, salt metin tabanlı modellere göre belirtilen görevlerde daha iyi sonuçlar elde edilmesine imkan sunduğunu göstermekte, bu bağlamda çok kipliliğin bileşimsel genelleme probleminde önemli bir katkı sunabileceğini göstermektedir.

Son olarak, bu tezde görüntü ve metinlerden oluşan çok kipli nasıl yapılar talimatlarının kavranması ve muhakemesi konusunda temel kısıtları ele almakta, gelecekte bu alanda yapılacak çalışmalara yön vermek için muhtemel araştırmaları istikametlerini belirtmekteyiz.

Keywords: Makine muhakemesi ve makine kavraması, nasıl yapılar talimatları, çok kiplilik, yordamsal anlama, bileşimsellik

CONTENTS

	<u>Page</u>
ABSTRACT	i
ÖZET	iii
CONTENTS	viii
TABLES	xi
FIGURES	xii
ABBREVIATIONS.....	xiv
1. INTRODUCTION	1
1.1. Scope of the Thesis	3
1.2. Contributions	9
1.3. Organization of the Thesis	10
2. MACHINE COMPREHENSION AND REASONING.....	11
2.1. Reading Comprehension	13
2.2. Visual Reasoning	14
2.3. Tasks	16
2.4. Question Types	18
2.5. Evaluation.....	21
2.6. Reading Comprehension Datasets	22
2.7. Visual Reasoning Datasets	27
3. COMPOSITIONAL GENERALIZATION	35
3.1. Background	36
3.2. Tasks	38
3.3. Datasets	41
3.4. Methods	47
3.4.1. Models.....	47
3.4.2. Learning Strategies.....	48
3.5. Research Challenges	51
4. MULTIMODAL COMPREHENSION OF COOKING RECIPES	55

4.1. RecipeQA Dataset	58
4.1.1. Data Collection.....	59
4.1.2. Questions and Answers	60
4.1.3. Dataset Statistics	61
4.2. Tasks	62
4.2.1. Textual Cloze.....	63
4.2.2. Visual Cloze	64
4.2.3. Visual Coherence.....	64
4.2.4. Visual Ordering	64
4.3. Experiments.....	66
4.3.1. Data Preparation	66
4.3.2. Baseline Models.....	67
4.3.3. Baseline Results.....	68
4.4. Related Work	69
4.5. Discussions	72
5. PROCEDURAL UNDERSTANDING OF COOKING RECIPES	74
5.1. Visual Reasoning in RecipeQA.....	77
5.2. Procedural Reasoning Networks	79
5.2.1. Input Module	80
5.2.2. Reasoning Module	81
5.2.3. Attention Module	83
5.2.4. Modeling Module	84
5.2.5. Output Module	84
5.3. Experiments.....	85
5.3.1. Entity Extraction	85
5.3.2. Training Details	85
5.3.3. Baselines	86
5.3.4. Results	86
5.4. Further Analysis	89
5.5. Discussions	90

6. COMPOSITIONAL GROUNDED UNDERSTANDING OF ACTIONS.....	94
6.1. Problem Formulation	96
6.1.1. Future Utterance Prediction Task	96
6.1.2. Action Anticipation Task	98
6.2. EK-100-SYS Dataset	98
6.2.1. Systematicity Splits (EK-100-SYS)	99
6.3. Models for Future Utterance Prediction	100
6.3.1. Text-only Unimodal Baseline (L)	101
6.3.2. Multimodal Baselines	101
6.3.3. Vision and Language (VL)	101
6.3.4. Audio and Language (AL)	103
6.3.5. Object and Language (OL)	103
6.3.6. Audio, Vision and Language (AVL)	103
6.3.7. Object, Audio and Language (OAL)	104
6.3.8. Pretrained Vision and Language (PVL)	104
6.4. Models for Action Anticipation	104
6.5. Experimental Setup	105
6.5.1. Evaluation Metrics	105
6.5.2. Training Details	105
6.6. Results	106
6.6.1. Future Utterance Prediction	106
6.6.2. Action Anticipation	107
6.7. Related Work	108
6.8. Experimental Setup	110
6.9. Discussions	111
7. CONCLUSION	113
7.1. Future work and open directions	115

TABLES

	<u>Page</u>
Table 2.1 List of Comprehension and Reasoning Datasets.....	12
Table 4.1 RecipeQA dataset statistics.	62
Table 4.2 Results for simple and neural models	68
Table 4.3 Comparison of multimodal machine comprehension datasets	69
Table 5.1 Comparison of PRN against baselines	87
Table 5.2 Training hyperparameter selection	90
Table 6.1 Future utterance prediction results	106
Table 6.2 Object-level attention accuracy	106
Table 6.3 Action anticipation results	108
Table 6.4 Model sizes and their training times.....	110
Table 6.5 Comparison of baselines in the action anticipation task for verbs	111
Table 6.6 Comparison of baselines in the action anticipation task for nouns	111

FIGURES

	<u>Page</u>
Figure 1.1 A recipe with both images and text	4
Figure 1.2 Different states of ingredients	6
Figure 1.3 A sweet potato recipe with 6 steps	7
Figure 1.4 Illustration of compositional generalization problem	9
Figure 2.1 An example story, and question-answer pair taken from Daily Mail dataset.....	14
Figure 2.2 Semantic segmentation from MSCOCO dataset.....	15
Figure 2.3 A natural image from MSCOCO and question-answer pairs from VQA dataset	16
Figure 2.4 A set of comprehension tasks from bAbI and VQA datasets	17
Figure 2.5 Yes/No question type	18
Figure 2.6 Multiple choice question type	19
Figure 2.7 W question type	19
Figure 2.8 Open-ended question type	20
Figure 2.9 Textual cloze style question	20
Figure 2.10 Visual cloze style question.....	21
Figure 2.11 Pointing question type	21
Figure 2.12 Factoid question-answer pair from QASENT	23
Figure 2.13 Examples from MCTest dataset	24
Figure 2.14 A story and question answer pair from CNN dataset.....	25
Figure 2.15 Example story from CBT dataset.....	26
Figure 2.16 Example story and a set of question-answer pairs from the SQuAD ...	27
Figure 2.17 A passage and related questions and their answers from NewsQA.....	28
Figure 2.18 Examples images and question-answer pairs from DAQUAR dataset..	29
Figure 2.19 A set of example images and descriptions from Visual Madlibs dataset	29
Figure 2.20 Examples from VQA dataset	30

Figure 2.21	A set of examples from FM-IQA dataset	30
Figure 2.22	Examples of grounded images and question-answer pairs from Visual7W dataset.....	31
Figure 2.23	Examples images and question-answer pairs from MovieQA dataset..	31
Figure 2.24	An examples image and generated questions from CLEVR dataset....	32
Figure 2.25	A few examples from Comics dataset.....	33
Figure 2.26	Multi-modal machine comprehension overview from TQA dataset....	33
Figure 3.1	Example from Table Lookup task	39
Figure 3.2	Example from gSCAN dataset	43
Figure 3.3	Illustration of a visual scene sample from CLEVR dataset	45
Figure 3.4	Illustration of EPIC-KITCHENS dataset	46
Figure 4.1	An illustrative text cloze style question	57
Figure 4.2	A recipe of ‘Creamy Coconut Chickpea Curry’ recipe with 9 steps....	59
Figure 4.3	Distribution of food categories across RecipeQA.....	62
Figure 4.4	Word clouds of the tokens of the recipes in RecipeQA	63
Figure 4.5	Sample question types from RecipeQA training set.....	65
Figure 4.6	Sample nearest neighbors.....	66
Figure 4.7	Sample groundtruth and model prediction results for a textual cloze style question	70
Figure 4.8	Sample question types taken from the RecipeQA test set.....	71
Figure 5.1	A recipe for preparing a cheeseburger, illustrating instructions and entity interactions	75
Figure 5.2	An illustration of our Procedural Reasoning Networks (PRN)	78
Figure 5.3	Self-attention weights between entities and instructions	81
Figure 5.4	t-SNE visualizations of learned embeddings	87
Figure 5.5	Step-aware entity representations	88
Figure 6.1	Systematicity setup overview for future utterance prediction task.....	97
Figure 6.2	Overview of multimodal Audio, Vision and Language baseline model	100
Figure 6.3	Future Utterance Prediction qualitative results	107
Figure 6.4	Atom and compound distributions for EK-100-SYS splits	110

ABBREVIATIONS

RNN	: R ecurrent N eural N etworks
CNN	: C onvolutional N eural N etworks
LSTM	: L ong S hort T erm M emory
GRU	: G ated R ecurrent U nits
DL	: D eep L earning
RC	: R eadin g C ompre h ension
MRC	: M achin e R eadin g C ompre h ension
VR	: V isual R easoning
MMC	: M ultimodal M achin e C ompre h ension
NLP	: N atural L anguag e P rocessin g
CV	: C omputer V ision
NLQ	: N atural L anguag e Q uestions
NLU	: N atural L anguag e U nderstanding
RGBD	: R ed G reen B lue D ept h
MT	: M achin e T ranslation
NMT	: N eural M achin e T ranslation
PRN	: P rocedural R easoning N etworks
QA	: Q uesti o n A nsw e ring
VQA	: V isual Q uesti o n A nsw e ring
AA	: A cti o n A nticipati o n
FUP	: F utur e U tt e rance P redic t ion

1. INTRODUCTION

In order to understand what is going around us we just need to look, for merely a second. In this short amount of time, the visual information flows through our eyes and through the lateral geniculate nucleus, a relay center located in the thalamus, and then reaches the visual cortex. The first cortical areas decode lines, edges, and contours. The parietal cortex takes the visual signals to locate where things are in 3-dimensional space. The inferotemporal cortex recognizes shapes, objects, and faces. In the blink of an eye, we understand what we are looking at. Humans, and in general animals are very good at understanding what they see. Seeing is merely a thing for us, we only need to look in order to see, as it comes off the shelf. Nonetheless, this is not the case for machines. For them to recognize what they see, one obvious thing to do is to replicate animal vision. This, however, is not enough to understand our visual world, but an important milestone for machines. In order to understand what they see, machines have to learn, through vigorous training, of shapes, edges, geometry, scale, occlusion, time and context, and maybe and more importantly their combinations, just like humans. There have been decades of work to understand the vision. One of the earliest works in understanding vision is Marr's approach to treating vision as an information processing system. In his work, Marr proposed three levels for a machine that carries out an information processing system, such as a vision task [1]. The first level is the computational theory, the second is representation and algorithm, and finally, the third level is hardware implementation each of which refers to what a system does, how does it do what it does and how the system is physically realized, respectively. Marr also proposed a representational framework for vision with three stages. In the first stage, that is the primal sketch, edges, bars, ends, and blobs are represented. In the second stage, which is a 2.5D sketch, the textures are acknowledged. Finally, in the third stage, the scene is visualized in a continuous, 3-dimensional map. Marr's theory kindled interest in understanding how our visual system works, along with a framework to process visual data. Visual perception and understanding how our visual system is important, as most of our brain is organized to process the visual information we receive, and thus, replicating human intelligence requires a

complete understanding of human vision. But, is understanding vision enough to understand human intelligence? Probably not. Besides our visual perception skills, language is also an important and unique ability and a natural way of communication for humans. Vision is about how we see the world, but language is about how we communicate and interact with it. Among other things, language is maybe what makes the human species superior in the whole animal kingdom. Although there is no consensus on how, why, when, and where language might have emerged [2], there are about 7000 languages spoken around the world [3]. Languages emerge, live, and die with people. Shaped with historical events, languages capture almost every aspect of human life, at the time they are being spoken. Although languages have been with us for thousands of years, humans started writing much later. With the invention of written language, humans have started to record everything around them. Through different mediums, we have produced and captured knowledge for thousands of years in written form. This leads to an accumulation of human knowledge, and today we are generating more than ever, adding up to that knowledge. Representing a language through graphic means, *i.e.* writing enables readers of that language to reconstruct the encoded contents. Reading in this regard is a very important tool for humans to acquire information and comprehend what is written. Understanding language is a very challenging task, as it often takes several years for an infant to learn a language from scratch. For thousands of years, mankind has been telling stories and giving instructions through spoken language. With the invention of written language, people started to write down what they deem to be important. Writing in this regard, helped preserve human knowledge, and subsequently to pass this knowledge from one generation to the next generation as well as to share it with different cultures. One of the earliest written forms of language is instructions, specifically food recipes. Nettle pudding, recently discovered in the UK is, to date, one of oldest recipes known in the world [4]. The findings indicate that the recipe was from circa 6000 BC and was possibly dating back to 8000 BC. This wild plant often found in the woods was an important source of food for the people of that era. The uncovered recipe gives directions to get rid of the famous stings of the nettle by boiling them in hot water and then preparing a mixture with barley.

Food recipes are not the only instructional information humans have preferred to record in the early ages. They also wanted to record other aspects of their life. One of the earliest surviving manuals known to date is discovered in modern Turkey in 1906, which is about how to train Hittite chariot horses and is from circa 1350 BC [5]. In this text, Kikkuli, a master horse trainer from the land of Mitanni, gives detailed instructions to follow each day to raise the best chariot horses, such as the training program, the food and water rations, *etc.* These instructions not only help us understand what the people of that times eat, or how they work but also give us how they used to live their lives by providing contextual information, how they collect food, preserve it, or what methods they follow to train their horses and most importantly give us a piece of adequate information to how to replicate that process. Modern how-to instructions have been using both images and text for many years now, such as the recipes in cookbooks or the how-to guides shared across the web. These instructional guides often use visual information, such as images or videos to aid textual directions. Although there are instructional guides with only one modality, be it text, videos, or images, the vast majority of modern how-to guides shared across the web are in a multimodal form that co-occurs with text and images at the same time often with a visual that illustrates the process described in the text.

For instance, in Fig. 1.1, the end product of the recipe is illustrated on the top left and ingredients are illustrated on the top right both as images and at the bottom the directions to follow as well as the ingredients are given. As seen from Fig. 1.1, different modalities are often used to aid the other modality.

1.1. Scope of the Thesis

Throughout this thesis, we will introduce the problem of comprehending multimodal how-to instructions from images and text. First and foremost, we will make an emphasis on why this problem is important and worth solving as well as possible ways to address the main challenges about it. There are a handful of reasons why how-to instructions with images and text are interesting to work on, and why understanding them is an interesting research



To make this salad, in the jar, put items in this order:

- Strawberry-Lime Vinaigrette (see recipe below—) use about two tablespoons of dressing per pint-sized jar)
- Cooked Quinoa
- Sliced Strawberries
- Sunflower Seeds
- Sliced Green Onions
- Crumbled Feta
- Baby Spinach

Depending on how juicy your strawberries are, you might want to put them down lower in the jar (before the quinoa) to keep them even further away from the spinach.

Figure 1.1 A recipe with both images and text tells almost everything one would need about how to prepare food. A Feta salad in a jar recipe taken from [6]. On the top left the final state is illustrated, top right shows the ingredients annotated with text. At the bottom, we see the directions to follow.

problem to solve. Instructions have been around us for centuries, be it in the form of recipes, or how-to guides, written on stone tablets or books, or else published on the web. Thus, there is a good chance that they will still be useful sources of information in the future as well. A second interesting aspect of understanding images and text in the form of instructions is that they explicitly contain the information to replicate a process from scratch. This is especially important for three reasons, first, the instructions are almost always written with a clear purpose, that is to replicate a process, therefore contain only the specific information to do so. Second, understanding the required steps of a process is an important milestone for machines because step-by-step instructions inherently divide a process into smaller chunks with each chunk followed by a successor chunk. The third interesting aspect of this problem is that modern how-to guides of any form almost always contain multimodal information, such as images, videos, and text. In particular, images and text often follow each other in subsequent steps, with no particular order. This is challenging but an interesting part of the problem as having multiple modalities might be useful to address some of the challenges that surface around this problem, yet this might give birth to newer challenges such as the possible ambiguities, biases, or semantic gaps that are introduced with modality shifts. Moreover, having multiple modalities brings another challenge, that is understanding the coreferences made in the text might refer to either the predecessor or the successor image or images. Another problem might be to skip mentioning an instruction in one modality and expect the reader to fill in this logical gap by inferring what is missing from the information provided from another modality, *e.g.* having a chopped tomato image and failing to mention chopping the tomatoes in the text and just continuing with “put the tomatoes in the pan”. For this particular case, we need to be able to make a coreference resolution but the problem here can not be simply solved by just mapping the word “tomatoes” to the tomatoes in the images, as the reader is forced to make a reasoning that the tomatoes should have been chopped because the tomatoes are different from the previous tomato images, thus they might have been transformed *i.e.* chopped. Nevertheless, failing to mention this in one modality forces the reader to fill in the semantic gap using only one modality.

¹<http://www.instructables.com/id/How-to-De-Seed-a-Tomato>

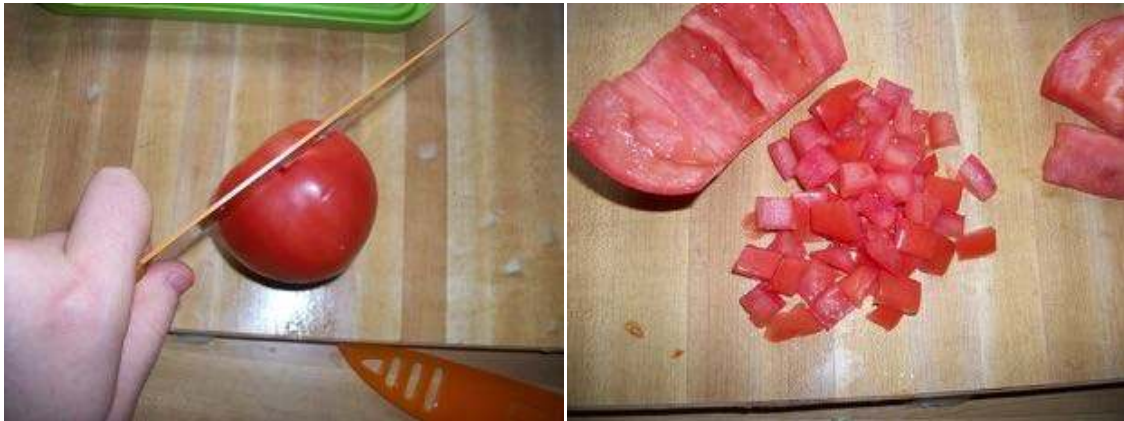


Figure 1.2 Different states of tomato, taken from Instructables.com ¹

One particular challenge we also observe in how-to instructions, is that often objects are referred to with no states, such as a tomato in good shape is cut into smaller parts, but still mentioned as a tomato. This is because people seem to refer to objects even though they change states, nevertheless constituting a problem as the text does not completely describe the image, as also illustrated in Fig. 1.2.

Another challenge that needs to be addressed while tackling this problem is that how-to guides might have directions in them telling the same story but in a different modality, although this might be seen as complementary for a direction in different modalities, the change in visual information and the change in textual information often can not be compared. For instance in a recipe, for an image that displays the ingredients as a first figure and as a list of those ingredients as the first textual direction, the next step might be to mix some of those ingredients, which would lead to a very different image from the previous image, but for the text modality, the direction might simply be “mix the pepper with the mince”. This is problematic for two reasons, first, the instructions in text modality do not simply describe the image but the image might be a product of the actions described in the textual directions. Therefore, one needs to understand and infer that following a set of instructions would transform the image to the subsequent image, although the images are not their immediate successors, but might have textual steps between them.

Alignment between images and text is a big problem in understanding how-to guides. This

is mainly because how-to instructions do not follow a fixed structure which makes them even harder to process, and finally leads to problems in aligning one modality with the other. In Fig. 1.3, a how-to instruction, in the form of the cooking recipe is taught in 6 steps but the first set of images demonstrates the final state of the how-to but not the initial state. Thus, this alignment should be considered while dealing with the sequences in the tutorial.

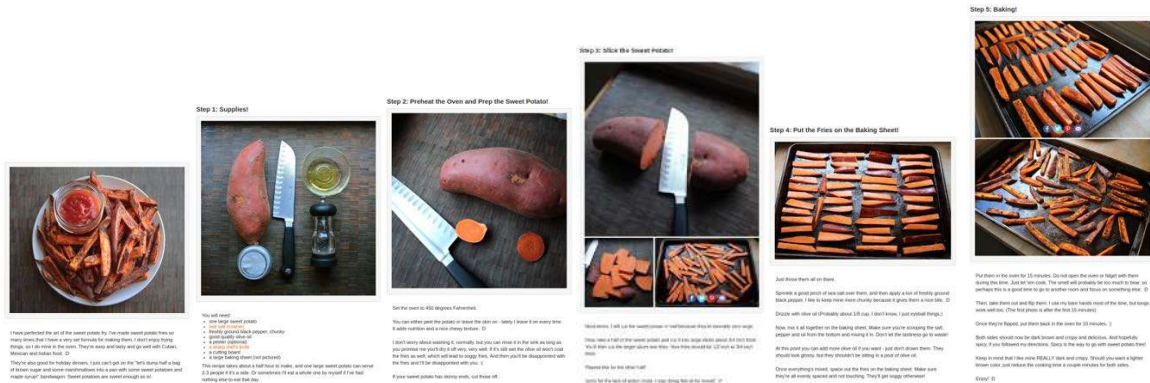


Figure 1.3 A sweet potato recipe with 6 steps, taken from Instructables.com (<http://instructables.com/id/Perfect-Oven-Sweet-Potato-Fries>)

In Fig. 1.3, we see a recipe written in 6 steps. Each of these steps contains at least one or more images related to the story told in the text in that step, which leads to some structure and a degree of alignment between images and text. Nevertheless, considering the vast amount of recipes on the web, we observe varying structures between the websites, or even through each instruction. Thus, it is a challenge to understand which part of the recipe is related to which image or images as well as to extract ingredients and tools in the recipes.

Natural language understanding and visual understanding have been two major challenges in the artificial intelligence field and there have been decades of work in both domains. Much recently, language and vision communities started working on problems that require a joint understanding of images and text. Most mediums, such as the web, books, emails, *etc.* have both images and text nested in each other. Thus, understanding how-to instructions often involve a joint understanding of multiple modalities of information *e.g.* images and text.

Clearly designed instructions are key to understanding and replicating a process. How-to guides in this sense, are great sources of instruction as people can replicate the same process

by just following the guide. Moreover, how-to guides are quite an interesting source of information as there are infinitely many instructions in the world, considering there is typically more than one way to do something, and in this regard, there might be infinitely many variations for almost any how-to guide. Instructional guides are also very challenging to understand, on the one hand, they often contain multimodal information such as images, and text in them, on the other hand from the beginning to the end of a how-to often there is very little structure or no structure at all. Additionally, the images in the how-to's might show various forms of the objects, tools, or ingredients, therefore implicit and explicit knowledge of the text might be required to understand the transformation or relationship between each image. This is also true for the text as the author of the guide might have chosen to skip some of the vital information in the text but expect the reader to infer this information from the visual content. In general, there are several challenges to this problem, but this problem is quite interesting firstly because it is a real-world problem, *i.e.* almost everyone has read a how-to before or tried to replicate it from scratch, and second, how-to guides have instructional information for adults to replicate a process in an often multimodal form which makes it even more interesting as understanding instructions would be a crucial goal for machines to have a generalized intelligence.

How-to guides, such as cooking recipes, typically consist of multiple steps involving various objects and entities, most of which interact with each other through different actions. Considering an action as a combination of a verb and an object or entity, being able to generalize to unseen compositions of these action compounds pose a great challenge that remains an open research problem. In particular, having been exposed to primitive elements such as “cut” and “tomatoe”, being able to comprehend a composition of those primitives *e.g.* “cut tomatoe”, is a challenging research problem as this particular compound might have never been observed during training time. Therefore, understanding and testing models’ compositional generalization abilities and exploring whether grounded textual instructions with images can help models to systematically generalize to novel compositions of actions remains an interesting challenge towards understanding multimodal how-to instructions. In this regard, understanding how visually grounded textual instructions might help models’





Inputs (images and textual instructions)		Targets (future textual instruction)
Inference		
		
		
		
	wash celery	close tap
		put down celery
		cut celery

Figure 1.4 Even though a model may be exposed to the primitives ‘wash’, ‘close’, ‘put down’, ‘cut’ and ‘celery’ during training, generalizing to the never before observed “cut celery” composition during inference pose a great challenge towards understanding how-to instructions, hence understanding whether multimodality can help models systematically generalize to unseen compositions yet remains an open research problem. (Image taken from EK100 [7].)

systematic generalization abilities, as illustrated in Fig. 1.4 yet remains an important research problem we will be investigating.

In summary, machine comprehension of how-to instructions with images and texts is a challenging but interesting real-life problem involving several tasks to be addressed. We briefly discussed a few of these challenges here and will be thoroughly addressing some of the main challenges related to this research problem throughout this thesis.

1.2. Contributions

In this thesis, we investigate multimodal machine comprehension of how-to instructions with images and text and review related work in the literature and point out current challenges toward understanding multimodal how-to’s, and propose methods to address some of these challenges. The main contributions of this thesis can be summarized as follows:

- We investigate machine comprehension and reasoning problems and review the previous literature to lay the grounds for understanding multimodal how-to instructions.
- We survey compositional generalization literature and highlight current research challenges and discuss its relation to understanding multimodal how-to instructions.

- We introduce a multimodal benchmark how-to instructions dataset comprised of cooking recipes with images and text.
- We propose novel methods for understanding multimodal procedures.
- We introduce a challenging multimodal compositional generalization setup and propose methods to benchmark and show the contribution of multimodality to significantly improve the current state of the art in understanding multimodal how-to instructions.

1.3. Organization of the Thesis

In the following chapters, we will discuss how to teach machines to understand what they read and see and to comprehend and reason by using something that has been around us for thousands of years, how-to instructions.

The organization of the thesis is as follows:

- Chapter 1 presents our motivation, contributions, and the scope of the thesis.
- Chapter 2 discusses the comprehension and reasoning problem.
- Chapter 3 reviews the compositional generalization problem and related studies.
- Chapter 4 introduces multimodal machine comprehension of cooking recipes.
- Chapter 5 demonstrates methods for procedural understanding of cooking recipes.
- Chapter 6 proposes methods for grounded compositional understanding of actions.
- Chapter 7 summarizes the thesis and discusses possible future research directions.

2. MACHINE COMPREHENSION AND REASONING

In recent years, we have seen a great deal of interest in artificial intelligence research. The availability of large-scale datasets especially in the vision and language fields enabled researchers from both Computer Vision and Natural Language Processing communities to attack problems that had never been proposed before. The availability of large-scale datasets motivated researchers to develop more complex models which could benefit from these datasets, and the progress in deep learning methods pushed the artificial intelligence and machine learning communities to tackle different problems. A variety of tasks have been proposed around these multimodal datasets [8–11]. In the vision and language domain, especially tasks such as image captioning and visual question answering have garnered a lot of interest. In the image captioning task, the main goal is to describe an image with the best possible caption. The studies around this task mainly fall into two categories, that is generative-based models, and retrieval models. The studies in the former category deal with generating new natural language descriptions from scratch [11, 12] whereas the studies that fall into the latter category aim to retrieve descriptive captions for the query image [13, 14]. Retrieval-based studies also deal with multi-modal retrieval by linking different modalities such as images and their related descriptions. The studies in this category use images as queries to retrieve a related caption among several candidates, and a natural language description to retrieve a relevant image for that description. In visual question answering [10, 15, 16], the main goal is to answer a question given an image. Although the quality of the models, improved steadily over time, due to the nature of these tasks, there have been several challenges that required solving problems jointly in both the image and text domains. For example, in the image captioning task it is problematic to measure how well the model performs. This is mainly because descriptions are hard to evaluate and how well the captions align with human judgment is a separate problem on its own. In visual question answering, although the problem is better formulated in terms of evaluation, casting the task as a classification raises other questions such as whether models actually learn to reason or just select answers which maximize a probability, even though they are out of context.

Question answering has been a well-established task in the language domain [17] and is referred to as reading comprehension, where the aim is to read a story, understand it, and finally be able to answer questions about the story [18] to measure how well a reader understands the given story. Visual question answering might be considered as an extension to this task where the narration in visual question answering is done via an image, whereas in reading comprehension narration is done via text. Understanding however is often measured in the same way. Asking a question and expecting a correct answer for the given context. Recently, comprehension and reasoning tasks have become very popular, especially in the text and image domain. Yet there is a huge gap between machines and humans for machines to achieve human-level comprehension and reasoning. Comprehension, in a broader context, refers to the ability to understand the meaning of something, and reasoning in the same regard, refers to drawing conclusions from facts. These are very high-level problems and are considered to be challenging, not just for machines but even for humans. In the following sections, we will discuss the comprehension and reasoning problem for both image and text domains and briefly describe prior efforts toward this direction, and lay the groundwork for our proposed research problem. We also provide an overview of the prominent datasets proposed in the literature for the comprehension and reasoning tasks in Table 2.1 which we will be discussing in this chapter.

Table 2.1 List of Comprehension and Reasoning Datasets

Dataset	Image Source	Question Source	Formulation	#Images	#Questions	Modality
SQuAD	-	Human	RC	-	100K	Text
MCTest	-	Human	RC	-	2640	Text
WikiQA	-	Query Logs	IR	-	3047	Text
TREC-QA	-	Query Logs	IR	-	1479	Text
CNN/Daily Mail	-	Summary + Cloze	RC	-	1.4M	Text
CBT	-	Cloze	RC	-	688K	Text
DAQUAR	NYU-Depth V2	Human	VQA	1,449	12,468	Image
Visual Madlibs	COCO	Human	VQA	10,738	360,001	Image
VQA	COCO	Human	VQA	204,721	614,163	Image
COCO-QA	COCO	Syntetic	VQA	117,684	117,684	Image
FM-IQA	COCO	Human	VQA	158,392	14,944	Image
Visual7W	COCO	Human	VQA	47,300	327,939	Image
MovieQA	Movies	Human	Movie QA	408 (Movies)	14,944	Image/Video/Text
TQA	CK12	Human	TQA	3,455	26,260	Image/Text
CLEVR	Rendered	Syntetic	VQA	100,000	999,968	Image

2.1. Reading Comprehension

Machine reading and comprehension task aims to teach machines, first how to read, and then question their understanding of what they read by asking questions about it. Machine reading comprehension in this regard aims to evaluate how well the machines comprehend the given text by measuring how accurate the machines are to answer questions, similar to the measures humans evaluate reading comprehension for human students. In this regard, using comprehension tests is appealing mainly because the performance of the reader is objectively gradable. Reading comprehension task goes back to Hirschman *et al.* [17], in which they collected a dataset of 115 reading comprehension tests that encompasses four grade levels, through third grade to sixth grade. Stories in this dataset are in the form of newspaper article and for each story there are five questions among the “who, what, when, where, why” types. In this work the Hirschman *et al.* showed that pattern matching approach could be used to select answers for “who, what, when, where, why” questions. This dataset sparked interest among researchers and their baseline was later improved by Riloff and Thelen [19] and Charniak *et al.* [20] using rule-based systems, and by Ng *et al.* [21] using a logistic regression based system. Wang *et al.* [22] used a neural network based approach on the same dataset but could not improve upon the rule-based baseline. More recently, Wang *et al.* [23] created QASENT dataset based on TREC-QA data wherein the questions are selected from TREC 8-13 QA tracks and selected sentences from question that share at least one or more non-stopwords. Richardson *et al.* [24] collected a dataset consists of 660 fictional stories for open-domain question answering. In this dataset, each story has 4 questions where the answer of each question exists in the story and the stories are limited to a level where young children would understand them, to reduce the common world knowledge required by the task. Yang *et al.* [25] created WikiQA dataset for open-domain question answering. In this dataset, there are 3047 questions sampled from Bing query logs and answers are sourced from Wikipedia passages. Weston *et al.* [26] created bAbI, a synthetic dataset for reading comprehension for a set of toy tasks. Much recently, Rajpurkar *et al.* proposed SQUAD dataset which consists of more than 100K questions posed over Wikipedia articles. Trischler

et al. [27] created NewsQA dataset from 12K CNN articles with 120K question answer pairs where where some of the questions might not have the correct answer in the related article.

Context:

The BBC producer allegedly struck by Jeremy Clarkson will not press charges against the “Top Gear” host, his lawyer said Friday. Clarkson, who hosted one of the most-watched television shows in the world, was dropped by the BBC Wednesday after an internal investigation by the British broadcaster found he had subjected producer Oisin Tymon “to an unprovoked physical and verbal attack.” ...

Query:

Producer X will not press charges against Jeremy Clarkson, his lawyer says.

Answer:

Oisin Tymon

Figure 2.1 An example story and question-answer pair taken from Daily Mail dataset [28]

Reading comprehension task is by nature challenging, mainly because of ambiguities, and biases and require a higher level of understanding and often common world knowledge to comprehend what is being read, as well as what is being asked. In Fig. 2.1, a typical setting is illustrated for the reading comprehension task where the story is given along with the query posed to the reader and the correct answer for that query. In order to answer the posed question correctly, a reader should be able to understand who the producer is, who attacked whom, and what the lawyer said, *etc.* Therefore, simple pattern-matching approaches would not suffice for this challenging task.

2.2. Visual Reasoning

Visual reasoning is a broad topic that takes its roots back from the well-known Summer Vision Project², aiming to identify objects by matching them with a vocabulary of known objects. The project is widely considered as the birth of artificial intelligence as a research field. In this regard, visual reasoning has been a major goal of computer vision from maybe

²<https://dspace.mit.edu/handle/1721.1/6125>

the very beginning which was kindled with the Summer Vision Project. Subsequently, visual reasoning requires an in-depth understanding of visual data, such as what is the scene, the actors in it, the actions, if there are any, the context, the objects, and their compositions, the relations and interactions between the actors and the objects, size, shape, orientation, geometry, scale, *etc.*



Figure 2.2 Semantic segmentation of pixels translates to understanding the correct labels of visual data. On the left, the original image is taken from MSCOCO dataset [9], and on the right semantic annotations are drawn onto the same image.

There have been decades of research done in understanding visual data. Visual understanding, in this regard often referred as to understanding the correct labels of pixels through semantic annotations as also illustrated in Fig. 2.2. But, to completely understand a scene we should be able to go beyond semantic labeling. Moreover, how we measure the degree of visual understanding is a problem on its own. One possible way to do that is to be able to answer any question about it [15]. A large body of work has been amassed to date for question answering, an established task in the language domain. Recently, through visual question answering, this task has been extended to the visual domain by replacing the passage or story in textual counterpart with an image which translates to changing the modality of the context from text to visual data [15]. Latest efforts in this direction sparked a lot of interest in visual reasoning and comprehension [10, 15, 16]. As visual reasoning and comprehension typically require interaction with humans, that is through natural language, this task practically requires dealing with multimodal data such as images or videos and text at the same time. In a typical visual reasoning task, a viewer is expected to understand the

visual data as well as to comprehend the question about it, and finally be able to provide an answer within the context of the visual data and the question. Such a scenario is illustrated in Fig. 2.3.

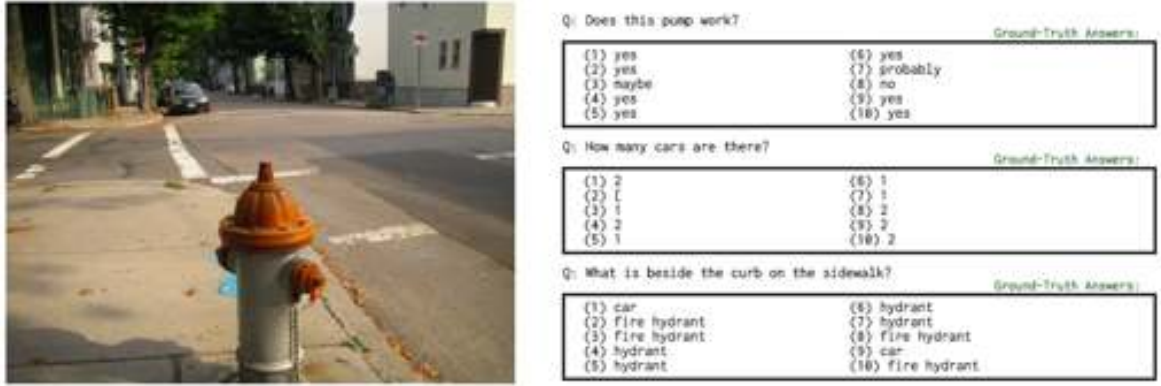


Figure 2.3 A natural image on the left taken from MSCOCO dataset [9] along with questions and their ground-truth answers from VQA dataset [10].

2.3. Tasks

Reading comprehension tasks requires understanding the given context, be it a story, a dialogue, or a paragraph, and subsequently being able to query that understanding by providing answers for what is being questioned. In particular, the query performed over the previous understanding would require the reader’s understanding of the given context in a variety of tasks such as assessing the ability to count, list, or else reasoning about the position and size *etc.* Visual reasoning is a challenging concept wherein a viewer needs to understand all the parts of the visual data, but also to understand the semantics of compositions of scenes, objects, entities, and events create. For the reading comprehension and visual reasoning tasks, a reader/viewer might need different forms of reasoning. We adopted and stratified a set of reasoning types from Trischler *et al.* [27] and Chen *et al.* [29] for the text domain and extended and transformed these types for multimodal comprehension of images and text.

- **Exact Match:** The answer is self-evident and can be found in the given context by just matching the surrounding words.

- **Paraphrasing:** The answer can be easily identified by paraphrasing the question.
- **Inference:** The answer can not be found easily by matching the available semantic information but must be inferred from the conceptual overlap of partial clues in the context.
- **Coreferencing:** The answer can not simply be matched but must be coreferenced from the available information.
- **Synthesis:** The answer can not be extracted from the context easily but must be synthesized from the available components.
- **Non-existent:** The answer can not be derived from the given context and common world knowledge is required.

Task 1: Single Supporting Fact Mary went to the bathroom. John moved to the hallway. Mary travelled to the office. Where is Mary? <i>A: office</i>	Task 2: Two Supporting Facts John is in the playground. John picked up the football. Bolt went to the kitchen. Where is the football? <i>A: playground</i>	Task 11: Basic Coreference Daniel was in the kitchen. Then he went to the studio. Sandra was in the office. Where is Daniel? <i>A: studio</i>	Task 12: Conjunction Mary and Jeff went to the kitchen. Then Jeff went to the park. Where is Mary? <i>A: kitchen</i> Where is Jeff? <i>A: park</i>
Task 3: Three Supporting Facts John picked up the apple. John went to the office. John went to the kitchen. John dropped the apple. Where was the apple before the kitchen? <i>A: office</i>	Task 4: Two Argument Relations The office is north of the bedroom. The bedroom is north of the bathroom. The kitchen is west of the garden. What is north of the bedroom? <i>A: office</i> What is the bedroom north of? <i>A: bathroom</i>	Task 13: Compound Coreference Daniel and Sandra journeyed to the office. Then they went to the garden. Sandra and John travelled to the kitchen. After that they moved to the hallway. Where is Daniel? <i>A: garden</i>	Task 14: Time Reasoning In the afternoon Julie went to the park. Yesterday Julie was at school. Julie went to the cinema this evening. When did Julie go after the park? <i>A: cinema</i> When was Julie before the park? <i>A: school</i>
Task 5: Three Argument Relations Mary gave the cake to Fred. Fred gave the cake to Bill. Jeff was given the milk by Bill. Who gave the cake to Fred? <i>A: Mary</i> Who did Fred give the cake to? <i>A: Bill</i>	Task 6: Yes/No Questions John moved to the playground. Daniel went to the bathroom. John went back to the hallway. Is John in the playground? <i>A: no</i> Is Daniel in the bathroom? <i>A: yes</i>	Task 15: Basic Deduction Sheep are afraid of wolves. Cats are afraid of dogs. Mice are afraid of cats. Gerrardo is a sheep. What is Gerrardo afraid of? <i>A: wolves</i>	Task 16: Basic Induction Lily is a swan. Lily is white. Bernhard is green. Grog is a swan. What color is Grog? <i>A: white</i>
Task 7: Counting Daniel picked up the football. Daniel dropped the football. Daniel got the milk. Daniel took the apple. How many objects is Daniel holding? <i>A: two</i>	Task 8: Lists/Sets Daniel picks up the football. Daniel drops the newspaper. Daniel picks up the milk. John took the apple. What is Daniel holding? <i>milk, football</i>	Task 17: Positional Reasoning The triangle is to the right of the blue square. The red square is on top of the blue square. The red sphere is to the right of the blue square. Is the red sphere to the right of the blue square? <i>A: yes</i> Is the red square to the left of the triangle? <i>A: yes</i>	Task 18: Size Reasoning The football fits in the suitcase. The suitcase fits in the cupboard. The box is smaller than the football. Will the box fit in the suitcase? <i>A: yes</i> Will the cupboard fit in the box? <i>A: no</i>
Task 9: Simple Negation Sandra travelled to the office. Fred is no longer in the office. Is Fred in the office? <i>A: no</i> Is Sandra in the office? <i>A: yes</i>	Task 10: Indefinite Knowledge John is either in the classroom or the playground. Sandra is in the garden. Is John in the classroom? <i>A: maybe</i> Is John in the office? <i>A: no</i>	Task 19: Path Finding The kitchen is north of the hallway. The bathroom is west of the bedroom. The den is east of the hallway. The office is south of the bedroom. How do you go from den to kitchen? <i>A: west, north</i> How do you go from office to bathroom? <i>A: north, west</i>	Task 20: Agent's Motivations John is hungry. John goes to the kitchen. John grabbed the apple there. Daniel is hungry. Where does Daniel go? <i>A: kitchen</i> Why did John go to the kitchen? <i>A: hungry</i>
 What color are her eyes? What is the mustache made of?	 How many slices of pizza are there? Is this a vegetarian pizza?	 Is this person expecting company? What is just under the tree?	 Does it appear to be rainy? Does this person have 20/20 vision?

Figure 2.4 A set of reading comprehension tasks and their examples described on bAbI dataset, taken from [30] on the top five rows, and on the bottom row a few example question answer pairs given for a typical visual reasoning task taken from VQA dataset [10].

In Fig. 2.4, a list of comprehension and reasoning tasks are given that assess a variety of abilities, such as reasoning about the size of things, making inductions between actions,

inferring when they happen, referring to them and making deduction of what a good sight means, *etc.*

Comprehension and Reasoning as we discussed so far, are challenging tasks and due to the nature of this problem, for both image and text domains, using RNNs including their variants such as LSTM or GRU and CNNs does not simply work well in practice, for this task. In particular, comprehension and reasoning as we discussed need to deal with complex information and use different types of reasoning, thus we need specific mechanisms to account for the abilities a machine would need to answer such questions which we will be discussing under relevant sections.

2.4. Question Types

In this section, we briefly go over the types of questions used commonly in comprehension and reasoning tasks. Each of these question types is often posed to query various forms of reasoning as well as abilities as we discuss in the previous Chapter 2.3..

True/False Questions. In True/False type questions the reader/viewer is expected to provide an answer by stating whether the facts in the question form a true statement or else false. Another variation of True/False type questions is Yes/No questions. In the case of a Yes/No type question, the reader/viewer is given a context (text/visual) and asked to assess whether the statement is true or false but the formulation of the question is slightly different from the original true/false statement as the answer should be in yes/no form.

John is in the playground.
Daniel picks up the milk.

Q: Is John in the classroom?

A: No

Q: Does Daniel have the milk?

A: Yes

Figure 2.5 Yes/No question and answer pairs taken from bAbI dataset [31]

Multiple Choice. One common way to assess a reader's/viewer's understanding of a context is to use multiple-choice tests. In multiple-choice type questions, the reader is given a set of options to choose from. In such a question the aim is to select the correct answer from the pre-selected answers [32–34].

Question: Where did James go after he went to the grocery store?

- A) his deck
- B) his freezer
- C) a fast food restaurant
- D) his room

Figure 2.6 A multiple choice question answer pair is selected from MCTest dataset [24]

These types of questions are relatively easier to answer as there are only a few options. Thus, as a precaution, the reader's ability to comprehend might be compared against random selection.

W Questions. In this type of question, which is often referred to as W questions, the questions are cast as Who, Where, What, When, Why, Which, and How, and the reader is asked to provide an answer that requires understanding a story, and actors in that story, as well as their relations while keeping track of different states in the given story. In W-type questions, typically Who questions assess the ability to find the related actor in a context, be it textual or visual. Such as 'who went to the kitchen', or 'who has a mustache in the image'. Where questions assess positional understanding, 'What' type of questions demand an understanding of the objects or entities in the context, 'When' type of questions evaluates the temporal understanding. 'How' type of questions are typically harder to answer as one needs to understand the relations and compositions in the context but also needs to make inferences using them.

Sam walks into the kitchen. Sam picks up an apple. Sam walks into the bedroom. Sam drops the apple. Q: Where is the apple? A. Bedroom	Brian is a lion. Julius is a lion. Julius is white. Bernhard is green. Q: What color is Brian? A. White	Mary journeyed to the den. Mary went back to the kitchen. John journeyed to the bedroom. Mary discarded the milk. Q: Where was the milk before the den? A. Hallway
--	--	---

Figure 2.7 Three W question and answer pairs taken from bAbI dataset [31]

Open Ended Questions In the open-ended type of questions, the answer might be in a free form and thus has no restrictions. The answers in open-ended question types might vary from one word to several words, and thus might be considered more realistic and challenging. In Fig. 2.8 an open-ended type question-answer pair is illustrated.

Passage: As you can imagine, it costs a lot of money to fix a head gasket leak because the mechanic is going to have to remove the head off the top of the engine, which will take a lot of labor hours. A repair like this can easily cost between \$1,000 and \$2,000. Another common source of a leak is the radiator. Leaks can form on several different places on the radiator, from cracks in the tanks to tubes that leak due to damage or corrosion. Average cost to replace a radiator is \$400 to \$600. Radiator caps can fail and leak, but that's definitely a do-it-yourself repair

Query: average cost to repair car radiator

Answer: Between \$400 and \$600

Figure 2.8 An open-ended question with the passage from MS MARCO dataset, with query id 10555 taken from development set [35]

Cloze Style Questions. Cloze-style questions fill in the blank type of questions. It is often obtained by randomly removing a word from a sentence and asking the reader to find what is missing given the context. These types of questions are not considered Natural Language Questions (NLQ) as they are not really questions posed by humans, but only transformed sentences in fill-in-the-blank form.

Sentence: Characters in Star Wars movies have gradually become more diverse.

Question: Characters in _____ movies have gradually become more diverse.

Answer: Star Wars

Figure 2.9 A cloze style question-answer pair is selected from CNN/Daily Mail dataset [28]

Visual cloze style question type is similar to the textual counterpart. In this question type, the reader should select the missing visual content from a set of candidates. In Fig. 2.10 a visual cloze style question is illustrated.

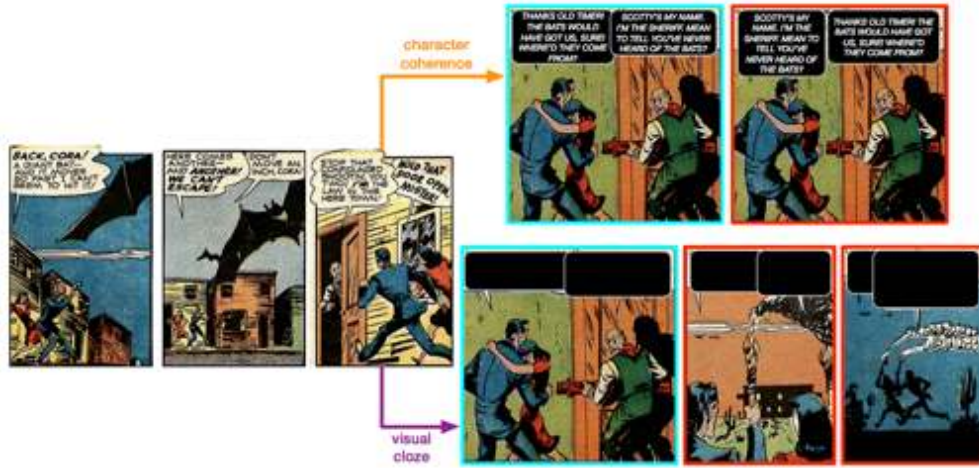


Figure 2.10 A visual cloze style question is taken from [36], in which the reader is expected to select the correct panel that would not break the sequence of panel semantically.

Pointing Type Questions. In pointing-type questions, the machine is asked to point to a location in the given context. This type of question is mostly suitable for visual questions where the machine might locate a point, but they might be adapted to other modalities as well.

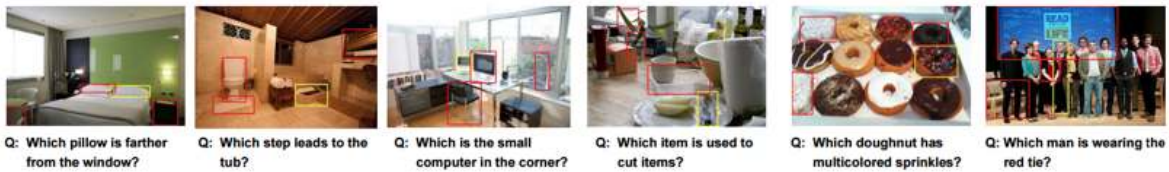


Figure 2.11 A set of examples for a pointing type question taken from Visual7W dataset [37]

For the visual context, pointing-type questions are typically asked by which question as in [37], but in general pointing type of questions might be used for circumstances one can point to, given the context is appropriate. A set of pointing-type questions is illustrated in Fig. 2.11.

2.5. Evaluation

Evaluation of machine comprehension and reasoning is extremely important as without a measure it is hard to say how well a model is doing. Evaluation of comprehension and reasoning tasks in this regard requires special attention as there are a variety of subtasks that

comprehension and reasoning models should solve and measure each of these tasks would need special care. Moreover, the nature of the context and how the models' understanding is questioned changes what is being measured. In the following, we will briefly describe the common evaluation methods for comprehension and reasoning-related tasks.

- **Accuracy:** Accuracy is the most widely used metric to evaluate the performance of comprehension and reasoning tasks. Accuracy is thus calculated by the ratio of correct answers to the total questions. Typically this metric is used in True/False question types, and multiple choice question types.
- **Word Match:** Word level matching aims to look at the exact string match between the answer and ground truth where only the correct matchings are used [38].
- **Automatic Metrics:** In this type of question, the evaluation is done by comparing how well the reader's answer match with the ground-truth answer by using automatic evaluation metrics for machine translation such as BLEU [39] and METEOR [40].
- **WUPS:** Wu-Palmer similarity (WUPS) [41] looks at the similarity of two words by considering the depths of their common subsequence in a taxonomy tree. The answer is considered correct if the similarity between the ground truth is above a threshold.
- **Human Judgment:** In the human judgment type of evaluation, the answers are presented to human judges along with the context and question-answer pairs. The judges then evaluate the results as in the Turing test, whether the results are coming from a machine or human [42].

2.6. Reading Comprehension Datasets

There are a variety of datasets proposed in the literature for reading comprehension tasks. Each dataset has its own merits as well as weaknesses. In the following, we will briefly discuss the publicly available reading comprehension datasets.

QASent. QASent dataset is proposed by [23] and is based on the TREC-QA data. The dataset consists of questions selected from TREC 8-13 QA tracks. The sentences in this dataset are selected to have one or more non-stopwords common with the questions.

Question: When did James Dean die?

Sentence: In 1955, actor James Dean was killed in a two-car collision near Cholame, Calif.

Answer: 1955

Figure 2.12 A factoid question and answer pair with supporting fact taken from [23]

MCTest. MCTest is a benchmark dataset for reading comprehension which contains 660 stories and each of these stories has 4 natural questions asked by humans wherein each question has 4 answers [24].

The dataset was constructed by crowd workers and each story was written by a different person and each story is independent of the rest. The dataset is split into two parts to manually correct crowdsourcing errors as MC160 and a larger one, MC500. One caveat of this dataset is that it is a small dataset to train deep learning models.

CNN/Daily Mail. CNN/Daily Mail datasets are collected from news web pages and contain real-life news articles [28]. CNN dataset contains over 90K news articles from CNN website, and on average it has 4 queries for each article resulting 380K question-story pairs. Daily Mail dataset has around 197K news articles wherein each article has 4 questions adding up to 880K story-question pairs.

Children’s Book Test (CBT). CBT dataset [43] contains stories taken from child books. The dataset is created from books that are freely available in Project Gutenberg ³. In each of the stories, there are 20 consecutive sentences and the 21st sentence is transformed into a question by removing a word from that sentence.

An example context and query from Alice in Wonderland by Lewis Carroll in the CBT dataset is illustrated in Fig. 2.15.

³<https://gutenberg.org>

Context: James the Turtle was always getting in trouble. Sometimes he'd reach into the freezer and empty out all the food. Other times he'd sled on the deck and get a splinter. His aunt Jane tried as hard as she could to keep him out of trouble, but he was sneaky and got into lots of trouble behind her back. One day, James thought he would go into town and see what kind of trouble he could get into. He went to the grocery store and pulled all the pudding off the shelves and ate two jars. Then he walked to the fast food restaurant and ordered 15 bags of fries. He didn't pay, and instead headed home. His aunt was waiting for him in his room. She told James that she loved him, but he would have to start acting like a well-behaved turtle. After about a month, and after getting into lots of trouble, James finally made up his mind to be a better turtle.

Q1) What is the name of the trouble making turtle?

- A) Fries
- B) Pudding
- C) James
- D) Jane

Q2) What did James pull off of the shelves in the grocery store?

- A) pudding
- B) fries
- C) food
- D) splinters

Q3) Where did James go after he went to the grocery store?

- A) his deck
- B) his freezer
- C) a fast food restaurant
- D) his room

Q4) What did James do after he ordered the fries?

- A) went to the grocery store
- B) went home without paying
- C) ate them
- D) made up his mind to be a better turtle

Figure 2.13 An example story and question-answer pairs related to that story is taken from MCTest [24].

There are 4 splits in the dataset, each classified as the type of the removed word from the question which are Named Entities, Common Nouns, Verbs, and Prepositions. For each question, 10 answers are selected from the story each having the same POS tag with the correct answer.

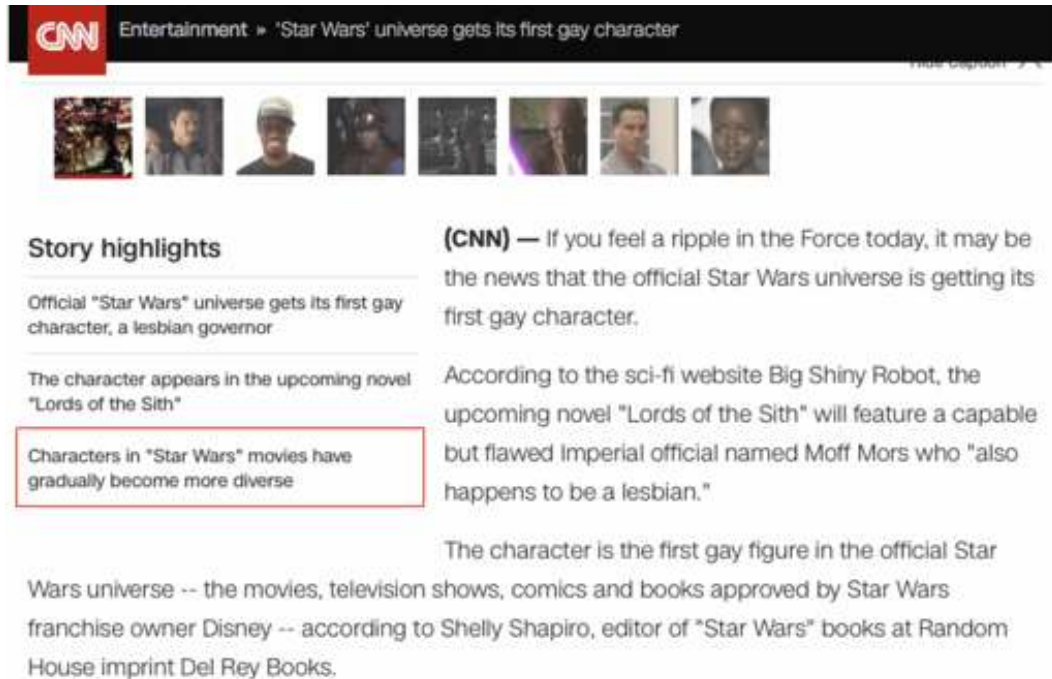


Figure 2.14 An example story taken from CNN/Daily Mail datasets where a question is selected from the highlights on the left marked as red box as cloze style [28]

Stanford Question Answering Dataset (SQuAD). SQuAD dataset [44] contains around 100K question-answer pairs obtained from 536 articles. Each story is a paragraph taken from those articles and questions are collected via crowdsourcing. The answers are also collected from crowd workers and there is more than one answer for each question. One strength of this dataset is that the questions are NLQ. The availability of multiple responses from crowd workers collected for each question provides human agreement information for answers. Moreover, the answers are open-ended, thus more challenging and realistic for reading comprehension tasks. One weakness of this dataset is the small size of the articles. Another caveat is that it uses short paragraphs for stories, which typically contain 4 to 5 sentences.

In Fig. 2.16, an example story and a set of question-answer pairs related to that story are shown from the SQuAD dataset.

NewsQA. NewsQA is a challenging benchmark dataset consisting of 12K articles collected from CNN and around 120K question-answer pairs. The question and answer pairs are

Context:

1 So they had to fall a long way .
2 So they got their tails fast in their mouths .
3 So they could n't get them out again .
4 That 's all .
5 " Thank you , " said Alice , " it 's very interesting .
6 I never knew so much about a whiting before . "
7 " I can tell you more than that , if you like , " said the Gryphon .
8 " Do you know why it 's called a whiting ? "
9 " I never thought about it , " said Alice .
10 " Why ? "
11 " IT DOES THE BOOTS AND SHOES . '
12 the Gryphon replied very solemnly .
13 Alice was thoroughly puzzled .
14 " Does the boots and shoes ! "
15 she repeated in a wondering tone .
16 " Why , what are YOUR shoes done with ? "
17 said the Gryphon .
18 " I mean , what makes them so shiny ? "
19 Alice looked down at them , and considered a little before she gave her answer .
20 " They 're done with blacking , I believe . "
Query: " Boots and shoes under the sea , " the XXXXX went on in a deep voice , " are done with a whiting " .

Candidates:

Alice—BOOTS—Gryphon—SHOES—answer—fall—mouths—tone—way—whiting

Answer: gryphon

Figure 2.15 An example story and query from Alice in Wonderland by Lewis Carroll from the CBT dataset [43]

written by humans in natural language. Questions in this dataset may not be answered as some of the questions do not have an answer in the related story. The answers may be multiple-word passages from the related story.

In Fig. 2.17, for a story collected from CNN, we see a headline, an article, and related questions and answers.

Context: One of the most famous people born in Warsaw was Maria Skłodowska-Curie, who achieved international recognition for her research on radioactivity and was the first female recipient of the Nobel Prize. Famous musicians include Władysław Szpilman and Frédéric Chopin. Though Chopin was born in the village of Żelazowa Wola, about 60 km (37 mi) from Warsaw, he moved to the city with his family when he was seven months old. Casimir Pulaski, a Polish general and hero of the American Revolutionary War, was born here in 1745.

Question: What was Maria Curie the first female recipient of?

Answers: Nobel Prize — Nobel Prize — Nobel Prize

Question: What year was Casimir Pulaski born in Warsaw?

Answers: 1745 — 1745 — 1745

Question: Who was one of the most famous people born in Warsaw?

Answers: Maria Skłodowska-Curie — Maria Skłodowska-Curie — Maria Skłodowska-Curie

Question: Who was Frédéric Chopin?

Answers: Famous musicians — musicians — Famous musicians

Question: How old was Chopin when he moved to Warsaw with his family?

Answers: seven months old — seven months old — seven months old

Figure 2.16 Example story and a set of question-answer pairs related to that story is shown from the SQuAD dataset [44]

2.7. Visual Reasoning Datasets

In the following, we discuss the most relevant datasets in the visual domain for comprehension and reasoning tasks.

DAQUAR. Dataset for Question Answering on Real-world images (DAQUAR), is one of the earliest question-answering datasets in the visual domain [15]. The dataset contains over 12K question-answer pairs of RGBD images generated by humans. DAQUAR dataset uses 1449 RGBD indoor images from NYU-Depth V2 dataset [45] with depth information and annotated semantic segmentations of 894 object classes.

Title: Soyuz Crew Endures Severe G-Forces on Re-entry

Context:

MOSCOW, Russia (CNN) – Russian space officials say the crew of the Soyuz space ship is resting after a rough ride back to Earth.

A South Korean bioengineer was one of three people on board the Soyuz capsule.

The craft carrying South Korea's first astronaut landed in northern Kazakhstan on Saturday, 260 miles (418 kilometers) off its mark, they said.

Mission Control spokesman Valery Lyndin said the condition of the crew – South Korean bioengineer Yi So-yeon, American astronaut Peggy Whitson and Russian flight engineer Yuri Malenchenko – was satisfactory, though the three had been subjected to severe G-forces during the re-entry.

Search helicopters took 25 minutes to find the capsule and determine that the crew was unharmed.

Officials said the craft followed a very steep trajectory that subjects the crew to gravitational forces of up to 10 times those on Earth.

Interfax reported that the spacecraft's landing was rough.

This is not the first time a spacecraft veered from its planned trajectory during landing.

In October, the Soyuz capsule landed 70 kilometers from the planned area because of a damaged control cable. The capsule was carrying two Russian cosmonauts and the first Malaysian astronaut. E-mail to a friend

Q1) Who was the capsule carrying?

Answer 1: two Russian cosmonauts and the first Malaysian astronaut.

Q2) Where did the Soyuz capsule land?

Answer 1: northern Kazakhstan

Q3) What distance from the target was the capsule?

Answer 1: 260 miles (418 kilometers)

Q4) What happened during the first landing?

Answer 1: landed in northern Kazakhstan on Saturday, 260 miles (418 kilometers) off its mark,

Answer 2: damaged control cable.

Q5) What was the target?

Answer not in story.

Figure 2.17 An example story and question answer pairs from NewsQA dataset [27].

In Fig. 2.18, a set of challenging examples is illustrated along with the question and answer pairs.



Figure 2.18 Example images and question-answer pairs denoted as QA and questions denoted as Q from DAQUAR dataset, taken from [15]

Visual MadLibs. Visual Madlibs dataset [16] consists of 10,738 images from MSCOCO dataset [9] with 360K focused image descriptions collected automatically with fill-in-the-blank questions.



Figure 2.19 Example images and descriptions from Visual Madlibs dataset, taken from [16]

In Fig. 2.19, a set of examples are illustrated along with their descriptions. Descriptions in each row refer to a different type of question in the dataset.

VQA. Visual Question Answering dataset [10], consists of 204,721 real images from MS COCO dataset [9] and 50K abstract scenes from [46, 47]. The dataset contains 614,163 and 150K questions for the real images and abstract scenes, respectively. In total there are around 10M answers in the first version of the VQA dataset.

In Fig. 2.20, a set of examples are illustrated from the VQA dataset for both the real images and abstract scenes.



Figure 2.20 Example images from VQA dataset, taken from [10]. For each row, images are provided at the top and below them, questions are provided in black, green, and blue responses referring to whether the answer is given by looking at the image or not, respectively.

FM-IQA.

FM-IQA dataset [42] contains 150K images from MSCOCO dataset [9] and 310K freestyle Chinese question-answer pairs, as well as their English translations for the purpose of multilingual image question-answering.



Figure 2.21 Example images and question-answer pairs in Chinese and their translations in English from FM-IQA dataset, taken from [42]

In Fig. 2.21, a set of examples are taken from the FM-IQA dataset where each image has both Chinese and English translations for question-answer pairs.

Visual7W. Visual7W dataset is a large-scale visual question-answering dataset with object groundings [37]. The dataset is collected on 47,300 MSCOCO dataset images [9] and has around 328K question-answer pairs with over 1.3M human-generated multiple-choices and over 561K object groundings from 36,579 categories. Each question in the dataset is a W question which is Who, Where, When, What, Why, How, and Which. In Fig. 2.22, examples are illustrated related to the aforementioned question types.



Figure 2.22 Examples from the Visual7W dataset taken from [37]. Each example is given with the context image and question-answer pairs along with the correct answers which are marked as green and the wrong answers, which are marked as red. At the top row, ‘Telling’ the type of questions are given where the answer modality is text. At the bottom row, ‘Pointing’ type of questions are given where the answer is a bounding box in the image, for which the bounding box is yellow and red for the correct and wrong answers respectively.

The main difference of this dataset is that it introduces a multimodal answer type for QA tasks, wherein the answer is a grounded image for the Which type of questions. For the rest of the questions, the answer is in textual modality.

MovieQA. MovieQA is a dataset that contains 15K multiple-choice questions and answers



Figure 2.23 Examples from the MovieQA dataset taken from [48]. For each example, a snapshot from a movie clip is given along with the questions and answers below the snapshot.

obtained from over 400 movies. The dataset aims to be a benchmark for story comprehension for both video and text [48].

Each question in this dataset has 5 answers, among them only one is correct. The dataset contains multiple sources of information such as movie clips, dialogs, and plots. Although some of the questions can be answered using only one modality, most of them require using multiple modalities, in particular videos and text.

CLEVR. CLEVR is a visual reasoning dataset [49] with 100K computer-rendered images and around 1M generated questions. The questions in the CLEVR dataset aim to test various visual reasoning tasks such as spatial relationship, counting, attribute identification, comparison, multiple attention, and logical operations.

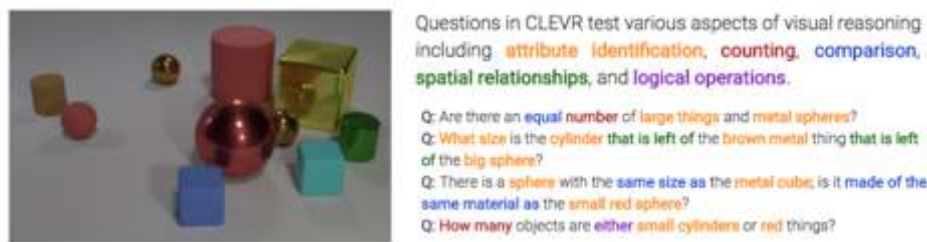


Figure 2.24 An example image and generated questions taken from CLEVR dataset [49]

In Fig. 2.24, for computer-rendered images, a set of questions are given where different coloring refers to attribute identification, counting, comparison, spatial relationships, and logical operations.

COMICS. COMICS dataset [36] consists of around 1.2M comic panels collected from comic books published during the “Golden Age” of American comics which spans from 1938 to 1954 period. These comic books are now publicly available because of copyright expiration. The dataset is created from 4000 highest-rated comic books from DCM (<http://digitalcomicmuseum.com>). The comic books are broken into smaller components and panels are extracted from each comic book leading to 1.2M panels with images and text told in a narrative manner.

TQA. Textbook Question Answering (TQA) dataset [50] was constructed from the open-source science curriculum of ck12 (<http://ck12.org>). The dataset consists of 1076 lessons from Life Science, Earth Science, and Physical Science textbooks. TQA dataset has 26,260



Figure 2.25 A few examples from Comics dataset taken from [36]

questions of which 12,567 of them with an accompanying diagram. The questions have multiple-choice answers ranging from 2 to 7 choices.

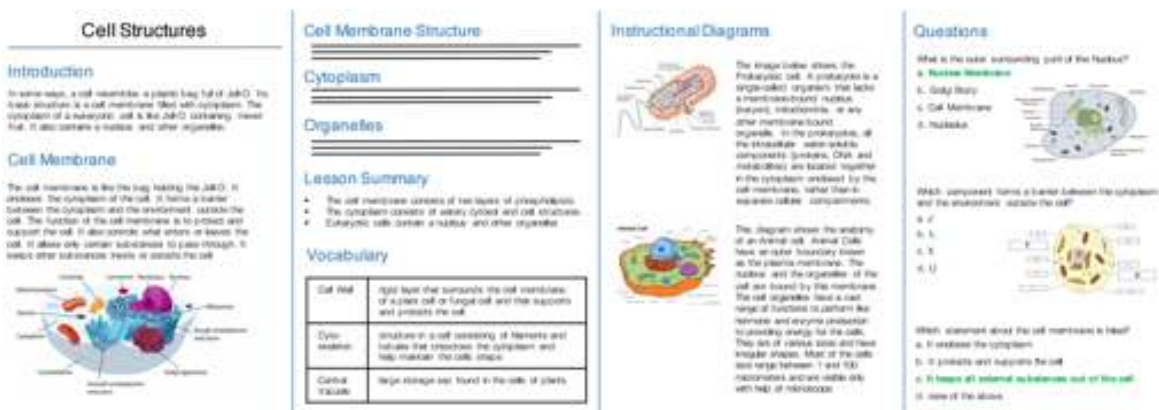


Figure 2.26 Overview of multi-modal machine comprehension from TQA dataset, taken from [50]

TQA is a challenging multi-modal dataset with various science concepts which aim to measure comprehension of both visual and textual data through the concepts taught in each lesson. The questions are divided into two categories as such whether they have a diagram accompanying them or not. Thus, a question with a diagram can only be answered with the text and the accompanying diagram. The questions with no diagram can be answered with only the given textual context. Although they are not a part of the dataset, lessons have links

to instructional videos, which might help researchers extract additional information about the lessons.

In Fig. 2.26, an overview of multi-model comprehension is illustrated where a lesson is given as a story along with textual and visual information.

3. COMPOSITIONAL GENERALIZATION

In this chapter, we survey compositional generalization literature and highlight current research challenges to lay the grounds for its connection to understanding multimodal how-to instructions based on our work⁴. We only made minor changes to fit the text in the narrative of the thesis and made small corrections in the text.

As the world we live in is inherently structured, many believe that capturing this compositional structure is a key component of artificial intelligence. In recent years, compositional generalization has become increasingly important since neural models are largely incapable of generalizing to novel compositions never observed during training. Hence, researchers have been investigating how the generalization ability of these models can be improved by exploring different aspects of compositionality. In this chapter, we provide a systematic review of the existing works on linguistic compositionality. In particular, we categorize and discuss existing tasks and datasets, and examine neural architectures and learning strategies proposed for compositional generalization. Finally, we conclude by pointing out some challenges and opportunities for future research.

Thanks to their compositional skills, humans can effortlessly understand and construct new utterances even though they encounter such novel compositions for the first time. For instance, once we understand the meaning of primitives such as “dax” and “twice”, we can easily infer what “dax twice” means [51]. Since its inception, compositionality has been studied in various aspects and sparked interest in different fields such as linguistics, philosophy, natural language processing, and computer vision. As humans can understand a potentially infinite number of novel combinations of utterances by only being exposed to a limited number of them, compositionality is regarded as a key towards solving the generalization problem. In particular, the systematicity aspect of compositionality has been an active research area as a common solution to maintaining the meanings of complex expressions by depending on their simple parts and the way those parts are organized.

⁴Under review

In the last few years, neural models have shown astounding performance on many tasks [52]. Nevertheless, there has been an ongoing debate around neural models’ ability to compositionally generalize to unseen instances. Due to their associative nature and inability to accurately represent systematic compositionality, researchers have argued that neural networks are not realistic representations of the mind [53–56]. Furthermore, despite a wealth of empirical evidence, there is little agreement on whether and how much neural networks can generalize compositionally [51, 57]. Despite the recent successes of state-of-the-art neural methods, neural networks often cannot capture the compositional structure; therefore, failing to generalize compositionally remains an open research challenge yet to be solved. This open research problem recently garnered much interest from different fields, such as language and vision. Recent work around compositional generalization sparked interest in the contribution of multimodality and grounded language processing toward systematic compositionality [58–60].

In this study, we survey the literature on compositional generalization by categorizing and discussing the existing tasks, datasets, neural models, and learning strategies proposed for improving linguistic compositionality and review related studies from the point of language generation. Moreover, we highlight current limitations and open research problems for future research directions. In particular, in Sec. 3.1., we review background work. Sec. 3.2. outlines the proposed tasks for measuring the compositional generalization abilities of learning systems. In Sec. 3.3., we underline the existing datasets in compositional generalization literature. In Sec. 3.4., we examine the prominent models and the learning strategies in the previous relevant research. Finally, in Sec. 3.5., we debate open research challenges and provide some discussions around the compositional generalization problem.

3.1. Background

Compositionality has been widely studied for many years. It studies ways to define the meaning of a sentence as a function of the meaning of its parts and the rules that are used to put those parts together. The principle of compositionality, which was introduced by Frege,

refers to the meaning of an expression being determined by the meanings of its parts [61]. It deals with describing the relationship between an unbounded number of sentences and a vast set of meanings from a finite set of rules. As a long-standing problem, compositionality has been investigated by the linguistic and philosophy communities and has been studied from different perspectives.

Compositionality was characterized by [62] from a syntactical point of view as “meaning of a whole is a function of the meanings of the parts and of the way they are syntactically combined”. From a different point of view, compositionality and systematic generalization have been viewed as standing problems to represent symbolic computation and human cognition [53, 63]. Since the beginning of the compositional generalization debate, researchers have studied the compositional abilities of neural models with mixed results (*e.g.* [51, 54, 64–71]) and little agreement on whether neural models can perform systematic compositionality or to what extent. Much recently, the compositionality problem has been investigated in various settings. [59] constructed the CLEVR-CoGenT dataset based on the CLEVR dataset to test models’ ability for compositional generalization on visual reasoning tasks. [72] investigated systematic generalization in a VQA-like setting. [73] studied systematicity and compositionality with a human-like number of examples. [74] examined picking up new concepts and applying them in test time by coupling previously learned concepts with new concepts in a meta-learning setup. [75] explored compositional generalization, to compose unseen combinations of concepts in an image captioning setting. [76] inspected the capacity of artificial neural networks in linguistic compositionality. Though neural models have been shown to generalize well across biased dataset splits, they are criticized for learning surface statistics. Towards addressing this problem, [77] described a heuristic to create similar distributions of primitives and different distributions of combinations to measure models’ compositional abilities on novel compositions. [60] explored a 2D grid world setting for situated language understanding using grounded instructions. [78] investigated compositionality in a novel word acquisition setting from narrated videos. [57] proposed an evaluation framework to test models’ compositional ability in five dimensions.

3.2. Tasks

Previous studies explored a collection of tasks to test models’ compositional ability. Below we highlight the prominent tasks from prior work to evaluate the compositional generalization properties of neural models.

Generalizing to Novel Compositions. This task is formulated to measure the compositional ability of models when all primitive components of the input have been observed during training, but the models are tested with novel compositions of primitive components – defined as *systematicity* by [57]. [51] introduced SCAN to measure the systematicity of models in a highly compositional textual navigation setting. They found that popular sequence-to-sequence models do not show systematic generalization skills. [79] inverted the SCAN task by swapping the input commands with action sequence outputs, introducing the NACS task. They confirmed the results of [51]. [77] explored a divergence-based approach and created MCD splits for SCAN, which the authors used to test models’ systematicity performance on the MCD splits.

Generalizing to Longer Sequences. [51] presented an experimental setup to test the generalization ability of recurrent sequence-to-sequence models trained on shorter action sequences while being tested against commands demanding longer action sequences using the SCAN dataset. Similarly, [57] tested the models’ generalization abilities of the PCFG-SET in a setup where the aim is to generalize to an unbounded length of sequences.

Machine Translation. [51] experimented with a machine translation (MT) task using a simple setup in which they trained a simple seq2seq model on short English-French sentence pairs that begin with English phrases such as “I am”, “he is”, “they are”, and their contractions. To examine compositionality with the introduction of a new word, the authors trained a fresh network after adding 1,000 repetitions of the sentence “I am daxy” to show that models will struggle with systematic compositionality when adding a new word to their vocabulary. [80] examined whether state-of-the-art neural machine translation (NMT) models can solve compositional tasks. As SCAN MT split only contains 8 words,

it is limited in its scope. For this, they curated a 216K-sized English-Chinese translation dataset and trained a Transformer based model. They found that even if the model has good metrics under traditional metrics, they fail to generalize compositionally. [81] claim that compositionality in natural language is much more multi-dimensional than rigid, artificially generated datasets and these datasets often focus on only strongly local, context-independent structures. As natural language involves *atomic compounds* (e.g. collocations, idioms) and global relations between words, a fine-tuned local-global approach to compositionality is indispensable. For this purpose, the authors formulate three different experimental setups for systematicity, substitutivity, and overgeneralization. Finally, the authors conclude that as natural language is not as algebraic as artificial data, a non-compositional word translation may not change the meaning of the whole. Hence, they state that MT is a more suitable task for compositionality than artificial transductions.

Table Lookup. [82] formulated the table lookup tasks to measure the compositional ability of sequence-to-sequence models. In this task, the objective is to map bit string inputs to function outputs, hence learning how to apply *functions* in a compositional setup. The authors experimented on the CTL dataset that maps atomic functions with input and output bit strings. A sample lookup task is depicted in Figure 3.1.

Atomic g			Atomic f			Composed fg		
00	\rightarrow	01	00	\rightarrow	11	00	\rightarrow	0110
01	\rightarrow	00	01	\rightarrow	10	01	\rightarrow	0011
10	\rightarrow	10	10	\rightarrow	01	10	\rightarrow	1001
11	\rightarrow	11	11	\rightarrow	00	11	\rightarrow	1100

Figure 3.1 Two atomic 2-bit lookup tasks f and g and their composition fg . Note that the composition function output starts with the output of the function g .

Image Captioning. The objective of this task is to automatically generate a description of the input image in natural language form [83]. The initial work on compositionality in image captioning research focused on triplet prediction [84], where each triplet is defined as subject-relation-object (SRO) present in a visual scene and the task is predicting unseen SRO combinations successfully in test time. [75] and [85] also focused on only systematicity,

but the task is composing unseen combinations of concepts by generating natural language captions. Recently, [86] also investigated models with a broader spectrum of compositional aspects such as productivity, substitutivity, and systematicity. [84] investigated whether state-of-the-art image captioning models that perform well on IID splits can generalize to novel compositions in terms of triplet predictions. The authors created a compositional split from MSCOCO dataset [87] such that test set triplets have zero probability distribution under train set, but test set is composed of the same atoms (*e.g.* subjects) as in train set. They found that state-of-the-art models [88] cannot generalize compositionally. [75] studied compositional generalization from the point of image captioning, focusing on systematicity. They created a new split of COCO dataset composed of novel combinations of adjective-noun and noun-verb pairs [89] and investigated to what extent image captioning models can generalize systematically. The authors found that state-of-the-art image captioning models [88], [90] failed to generalize systematically, mainly due to language generation components based on textual distributions that cannot incorporate a compositional perspective. By creating a multi-task setup combining captioning and image-sentence ranking on top of the Bottom-Up and Top-down attention model from [90], they achieve better results.

Visual Question Answering. In the Visual Question Answering (VQA) task, the reasoning and understanding capabilities of the models are examined by the answers they provide to the questions about visual scenes. Work on compositionality in VQA can be divided into two: where the image distribution in the dataset is changed, or the question distribution in the dataset is changed. Independent of the division, compositional VQA tasks test spatial reasoning, quantification, and comparison in general. CLEVR-CoGenT [59] is a compositional VQA task that requires the listed skills above from a model. By using mutually exclusive colors for cubes and cylinders in the train set and swapping these colors in the test set, the dataset measures compositional skills. CLOSURE [91] is another VQA task derived from CLEVR. In CLOSURE, the change in image distribution is modified with a change in question distribution, which increases the complexity of the task by enabling multiple referring expressions. In VQA-CP, [92] find that VQA models heavily rely on priors in data while producing correct results on non-compositional datasets, which inflates

models compositional understanding of the real-world phenomenon.

Cloze. In a compositional generalization setup, [78] investigated the acquisition of words from visual scenes. They measured how effectively the representations generalize to new verb and noun combinations that aren't part of the training set. They employed the cloze task for model evaluation but require the model to predict both a verb and a noun rather than just one word. The findings were then divided based on whether or not the compositions were visible during training. According to the authors, there is a significant performance difference between compositions that have been seen before and those that haven't. However, the gap is considerably lower because their model is explicitly trained for generalization (nearly twice as small). Additionally, their method significantly outperforms baselines for both known and unique compositions. Furthermore, even though their model is trained on three orders of magnitude less training data than pre-trained BERT [93], their approach can outperform or match BERT's performance.

3.3. Datasets

In this section, we review the datasets and their characteristics, which enable research on the compositional generalization of neural models.

SCAN. [51] introduced SCAN, a synthetically generated dataset that consists of navigation commands associated with action sequences such that the command "*jump thrice*" is mapped to actions JUMP JUMP JUMP and "*turn around left*" is mapped to LTURN LTURN LTURN LTURN. In this paper, the goal of the SCAN tasks is to translate the given commands in simplified natural language into a sequence of actions, evaluating the generalization abilities of recurrent sequence-to-sequence models.

CTL. [82] introduced a setup to measure compositional generalization of neural networks by artificially generating compositional table lookup (CTL) using a simple binary representation *e.g.* if $g(000) = 010$, $g(010) = 001$, $c(001) = 000$ and $c(010) = 101$, function compositions are applied as $cg(000) = 101$, $gg(000) = 001$, $cgc(001) = 101$. In this study, lookup tasks were formulated as, given the atomic functions such that g and c , the objective is to predict

the output of their composition cg . Here, the output of the composition includes the output of the intermediate step in this case applying function g , then f to predict the final output cg .

PCFG-SET. [57] introduced a synthetically generated dataset in which sequences are generated by a probabilistic context-free grammar (PCFG) where input sequences should be translated into output sequences that represent their meanings. PCFG-SET contains three types of words: words for unary functions (*e.g.* copy, reverse, swap) and binary functions (*e.g.* append, prepend, remove_first), and elements these functions can be applied to (*e.g.* A, B, A1, B1), and a separator to separate the arguments of a binary function (.). Particularly, for the given input sequence “repeat A B C” the expected output is A B C A B C and for “append swap F G, repeat I J” the target is G F I J I J.

CFQ. [77] introduced *Compositional Freebase Questions* (CFQ), a procedurally generated dataset, where the objective is translating natural language questions into SPARQL queries against the Freebase. For example, the question “Did a male film director produce and edit M1?” corresponds to the query `SELECT count (*) WHERE { ?x0 ns:film.producer.film M1 . ?x0 ns:film.editor.film M1 . ?x0 ns:people.person.gender m_05zppz }`. The main idea behind the generation of CFQ is *distribution-based compositionality assessment* (DBCA), where the train and test split display similar atom (*e.g.* entities, question patterns) distributions while the compound distributions between train and test splits are maximized using *maximum compound divergence* (MCD) heuristic. It has been empirically shown that there is a significant negative correlation between the model accuracy and the compound divergence.

NACS. [79] introduced NACS by leveraging the SCAN dataset. In particular, they modified the SCAN dataset so that action sequences are translated into navigation commands rather than navigation commands being translated into action sequences, such as translating `LTURN` into “turn left”, instead of vice versa. This simple modification makes the SCAN dataset more difficult in terms of compositional generalization for seq2seq models.

Mathematics. [94] presented the Mathematics dataset, comprised of mathematics question-answer pairs up to university level in textual form (*e.g.* solve $3(x-5) = x+3$ for

\times , where the answer is 9). Compared to other datasets like SCAN, Mathematics focuses on mathematical reasoning rather than language transduction to test the compositional skills of a model. This large, automatically generated dataset is divided into 56 modules (*e.g.* linear equations in two variables, addition or summation, differentiation) and has 2 million training examples with 10K test examples for each module. Moreover, the dataset includes extrapolation modules to test compositional generalization.

gSCAN. [60] presented gSCAN, a dataset based on SCAN [51] for measuring the compositional skills of models in a grounded learning setting. In natural languages the interpretation of linguistic data is grounded in the real world, introducing the contextual sensitivity notion. However, SCAN imitates a navigation task in only in textual form, where the goal is to learn an interpretation function with only the linguistic structure. Hence, SCAN lacks this type of contextual sensitivity. gSCAN alleviates this shortcoming by grounding textual information with visual information in a grid world setup (see Figure 3.2), measuring both the compositionality and contextual sensitivity of models.

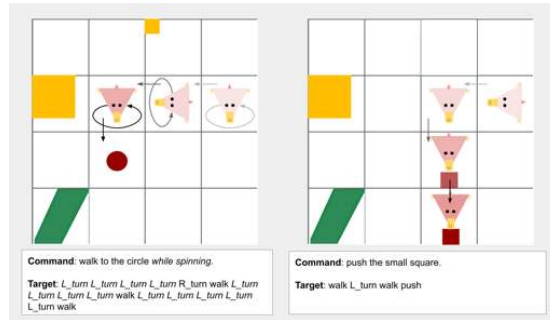


Figure 3.2 Illustration of a sample from gSCAN showing actions walking while spinning and pushing. Image taken from [60].

For example, *Walk to the small circle* command and an accompanying visual representation of a grid world with the target object, several distracting objects, and the agent altogether form a data sample. All objects can have different sizes, shapes, and colors, which implies that *small circle* in a visual representation is context-dependent. The introduction of context sensitivity enables testing broader types of compositional generalization.

ReaSCAN. [95] presented ReaSCAN, a dataset based on gSCAN trying to alleviate its limitations. The authors argue that there are four potential downsides of the gSCAN dataset,

namely irrelevance of word order, biased distractor object sampling, a small number of effective distractors, and its limits in testing linguistic compositionality. Irrelevance of word order makes bag-of-words models adequate for encoding gSCAN without any compositional understanding. In gSCAN, all compounds are not needed to generate the correct action sequence which raises the question of whether this can help models to attain linguistic compositional skills effectively. Hence, ReaSCAN has a more complex structure achieved by multiple relational clauses and that encourages models to display compositional skills (*e.g.* **gSCAN** \rightarrow push the big yellow circle cautiously, **ReaSCAN** \rightarrow push the big yellow circle that is in the same column as a blue square and in the same row as a red circle). Moreover, in ReaSCAN, commands become ambiguous when permuted, subsequently a random distractor sampling is used to make ReaSCAN more robust.

CLEVR-CoGenT. CLEVR [59] is a visual question-answering benchmark that measures a spectrum of visual reasoning skills. As CLEVR is artificially generated, it addresses the problem of inherent biases that can elicit in real-world data. The dataset consists of question-image-answer triplets, where each image is a scene from a 3D-rendered world with multiple objects. The questions require complex multi-step reasoning about the contents of rendered images. The objects are made of two different materials (matte “rubber” and shiny “metal”) and can have 8 colors (*e.g.* yellow, red), 2 sizes (small and big), 3 shapes (sphere, cube, and cylinder), and an absolute position in the 3D scene. An example question-image-answer triplet could then be an image of a 3D visual scene (see Figure 3.3), the question “*What color is the cylinder in front of the yellow cube?*” and, the answer “brown”.

CLEVR tests a range of reasoning abilities such as spatial reasoning, comparison, and quantification. In this dataset, authors proposed the CLEVR-CoGenT split, created to test the compositional generalization abilities of models. In the CLEVR-CoGenT train split, all cubes are gray, blue, brown, and yellow, whereas all cylinders are red, green, purple, or cyan and to measure compositional generalization, colors between cubes and cylinders are swapped in the test set, while spheres colors can be any color from the given color palette.



Figure 3.3 Illustration of a visual scene sample from CLEVR dataset. Image taken from [59].

CLOSURE. [91] constructed CLOSURE, a visual question-answering dataset based on CLEVR [59], which tests spatial, logical, and quantification skills of a model. While CLEVR-CoGenT tests models’ generalization skills under changing image distributions, on the contrary, CLOSURE tests models’ generalization skills under changing question distributions caused by the introduction of novel question compounds. The authors argue that the main complexity of CLEVR is how the present objects are referred to and to augment this complexity, they defined 7 different CLOSURE tests. In this scope, authors used three types of *referring expression* (RE) from CLEVR (*e.g.* simple, complex, and logical REs) where RE is a noun phrase referring to one or more objects. Then, by composing novel questions similar to CLEVR with matching object properties (*e.g.* small sphere that is the same color as the big cube) they constructed the CLOSURE dataset. The questions in this dataset follow the same structure as the CLEVR questions, however, the CLOSURE questions have zero probability distribution under CLEVR data.

SQOOP. SQOOP [72] is a synthetic visual question-answering dataset to test the compositional skills of models in terms of spatial reasoning. The dataset is composed of image-question-answer triplets, where each 64x64 image contains 5 distinct alphanumeric characters, and each question has a binary answer (yes/no). Each question has the following structure: $(\text{CHAR}_a \text{ SP_RELATION } \text{CHAR}_b)$ where CHARs are distinct alphanumeric characters, and SP_RELATION is an element from the set $\{\text{LEFT_OF}, \text{RIGHT_OF}, \text{ABOVE}, \text{BELOW}\}$. These properties make SQOOP simple and maybe not representative of

the real world, but they measure the spatial reasoning of models in an unbiased and isolated manner, which is beneficial.

COGS. [96] introduced a procedurally generated semantic parsing dataset (COGS) based on a subset of English. COGS examples consist of a natural language sentence, and a corresponding logical form (*e.g.* A frog hopped. $\rightarrow \text{frog}(\text{x1}) \text{ AND } \text{hop}.\text{agent}(\text{x2}, \text{x1})$). COGS embodies a large spectrum of syntactic structures and semantic representations (by using lambda calculus) that emerge in natural languages. Hence, the authors suggest that COGS can be a better benchmark for compositionality compared to SCAN, as it encompasses more systematic subcomponents that are present in natural languages. The evaluation part of COGS contains several systematic gaps that can only be addressed by compositional generalization such as new combinations of known syntactic structures or new combinations of known words. The dataset consists of three sets: a train set, an independent and identically distributed (IID) validation set, and an out-of-distribution (OOD) test set.

EPIC-Kitchens. [97] introduced an egocentric video benchmark dataset recorded by 32 participants in their native kitchen environments capturing non-scripted daily household activities such as cooking, cleaning, *etc.* This dataset is then extended by [98] to 100 hours of videos captured by 45 participants with denser annotations and fine-grained actions (see Fig. 3.4 for an overview of the dataset).



Figure 3.4 Illustration of EPIC-KITCHENS dataset. Image taken from [98].

Even though the EPIC-Kitchens dataset was introduced for mainly object detection, action recognition, and action anticipation tasks, it was later utilized for measuring the

compositional abilities of neural networks on a word acquisition task [78] leveraging its multimodal nature.

VQA-CP. [92] constructed a real-world visual question-answering dataset from the point of compositional generalization where the types of questions in the test split are of different distribution than the types of questions found in the train split. The authors discovered that VQA models exploit the biases in real-world data while achieving near-perfect results on non-compositional datasets and do not show any compositional understanding of underlying phenomena.

3.4. Methods

In this section, we review the popular models proposed for addressing compositional generalization and discuss learning strategies used in the existing work.

3.4.1. Models

Some researchers have tackled the compositional generalization problem with different neural architectures. In the following, we highlight the most prominent models investigated in the existing studies.

Neural Sequence Models. Recurrent neural networks have been widely used for scanning text and have been extensively used in the literature for solving various NLP tasks. [82] experimented with standard recurrent neural networks (RNN) [99] without any special architectural constraints. [51] described a simple sequence-to-sequence framework to tackle the compositional generalization problem on the SCAN dataset.

Convolutional Neural Models. Convolutional neural networks (CNN) initially proposed to process images, have been shown to work well in many vision problems [52]. After the early success of CNNs, convolutional architectures have been applied to solving various problems and adopted to process text input in NLP research [100]. [101] proposed a convolutional

model for sequence-to-sequence learning where the authors discuss advantages over a recurrent neural model for the seq2seq setting. Recently, [57] explored the compositional generalization abilities of the aforementioned convolutional neural model as a baseline on PCFG SET tasks.

Neural Module Networks. [58] approached compositionality from a structural point of view and proposed a dynamic neural module network for answering queries given only (question, world, answer) triples as training data, where the model learns to assemble neural networks on the fly from an inventory of neural models, and simultaneously learns weights for these modules so that they can be composed into novel structures. In a similar direction, [102] proposed an end-to-end module network in a visual question-answering set up to address the compositional aspect of the task.

Transformers. [103] introduced the transformer architecture, which then is widely adopted by the NLP community and has been shown to yield state-of-art results in many NLP tasks. The transformer model consists of a stack of encoder layers and a stack of decoder layers. Each stack of encoder and decoder layers has its corresponding embeddings. The model also utilizes a self-attention mechanism which works as a function to map a query and a set of key-value pairs to an output. Coupled with self-attention, positional encodings in the transformer model allow the model to process input in parallel, which gives the model an advantage over recurrent models where models need to process tokens in a sequential manner. Recent studies in compositional generalization literature explored different aspects of the Transformer model and its generalization capacity across different compositional tasks and datasets [104–107]

3.4.2. Learning Strategies

In the literature, there has been a growing body of work in which researchers investigated a different aspect of the compositional generalization problem. Below we review different learning strategies explored in the compositional generalization works.

Meta-learning. [74] investigated systematicity in a meta-learning setup. In this study, the author proposed a seq2seq architecture as a backbone coupled with a memory unit to solve SCAN tasks and demonstrates that memory-augmented neural networks can be trained to generalize compositionally through this setup. The meta-learning setup consists of meta-training and test phases. In the meta-training phase, models observe episodes of primitive instructions such as “jump”, “run”, “walk”, and “look” and learn the corresponding meanings for these primitives JUMP, RUN, WALK, LOOK. In test time, support items are loaded into the memory, and models are required to learn new meanings of primitives during test time to predict the target commands. Different from the regular seq2seq training setup, the models are provided support items in each training episode, which then are used to form a context to improve models’ compositional generalization abilities while also utilizing a self-attention mechanism with an external memory. Another work that leverages meta-learning is [78] on the EPIC-Kitchens [97] and Flickr30K [108] datasets in which the authors provide reference primitives to aid the models both during training time and test time. In this setup, the models are not only trained with training data but also reference instances are provided to support the models. The same procedure is used in the test phase, where the models are tasked with predicting the masked target word while also being provided a set of reference instances that contains the target word. While the objective is to generalize to novel compositions, this study diverges from the usual compositional generalization problem as the main goal is to acquire words from the visual scene and to generalize to unseen compositions of these primitives while being supported with reference instances that the models have access to during both the training as well as the test time.

Supervised Learning. Much recently, [109] formulated the compositional generalization problem as a classification problem and converted a natural language sequence-to-sequence dataset to a classification dataset that requires compositional generalization abilities. In this study, the authors converted the CFQ dataset into a binary classification dataset, where the input is a question and a SPARQL query, then the objective is to decide whether the given two inputs have the same meaning or not. Diverging from most of the sequence-to-sequence tasks in the literature, the authors describe a negative sampling strategy and construct negative

samples for every input to train and evaluate the models. Here, a random sampling strategy for constructing incorrect predictions was followed by another approach, where a pre-trained network is used to retrieve incorrect examples to choose hard examples aiming to make the task more challenging, hence requiring the models to leverage compositional generalization abilities to correctly predict the target.

Pre-training. Lately, Transformer-based architectures [103] have been shown to achieve state-of-the-art results in many NLP tasks. Thanks to their generalization capacity, transformer-based pre-trained models such as BERT and its variants are hugely adopted in the literature and applied to solving different tasks. These pre-trained models are often trained on large datasets and are widely used for solving different tasks, such as semantic labeling, sentiment classification, and language generation. Some researchers leveraged these models as pre-trained encoders and applied them to solving various tasks in NLP. One of the recent works in this direction is proposed by [104], in which the authors investigated the models’ generalization abilities in a compositional generalization setup. They explored how transformer models perform in compositional generalization tasks against different compositional generalization datasets and how the design decisions in transformers, such as position encoding, decoder type, and weight sharing, impact models’ compositional generalization abilities. The authors reported that they achieve state-of-the-art results in COGS and PCFG datasets.

Augmentation. Data augmentation is a strategy recently researchers explored from the point of improving the compositional generalization performance of models. A replacement-based strategy was employed by [110], where they explored a synthetic augmentation procedure to construct examples for downstream tasks such as semantic parsing using compositionality. In this approach, fragments of original data instances are replaced with fragments from other instances where samples are chosen with similar contexts. Another data augmentation strategy was introduced by [111]. In particular, they created new synthetic examples by randomly combining parts of two sentences from the training data to encourage models to rely on compositions of sentence segments to predict the output. [112] proposed a data augmentation procedure to improve compositional

generalization in neural sequence models by recombining fragments of training instances to reconstruct other instances and resample model outputs to pick high-quality synthetic samples. Recently, [105] studied one-shot primitive generalization introduced by the SCAN benchmark. They show that typical seq-to-seq models can obtain near-perfect generalization performance by changing the training distribution in simple and intuitive manners. Furthermore, they carried out thorough empirical analyses demonstrating that the traditional seq-to-seq models’ generalization capability is largely underestimated. A more general conclusion of their research is that, even though systematicity must be preserved when creating such benchmarks, it is crucial to carefully examine various setup parameters in order to make meaningful judgments about a model’s generalization capabilities. In this work, the authors argue that the proposed approach preserves the systematic differences between train and test sets while improving the performance of the seq2seq models, which some of the earlier work ignored while trying to improve compositional generalization performance but break the systematicity aspect of the setup.

Prompting. [113] proposed the prompting technique to improve the performance of language models for various tasks. In particular, the authors discuss that large language models can be used to solve complex mathematical reasoning tasks. In this work, the authors propose prompting with chain-of-thought, allowing models to use intermediate reasoning steps, hence demonstrating that prompting strategy can significantly improve the performance of language models on a different task, in addition to providing better compositional generalization.

3.5. Research Challenges

In this section, we discuss the open research challenges toward understanding compositional generalization. In particular, we review these challenges in the following orthogonal dimensions: *natural datasets*, *violation compositionality*, *a general evaluation benchmark*, *multimodality*, *multilinguality*, and *interpretability*.

Natural Datasets. Much of the previous work investigated compositional generalization problem using artificially generated datasets due to their ease of creation with the predetermined rules and constraints. These synthetic dataset construction strategies followed by many researchers enable them to build larger datasets that adhere to a set of rules and constraints (*e.g.* achieving different distribution for compositions across different splits *etc.*). On the other hand, it remains a challenge to achieve a certain distribution for natural datasets due to the lack of labels or even the existence of certain data. Even though there has been a considerable amount of effort in understanding and assessing compositional generalization problem with natural data, and an interest in this direction, we consider this to be an important open research problem to be addressed in future work.

Violating Compositionality Constraints. We should emphasize that diverging from the regular systematic generalization setup, where the models are expected to generalize to novel compositions and are evaluated against compositional generalization constraints, the pre-training in transformer model training naturally does not adhere to these constraints as the models during their pre-training phase could have been already exposed to the compositions which they are evaluated in test time. Nevertheless, it is still worth investigating the compositional generalization abilities of these models and exploring how pre-trained transformer models perform in various compositional generalization tasks. This could be an interesting direction for transformer-based compositional generalization research for future work. In another direction, researchers have been investigating data augmentation strategies to alleviate the out-of-distribution challenge we observe in the compositional splits. We should highlight that, little has been done in previous studies to enforce the compositional nature of datasets. This particular problem of data augmentation without breaking the compositional aspect of the problem in compositional splits remains an open challenge and is a future research direction.

A General Evaluation Benchmark. As compositional generalization research has been rekindled in the past decade, there is no general evaluation benchmark to test the compositional skills of a model. The lack of well thought out general benchmark results in a vicious loop, where a new benchmark is curated to remedy problems in an earlier benchmark,

a new symbolic/neural model performs great on this specific dataset while previous models cannot, then a new benchmark is created that the new model does not perform well. For natural language understanding (NLU) and language generation (LG) tasks, there exist de facto general evaluation benchmarks (*e.g.* GLUE for NLU and GEM for LG) [114], [115], [116]. By following a similar path, a systematic way to follow the progress in compositional models can be created. A general evaluation benchmark quite likely would accelerate the research and increase the quality and robustness of future studies.

Multimodality. The contribution of multimodality to the compositional generalization problem has become an interesting research problem researchers recently started to explore. However, due to limitations with natural datasets explored for compositional generalization, whether multimodality can help models systematically generalize and how multimodality can help solve compositional generalization, yet remains an open research problem. Nevertheless, we believe exploring multimodality, in a grounded language understanding and generation setup could be a fruitful research direction as we humans leverage multimodal cues in the real world to generalize compositionally. Toward this goal, we investigate the contribution of multimodality in compositional generalization for how-to instructions in Chapter 6..

Multilinguality. Multilinguality and its connections to compositional generalization is another research direction that remains an under-explored topic yet. There has been little done in the past to understand the impact of models’ compositional generalization abilities in multilingual settings, even though there has been a plethora of work related to multilinguality in the past years. From this point of view, this makes understanding models’ compositional generalization skills on multilinguality an exciting research direction for future work.

Interpretability. Neural networks have been largely criticized in the literature because of their black-box nature and their lack of interpretability. A body of work focuses on interpretability in machine learning models, but we consider this a significant research problem for compositional generalization since little has been done to understand how

models compositionally generalize and why. We consider this could be an interesting research direction for future work.

4. MULTIMODAL COMPREHENSION OF COOKING RECIPES

Understanding and reasoning about cooking recipes is a fruitful research direction toward enabling machines to interpret procedural text. In this chapter, we introduce RecipeQA, a dataset for multimodal comprehension of cooking recipes, based on our work,

- RecipeQA: A Dataset for Multimodal Comprehension of Cooking Recipes, S. Yagcioglu, A. Erdem, E. Erdem, N. Ikizler-Cinbis. EMNLP 2018, Brussels, Belgium, Oct. 2018.

We only made minor changes to fit the text in the narrative of the thesis and made small corrections and changes in the text and figures.

RecipeQA comprises approximately 20K instructional recipes with multiple modalities such as titles, descriptions, and aligned sets of images. With over 36K automatically generated question-answer pairs, we design a set of comprehension and reasoning tasks that require a joint understanding of images and text, capturing the temporal flow of events, and making sense of procedural knowledge. Our preliminary results indicate that RecipeQA will serve as a challenging test bed and an ideal benchmark for evaluating machine comprehension systems. The data and leaderboard are available at <http://hucvl.github.io/recipeqa>.

There is a rich literature in natural language processing (NLP) and information retrieval on question answering (QA) [117], but recently deep learning has sparked interest in a special kind of QA, commonly referred to as reading comprehension (RC) [18]. The aim in RC research is to build intelligent systems with the abilities to read and understand natural language text and answer questions related to it [118]. Such tests are appealing as they require a joint understanding of the question and the related passage (*i.e.* context), and moreover, they can analyze many different types of skills in a rather objective way [119].

Despite the progress made in recent years, there is still a significant performance gap between humans and deep neural models in RC, and researchers are pushing forward our understanding of the limitations and capabilities of these approaches by introducing new datasets. Existing tasks for RC mainly differ in two major respects: the question-answer formats, *e.g.* cloze (fill-in-the-blank), span selection or multiple choice, and the text sources they use, such as news articles [27, 28], fictional stories [43], Wikipedia articles [44, 120, 121] or other web sources [122]. A popular topic in computer vision closely related to RC is Visual Question Answering (VQA) in which context takes the form of an image in the comprehension task, where recent datasets have also been compiled, such as [10, 16, 49, 123], to name a few.

More recently, research in QA has been extended to focus on the multimodal aspects of the problem where different modalities are being explored. Tapaswi *et al.* [48] introduced MovieQA where they concentrate on evaluating automatic story comprehension from both video and text. In COMICS, Iyyer *et al.* [36] turned to comic books to test understanding of closure, and transitions in the narrative from one panel to the next. In AI2D [124] and FigureQA [125], the authors addressed comprehension of scientific diagrams and graphical plots. Last but not least, Kembhavi *et al.* [50] has proposed another comprehensive and challenging dataset named TQA, which is comprised of middle school science lessons of diagrams and texts.

In this study, we focus on *multimodal machine comprehension of cooking recipes* with images and text. To this end, we introduce a new QA dataset called *RecipeQA* that consists of recipe instructions and related questions (see Fig. 4.1 for an example text cloze style question).

There are a handful of reasons why understanding and reasoning about recipes is interesting. Recipes are written with a specific goal in mind, which is to teach others how to prepare a particular food. Hence, they contain immensely rich information about the real world. Recipes consist of instructions, wherein one needs to follow each instruction to successfully complete the recipe. As a classical example in introductory programming classes, each recipe

Text Cloze Style Question	Context Modalities: Images and Descriptions of Steps
<p>Recipe: Last-Minute Lasagna</p> <ol style="list-style-type: none"> 1. Heat oven to 375 degrees F. Spoon a thin layer of sauce over the bottom of a 9-by-13-inch baking dish. 2. Cover with a single layer of ravioli. 3. Top with half the spinach half the mozzarella and a third of the remaining sauce. 4. Repeat with another layer of ravioli and the remaining spinach mozzarella and half the remaining sauce. 5. Top with another layer of ravioli and the remaining sauce not all the ravioli may be needed. Sprinkle with the Parmesan. 6. Cover with foil and bake for 30 minutes. Uncover and bake until bubbly, 5 to 10 minutes. 7. Let cool 5 minutes before spooning onto individual plates. 	
Question	Choose the best text for the missing blank to correctly complete the recipe Cover. _____. Bake. Cool, serve.
Answer	A. Top, sprinkle B. Finishing touches C. Layer it up D. Ravioli bonus round

Figure 4.1 An illustrative text cloze style question (context, question and answer triplet). The context is comprised of recipe descriptions and images where the question is generated using the question titles. Each paragraph in the context is taken from another step, as also true for the images. The bold answer is the correct one.

might be seen as a particular way of solving a task and in that regard can also be considered an algorithm. We believe that recipe comprehension is an elusive challenge and might be seen as an important milestone in the long-standing goal of artificial intelligence and machine reasoning [126, 127].

Among previous efforts towards multimodal machine comprehension [36, 48, 50, 124, 125], our study is closer to what Kembhavi *et al.* [50] envisioned in TQA. Our task primarily differs in utilizing a substantially larger number of images – the average number of images per recipe in RecipeQA is 12 whereas TQA has only 3 images per question on average. Moreover, in our case, each image is aligned with the text of a particular step in the corresponding recipe. Another important difference is that TQA contains mostly diagrams or textbook images whereas RecipeQA consists of natural images taken by users in unconstrained environments.

Some of the important characteristics of RecipeQA are as follows:

- There are arbitrary numbers of steps in recipes and images in steps, respectively.
- There are different question styles, each requiring a specific comprehension skill.
- There exists high lexical and syntactic divergence between contexts, questions and answers.
- Answers require understanding procedural language, in particular keeping track of entities and/or actions and their state changes.
- Answers may need information coming from multiple steps (*i.e.* multiple images and multiple paragraphs).
- Answers inherently involve multimodal understanding of image(s) and text.

To sum up, we believe RecipeQA is a challenging benchmark dataset that will serve as a test bed for evaluating multimodal comprehension systems. In this section, we present several statistical analyses on RecipeQA and also obtain baseline performances for a number of multimodal comprehension tasks that we introduce for cooking recipes.

4.1. RecipeQA Dataset

The Recipe Question Answering (RecipeQA) dataset [128] is a challenging multimodal dataset that evaluates reasoning over real-life cooking recipes. It consists of approximately 20K recipes from 22 food categories and over 36K questions.

Fig. 4.2 shows an illustrative cooking recipe from our dataset. Each recipe includes an arbitrary number of steps containing both textual and visual elements. In particular, each step of a recipe is accompanied by a ‘title’, a ‘description’, and a set of illustrative ‘images’ that are aligned with the title and the description. Each of these elements can be considered as a different modality of the data. The questions in RecipeQA explore the multimodal

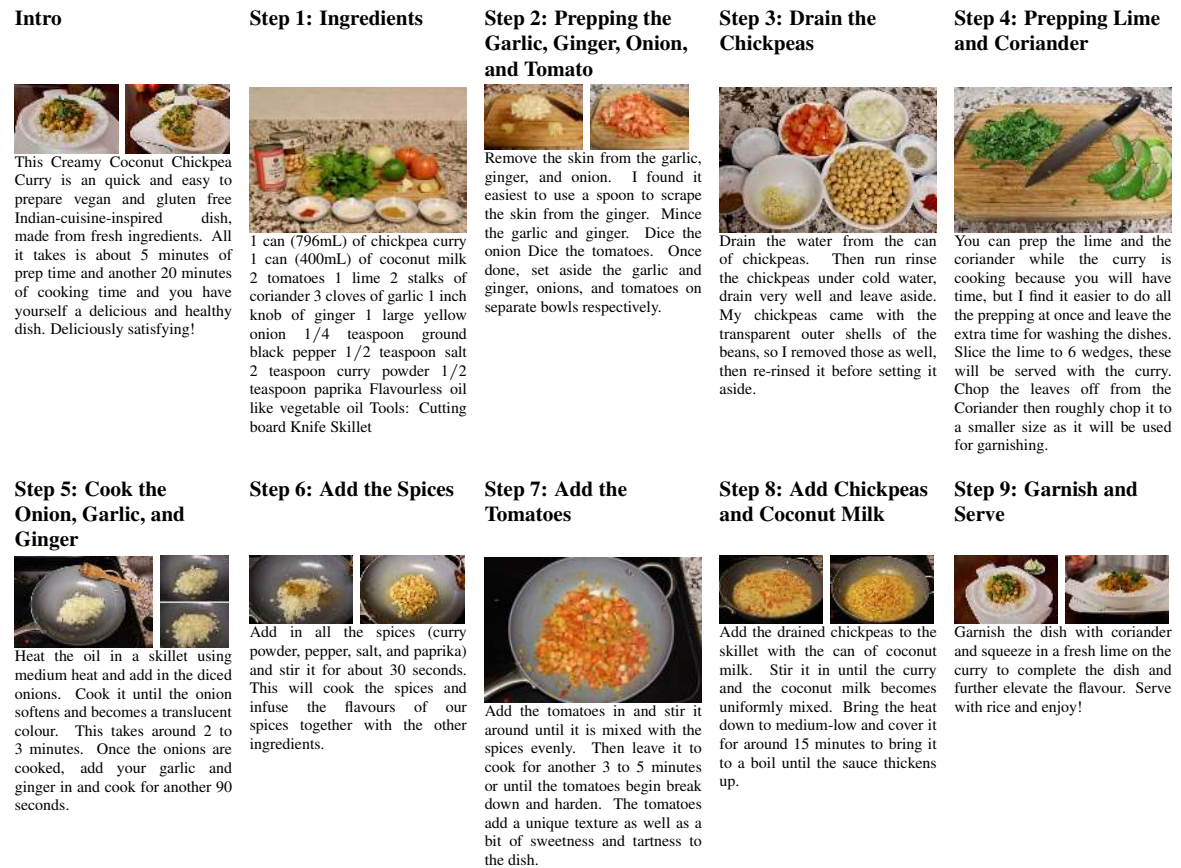


Figure 4.2 A recipe of ‘Creamy Coconut Chickpea Curry’ with 9 steps, taken from Instructables.

aspects of the step-by-step instructions available in the recipes through a number of specific tasks that are described in Sec. 4.2., namely *textual cloze*, *visual cloze*, *visual coherency* and *visual ordering*.

4.1.1. Data Collection

We consider cooking recipes as the main data source for our dataset. These recipes were collected from Instructables ⁵, which is a how-to website where users share all kinds of instructions including but not limited to recipes.

We employed a set of heuristics that helped us collect high-quality data in an automatic manner. For instance, while collecting the recipes, we downloaded only the most popular

⁵All materials from <http://instructables.com> were downloaded in April 2018.

recipes by considering their popularity as an objective measure for assessing the quality of a recipe. Our assumption is that the most viewed recipes contain less noise and include easy-to-understand instructions with high-quality illustrative images.

In total, we collected about 20K unique recipes from the food category of Instructables. We filtered out non-English recipes using a language identification [129] and automatically removed the ones with unreadable contents such as the ones that only contain recipe videos. Finally, as a post-processing step, we normalized the description text by removing non-ASCII characters from the text.

4.1.2. Questions and Answers

For machine comprehension and reasoning, forming the questions and the answers is crucial for evaluating the ability of a model in understanding the content. Prior studies employed natural language questions either collected via crowdsourcing platforms such as SQuAD [44] or generated synthetically as in CNN/Daily Mail [28]. Using natural language questions is a good approach in terms of capturing human understanding, but crowdsourcing is often too costly and does not scale well as the size of the dataset grows. Synthetic question generation is a low-cost solution, but the quality of the generated questions is subject to question.

RecipeQA includes structured data about the cooking recipes that consist of step-by-step instructions, which helps us generate questions in a fully automatic manner without compromising the quality. Our questions test the semantics of the instructions of the recipes from different aspects through the tasks described in Sec. 4.2.. In particular, we generate a set of multiple-choice questions (the number of choices is fixed as four) by following a simple procedure that applies to all of our tasks with slight modifications.

In order to generate question-answer-context triplets, we first filtered out recipes that contain less than 3 steps or more than 25 steps. We also ignored the initial step of the recipes as our preliminary analysis showed that the first step of the recipes almost always is used by the authors to provide a narrative, *e.g.* why they love making that particular food, or how it makes

sense to prepare food for some occasion and often is not relevant to the recipe instructions. In addition, we automatically removed some indicators such as step numbers that explicitly emphasize temporal order from the step titles while generating questions.

Given a task, we first randomly select a set of steps from each recipe and construct our questions and answers from these steps according to the task at hand. In particular, we employ the modality that the comprehension task is built upon to generate the candidate answers and use the remaining content as the necessary context for our questions. For instance, if the step titles are used within the candidate answers, the context becomes the descriptions and the images of the steps. As the average number of steps per recipe is larger than four, using this strategy, we can generate multiple context-question-answer triplets from a single recipe.

Candidate answers can be generated by selecting the distractors at random from the steps of other recipes. To make our dataset more challenging, we employ a different strategy and select the distractors from the relevant modalities (titles, descriptions or images), which are not too far or too close from the correct answer. Specifically, we employ the following simple heuristic. We first find k nearest neighbors ($k = 100$) from other recipes. We then define an adaptive neighborhood by finding the closest distance to the query and removing the candidates that are too close. The remaining candidates are similar enough to be adversarial but not too similar to semantically substitute for the groundtruth. Finally, we randomly sample distractors from that pool. Details of the question generation procedure for each of the tasks are given in Sec. 4.2..

4.1.3. Dataset Statistics

RecipeQA dataset contains approximately 20K cooking recipes and over 36K question-answer pairs divided into four major question types reflecting each of the tasks at hand. The data is split into non-overlapping training, validation and test sets so that one set does not include a recipe and/or questions about that recipe which are available in other sets. There are 22 different food categories across our dataset whose distribution is shown in

Fig. 4.3. While splitting the recipes into sets, we take into account these categories so that all the sets have a similar distribution of recipes across all the categories.

Table 4.1 RecipeQA dataset statistics.

	train	valid	test
# of recipes	15847	1963	1969
... avg. # of steps	5.99	6.01	6.00
... avg. # of tokens (titles)	17.79	17.40	17.67
... avg. # of tokens (descr.)	443.01	440.51	435.33
... avg. # of images	12.67	12.74	12.65
# of question-answers	29657	3562	3567
... textual cloze	7837	961	963
... visual cloze	7144	842	848
... visual coherence	7118	830	851
... visual ordering	7558	929	905

In Table 4.1, we show detailed statistics about our RecipeQA dataset. Moreover, to visualize the token frequencies, we also provide the word clouds of the titles and the descriptions from the recipes in Fig. 4.4.



Figure 4.3 Distribution of food categories across RecipeQA.

4.2. Tasks

RecipeQA includes four different types of tasks: (1) Textual cloze, (2) Visual cloze, (3) Visual coherence, and (4) Visual ordering. Each of these tasks requires different reasoning

that depends on the WMD [130] distance measure. In Fig. 4.1, we provide a sample text cloze question from RecipeQA generated automatically in this way.

4.2.2. Visual Cloze

Visual cloze-style questions test a skill similar to that of a textual cloze task with the difference that the missing information in this task resides in the visual domain. Here, just like the textual cloze task, for a recipe, we randomly select a step, hide its representative image, and ask to infer this image amongst the multiple choices. The context for this task is all textual and is in the form of a sequence of titles and descriptions. To construct the distractor images, we use Euclidean distances of 2048-d *pool5* features extracted from a ResNet-50 [131] pre-trained on ImageNet classification task. We show a sample visual cloze style question in Fig. 4.5 (second row).

4.2.3. Visual Coherence

Visual coherence style questions test the capability to identify an incoherent image in an ordered set of images given the titles and descriptions of the corresponding recipe as the context. Hence, to be successful at this task, a system needs to not only understand the relations between candidate steps but also align and relate different modalities existing in the context and the answers. While generating the answer candidates for this task, we randomly select a single representative image from a single step and replace this image with a distractor image by employing the distractor selection strategy used for the visual cloze task. In Fig. 4.5 (third row), we provide a sample visual coherence style question from RecipeQA.

4.2.4. Visual Ordering

Visual ordering questions test the ability of a system in finding a correctly ordered sequence given a jumbled set of representative images of a recipe. As in the previous visual tasks, the context of this task consists of the titles and descriptions of a recipe. To successfully complete

















Context Modalities: Titles and Descriptions of Steps for Recipe: Bacon Sushi	
	<p>Step 1: What You'll Need This recipe makes enough bacon sushi to feed 2 - 4 people. 2 x 500g(1 lb.) packages of bacon (I chose an applewood smoked bacon, but any type would work)...</p> <p>Step 2: Cooking the Bacon The bacon "nori" will have to be partially cooked before it can be rolled with the risotto filling. Preheat the oven to 350 degrees F. Lay half a package of bacon on...</p> <p>Step 3: Making the Risotto Filling I once made risotto with sushi rice, since I had no Arborio rice on hand, and I decided that the starchiness was similar in the two...</p> <p>Step 4: Jazzing Up the Risotto Risotto is a wonderfully customizable dish, and a quick search on the internet will result in a multitude of variations. Here are two of my favorites: Asian mushroom risotto. 1 tbsp. oil. 1 package...</p> <p>Step 5: Rolling the Sushi Cover the sushi rolling mat with a large piece of aluminum foil as protection from the risotto and bacon grease. (You don't want your next sushi dinner tasting like bacon. Or maybe you do...) Lay the stri...</p> <p>Step 6: Baking and Slicing Preheat the oven to 350 degrees F. Place the aluminum foil-covered sushi rolls in the oven and bake for 20 minutes. This will warm all the ingredients and crisp the bacon a little more. It will also melt a...</p> <p>Step 7: And You're Done! Serve the sushi with a light crispy vegetable side dish, such as refreshing cucumber sticks, or a green salad. White wine makes an excellent compliment to the meal, especially if it is the same wine used in ...</p>
	<p>Question Choose the best image for the missing blank to correctly complete the recipe</p> <div>     </div> <div>     </div> <p>Answer A. B. C. D.</p>
	<p>Question Select the incoherent image in the following sequence of images</p> <div>     </div> <p>Answer A. B. C. D.</p>
	<p>Question Choose the correct order of the images to make a complete recipe</p> <div>     </div> <p>(i) (ii) (iii) (iv)</p> <p>Answer A. (iv)-(iii)-(ii)-(i) B. (iv)-(iii)-(i)-(ii) C. (i)-(ii)-(iii)-(iv) D. (ii)-(iv)-(i)-(iii)</p>

Figure 4.5 Sample question types (context, question and answer triplet) taken from the RecipeQA training set. The correct answers are shown with green frames or in bold.

this task, the system needs to understand the temporal occurrence of a sequence of recipe steps and infer temporal relations between candidates, *i.e.* boiling the water first, putting the spaghetti next so that the ordered sequence of images aligns with the given recipe. To generate answer choices, we simply use random permutations of the illustrative images in the recipe steps. In Fig. 4.5 (last row), we illustrate this visual ordering task through an example question. Here, we should note that a similar task has been previously investigated by [132] for visual stories where the task is to order a jumbled set of aligned image-description pairs.

4.3. Experiments

4.3.1. Data Preparation

Ingredient Detection. We employed the method proposed in [133] to detect recipe ingredients. To learn more effective word embeddings, we transformed the ingredients with compound words such as *olive oil* into single-word ingredients with a proper hyphenation as *olive_oil*.

Textual Embeddings. We trained a distributed memory model, namely Doc2Vec [134], and used it to learn word-level and document-level embeddings while encoding the semantic similarity by taking into account the word order within the provided context. In this way, we can represent each word, sentence, or paragraph by a fixed-sized vector. In our experiments, we employed 100-d vectors to represent all of the textual modalities (titles and descriptions). We made sure that the embeddings encode semantically useful information by exploring nearest neighbors (see Fig. 4.6 for some examples.)

Query	Nearest Neighbor
Then add the green onion and garlic.	Then add the white onion, red pepper and garlic.
It will thicken some while it cools	Some cornflour to thicken.
Slowly whisk in the milk, scraping the bottom and sides with a heatproof spatula to make sure all the dry ingredients are mixed in.	Stir the dry ingredients in, incrementally, mixing on low speed and scraping with a spatula after each addition.

Figure 4.6 Sample nearest neighbors from the embeddings by the trained Doc2Vec model.

Visual Features. We used the final activation of the ResNet-50 [131] model trained on the ImageNet dataset [135] to extract 2048-d dense visual representations. Then, we further utilized an autoencoder to decrease the dimension of the visual features to 100-d so that they become compatible in size with the text embeddings.

4.3.2. Baseline Models

Neural Baselines. For our neural baselines, we adapted the Impatient Reader model in [28], which was originally developed only for the cloze style text comprehension questions in the CNN/Daily Mail dataset. In our implementation, we used a uni-directional stacked LSTM architecture with 3 layers, in which we feed the context of the question to the network in a sequential manner. Particularly, we preserve the temporal order of the steps of the recipe while feeding it to the neural model, by mimicking the most common reading strategy – reading from top to bottom. For the multimodal setting, since images are represented with vectors that are of the same size as the text embeddings, we also feed the images to the network in the same order they are presented in the recipe.

In order to account for different question types, we employ a modular architecture, which requires small adjustments to be made for each task. For instance, we place the candidate answers into the query for the cloze style questions or remove the candidate answer from the query for the visual coherence type questions. In training our Impatient Reader baseline model, we use a cosine similarity function and employed the *hinge ranking loss* [136] as follows:

$$L = \max\{0, M - \cos(q, a_+) + \cos(q, a_-)\} \quad (1)$$

where M is a scalar denoting the margin, a_+ represents the ground truth answer, and a_- corresponds to an incorrect answer which is sampled randomly from the whole answer space. For all of our experiments, we select M as 1.5 and employ a simple heuristic to prevent overfitting by following an early stopping scheme with patience set to 10 against the validation set accuracy after the initial epoch. For the optimizer, we use ADAM and set the

learning rate to $1e - 3$. The training took around 18 to 24 hours on GTX 1080Ti on a single GPU. We did not perform any hyperparameter tuning.

Simple Baselines. We adapt the Hasty Student model described in [48], which does not consider the provided context and simply answers questions by only looking at the similarities or the dissimilarities between the elements in questions and the candidate answers. For the textual cloze task, each candidate answer is compared against the titles or descriptions of the steps by using WMD [130] distance, where such distances are averaged. Then, the choice closest to all of the question steps is selected as the final answer. For the visual cloze task, a similar approach is carried out by considering images instead of text using deep visual features. For the visual coherence task, since the aim is to find the incoherent image among other images, the final answer is chosen as the most dissimilar one to the remaining images on average. Lastly, for the visual ordering task, first, the distances between each consecutive image pair in a candidate ordering of the jumbled image set is estimated. Then, each candidate ordering is scored based on the average of these pairwise distances, and the choice with the minimum average distance is set as the final answer. In all these simple baseline models, we use the cosine distance to rank the candidates.

4.3.3. Baseline Results

We report the performance of the baseline models in Table 4.2 which indicates the ratio of correct answers against the total questions in the test.

Table 4.2 Results for simple and neural models on the test set of RecipeQA dataset.

	Visual Cloze	Textual Cloze	Visual Coherence	Visual Ordering
Hasty Student	27.35	26.89	65.80	40.88
Impatient Reader (Text only)	–	28.03	–	–
Impatient Reader (Multimodal)	27.36	29.07	28.08	26.74

For the textual cloze, the comparison between text-only and multimodal Impatient Reader models shows that the additional visual modality helps the model to understand the question

better and to provide more accurate answers. While for the cloze style questions, the Impatient Reader outperforms the Hasty student, for the visual coherence and visual ordering style questions Hasty student gives way better results. This demonstrates that better neural models are needed to be able to effectively deal with these kinds of questions. Some qualitative examples are provided in Fig. 4.7 and in Fig. 4.8.

4.4. Related Work

Question Answering has been studied extensively in the literature. With the success of deep learning approaches in question answering, comprehension, and reasoning aspects of the task has attracted researchers to investigate QA as a medium to measure intelligence. Various datasets and methods have been proposed for measuring different aspects of the comprehension and reasoning problem. Each dataset has its own merits as well as weaknesses. Recently, a thorough analysis by [29] revealed that the required reasoning and inference level was quite simple for CNN/Daily Mail dataset [28]. To make reasoning task more realistic, new datasets such as SQuAD [44], NewsQA [27], MSMARCO [35], CLEVR [49], COMICS [36] and FigureQA [125] have been proposed.

In the following, we briefly discuss the publicly available datasets that are closely related to our problem and provide an overview in Table 4.3.

Table 4.3 Comparison of the RecipeQA dataset to other multimodal machine comprehension datasets.

Dataset	#Images	#Questions	Modality
COMICS	1.2M	750K	Image/Text
MovieQA	408	14,944	Image/Video/Text
TQA	3,455	26,260	Image/Text
RecipeQA	250,730	36,786	Image/Text

The closest works to ours are [36], [48] and [50] where data multi-modality is the key aspect. COMICS dataset [36] focuses on comic book narratives and explores visual cloze style questions, introducing a dataset consisting of drawings from comic books. The dataset is constructed from 4K Golden Age (1938-1954) comic books from the Digital Comics





Context Modalities: Images and Descriptions of Steps	
Recipe: Grans-Green-Tomato-Chutney	
<p>Step 1: Ingredients: 2.5kg green tomatoes, roughly chopped 0.5kg onions, finely sliced 4 tsp / 30g salt 1L malt vinegar 0.5kg soft light brown sugar 250g sultanas, 1 1...</p> 	
<p>Step 2: Finely slice your onions and washed green tomatoes, cutting out any bad bits. Add to a large bowl and stir. Add the 4 teaspoons of salt, stir again and then cover with food wrap or a large plate and leave overnight. This will draw out lots of the tomato juices and...</p> 	
<p>Step 3: The next day...Place the litre of vinegar into a large pan. Add the 500g of light brown soft sugar and stir over a medium heat until all the sugar has dissolved. Bring to the boil...</p> 	
<p>Step 4:</p>	
<p>Step 9: While the jars cool, write some labels showing the date, content and maker. Once cool, add the lids and stick on the labels. ...</p> 	
Textual Cloze Style Question	<p>Question Choose the best text for the missing blank to correctly complete the recipe Ingredients. _____. Drain and Add the Tomatoes and Onions. Preparing Your Jars.</p>
	<p>Answer</p>
	<p>A. Sultanas B. Spicy Tomato Chutney. C. Cover and Slice. D. Enjoy.</p>
	<p>Hasty Student: Cover and Slice</p>
	<p>Neural Baseline (Text only): Sultanas</p>
	<p>Neural Baseline (Multimodal): Sultanas</p>

Figure 4.7 Sample groundtruth and model prediction results for a textual cloze style question (context, question and answer triplet) taken from the RecipeQA test set (Question Id: 1000-12665-0-3-4-6). Here, the context is comprised of step descriptions and images where the questions are generated using the step titles in the recipe. The correct answer is in green. The answers selected by the neural models are correct, marked as green whereas Hasty Student's prediction is wrong and marked as red.

















Context Modalities: Titles and Descriptions of Steps	
Recipe: Peppermint-Patty-Pudding-Shot	
<p>Step 1: Gather Ingredients To make peppermint patty pudding shots you will need: 1 small box of chocolate pudding 3/4 cup of milk 3/4 cup of peppermint schnapps 1 tub of cool whip...</p> <p>Step 2: Mixing of Ingredients First whisk together milk and pudding. Once that is combined add in the peppermint schnapps. Then fold in the cool whip...</p> <p>Step 3: Prep for Serving I then scoop the pudding into small plastic cups with lids. I buy them from a local Chinese restaurant, they are the perfect size. Throw these in the freezer until ...</p> <p>Step 4: Serve Pull them out of the freezer and sprinkle with the crushed peppermint. You can either lick them out of the cup or eat with a spoon :) I hope you enjoy them as much as we did ...</p>	
Visual Cloze Style Question	<p>Question Choose the best image for the missing blank to correctly complete the recipe</p> <div>     </div> <p>Answer</p> <div>     </div> <p>A. B. C. D.</p> <p>Hasty Student: C</p> <p>Neural Baseline (Multimodal): C</p>
Visual Coherence Style Question	<p>Question Select the incoherent image in the following sequence of images</p> <div>     </div> <p>Answer</p> <p>A. B. C. D.</p> <p>Hasty Student: C</p> <p>Neural Baseline (Multimodal): B</p>
Visual Ordering Style Question	<p>Question Choose the correct order of the images to make a complete recipe</p> <div>     </div> <p>(i) (ii) (iii) (iv)</p> <p>Answer A. (i)-(ii)-(iii)-(iv) B. (iii)-(i)-(iv)-(ii) C. (ii)-(iv)-(iii)-(i) D. (i)-(iii)-(iv)-(ii)</p> <p>Hasty Student: B</p> <p>Neural Baseline (Multimodal): A</p>

Figure 4.8 Sample question types taken from the RecipeQA test set. The correct answers are shown with green frames or in green. Wrong answers are marked as red.

Museum and contains 1.2M panels with 2.5M textboxes. Three tasks are evaluated in this context, namely text cloze, visual cloze, and character coherence. MovieQA dataset [48], comprises 15K crowdsourced questions about 408 movies. It consists of movie clips, subtitles, and snapshots and is about comprehending stories about movies. TQA dataset [50], has 26K questions about 1K middle school science lessons with 3.5K images, mostly of diagrams and aims at addressing middle school knowledge acquisition using both images and text. Since the audience is middle school children, it requires limited reasoning.

RecipeQA substantially differentiates from the previous work in the following way. Our dataset consists of natural images that are taken by anonymous users in unconstrained environments, which is a major diversion from COMICS and TQA datasets.

It should also be noted that there has been a long history of research involving cooking recipes. Recent examples include parsing of recipes [137, 138], aligning instructional text to videos [139, 140], recipe text generation [141], learning cross-modal embeddings [133], tracking entities and action transformations in recipes [142].

Finally, to our best knowledge, there is no dataset focusing on “how-to” instructions or recipes; hence, this work will be the first to serve multimodal comprehension of recipes having an arbitrary number of steps aligned with multiple images and multiple sentences.

4.5. Discussions

In this chapter, we presented RecipeQA, a dataset for multimodal comprehension of cooking recipes, which consists of roughly 20K cooking recipes with over 36K context-question-answer triplets. To our knowledge, RecipeQA is the first machine comprehension dataset that deals with understanding procedural knowledge in a multimodal setting. Each one of the four question styles in our dataset is specifically tailored to evaluate a particular skill and requires connecting the dots between different modalities. Results of our baseline models demonstrate that RecipeQA is a challenging dataset and hopefully will be useful for other researchers to promote the development of new methods for multimodal

machine comprehension. We also hope that RecipeQA will serve other purposes for related research problems on cooking recipes as well.

5. PROCEDURAL UNDERSTANDING OF COOKING RECIPES

In this chapter, we explore the problem of comprehending procedural commonsense knowledge and describe a novel architecture. In particular, we propose Procedural Reasoning Networks (PRN) for understanding procedural commonsense knowledge, based on our work,

- Mustafa Sercan Amac, Semih Yagcioglu, Aykut Erdem, and Erkut Erdem. Procedural reasoning networks for understanding multimodal procedures. In Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), pages 441–451. 2019.

We only made minor changes to fit the text in the narrative of the thesis and made small corrections and changes in the text and figures and added further discussion regarding related work that report results with the RecipeQA dataset.

Procedural understanding of cooking recipes is a challenging task as it requires identifying key entities, keeping track of their state changes, and understanding temporal and causal relations. Contrary to most of the previous work, in this study, we do not rely on strong inductive bias and explore the question of how multimodality can be exploited to provide a complementary semantic signal. Towards this end, we introduce a new entity-aware neural comprehension model augmented with external relational memory units. Our model learns to dynamically update entity states in relation to each other while reading the text instructions. Our experimental analysis on the visual reasoning tasks in the recently proposed RecipeQA dataset [128] reveals that our approach improves the accuracy of the previously reported models by a large margin. Moreover, we find that our model learns effective dynamic representations of entities even though we do not use any supervision at the level of entity states.

A great deal of commonsense knowledge about the world we live in is procedural in nature and involves steps that show ways to achieve specific goals. Understanding and reasoning

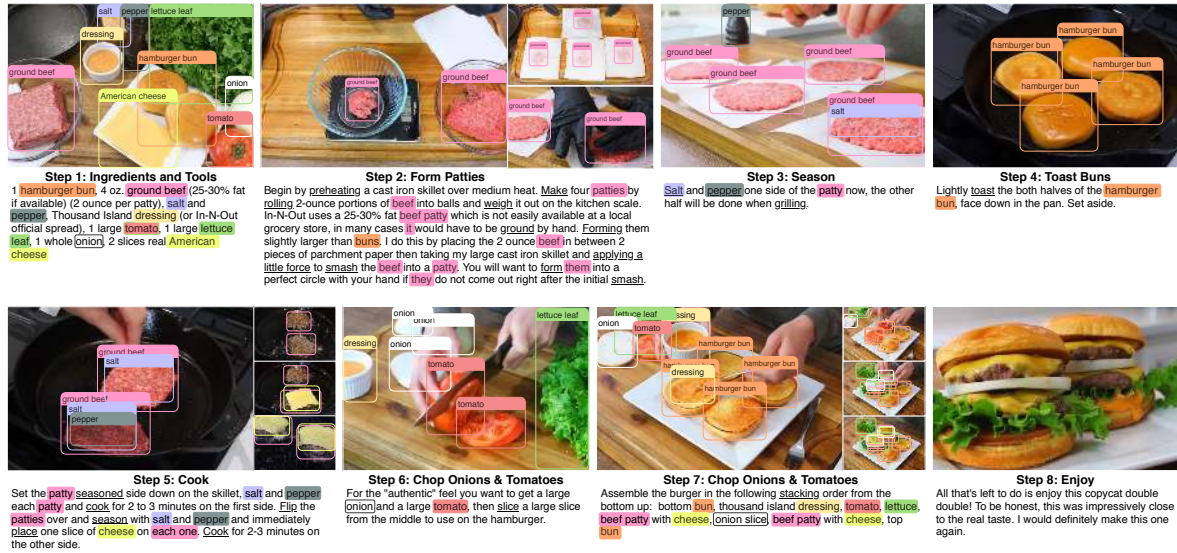


Figure 5.1 A recipe for preparing a cheeseburger (adapted from the cooking instructions available at <https://instructables.com/id/In-N-Out-Double-Double-Cheeseburger-Copycat>). Each basic ingredient (entity) is highlighted by a different color in the text and with bounding boxes on the accompanying images. Over the course of the recipe instructions, ingredients interact with each other, and change their states with each cooking action (underlined in the text), which in turn alter the visual and physical properties of entities. For instance, the tomato changes its form by being sliced up and then stacked on a hamburger bun.

about procedural texts (e.g. cooking recipes, how-to guides, scientific processes) are very hard for machines as it demands modeling the intrinsic dynamics of the procedures [128, 143, 144]. That is, one must be aware of the entities present in the text, infer relations among them and even anticipate changes in the states of the entities after each action. For example, consider the cheeseburger recipe presented in Fig. 5.1. The instruction “salt and pepper each patty and cook for 2 to 3 minutes on the first side” in Step 5 entails mixing three basic ingredients, the ground beef, salt and pepper, together and then applying heat to the mix, which in turn causes chemical changes that alter both the appearance and the taste. From a natural language understanding perspective, the main difficulty arises when a model sees the word *patty* again at a later stage of the recipe. It still corresponds to the same entity, but its form is totally different.

Over the past few years, many new datasets and approaches have been proposed that address this inherently hard problem [143–146]. To mitigate the aforementioned challenges, the

existing works rely mostly on heavy supervision and focus on predicting the individual state changes of entities at each step. Although these models can accurately learn to make local predictions, they may lack global consistency [145, 146], not to mention that building such annotated corpora is very labor-intensive. In this work, we take a different direction and explore the problem from a multimodal standpoint. Our basic motivation, as illustrated in Fig. 5.1, is that accompanying images provide complementary cues about causal effects and state changes. For instance, it is quite easy to distinguish raw meat from cooked one in the visual domain.

In recent years, tracking entities and their state changes have been explored in the literature from a variety of perspectives. In an early work, [147] proposed a dynamic memory-based network that updates entity states using a gating mechanism while reading the text. [148] presented a more structured memory-augmented model which employs memory slots for representing both entities and their relations. Pavez *et al.* [149] suggested a conceptually similar model in which the pairwise relations between attended memories are utilized to encode the world state. The main difference between our approach and these works is that by utilizing relational memory core units we also allow memories to interact with each other during each update.

Perez *et al.* [150] showed that similar ideas can be used to compile supporting memories in tracking dialogue state. Wang *et al.* [151] has shown the importance of coreference signals for reading comprehension tasks. More recently, [152] introduced a specialized recurrent layer that uses coreference annotations for improving reading comprehension tasks. On the language modeling task, [153] proposed a language model which can explicitly incorporate entities while dynamically updating their representations for a variety of tasks such as language modeling, coreference resolution, and entity prediction.

Our work builds upon and contributes to the growing literature on tracking state changes in procedural text. Bosselut *et al.* [143] presented a neural model that can learn to explicitly predict state changes of ingredients at different points in a cooking recipe. Mishra *et al.* [144] proposed another entity-aware model to track entity states in scientific processes. Tandon

et al. [145] demonstrated that the prediction quality can be boosted by including hard and soft constraints to eliminate unlikely or favor probable state changes. In follow-up work, [146] exploited the notion of label consistency in training to enforce similar predictions in similar procedural contexts. Das *et al.* [154] proposed a model that dynamically constructs a knowledge graph while reading the procedural text to track the ever-changing entities states. As discussed in the introduction, however, these previous methods use a strong inductive bias and assume that state labels are present during training. In our study, we deliberately focus on unlabeled procedural data and ask the question: Can multimodality help to identify and provide insights into understanding state changes?

In particular, we take advantage of the proposed RecipeQA dataset [128], for multimodal comprehension of cooking recipes, and ask whether it is possible to have a model which employs dynamic representations of entities in answering questions that require multimodal understanding of procedures. To this end, inspired from [155], we propose Procedural Reasoning Networks (PRN) [156] that incorporate entities into the comprehension process and allow us to keep track of entities, understand their interactions, and accordingly update their states across time. We report that our proposed approach significantly improves upon previously published results on visual reasoning tasks in RecipeQA, which test understanding causal and temporal relations from images and text. We further show that the dynamic entity representations can capture the semantics of the state information in the corresponding steps.

5.1. Visual Reasoning in RecipeQA

In our study, we particularly focus on the visual reasoning tasks of RecipeQA, namely *visual cloze*, *visual coherence*, and *visual ordering* tasks, each of which examines a different reasoning skill ⁶. We briefly describe these tasks below.

Visual Cloze. In the visual cloze task, the question is formed by a sequence of four images from consecutive steps of a recipe where one of them is replaced by a placeholder. A model

⁶We intentionally leave the textual cloze task out from our experiments as the questions in this task does not necessarily need multimodality.

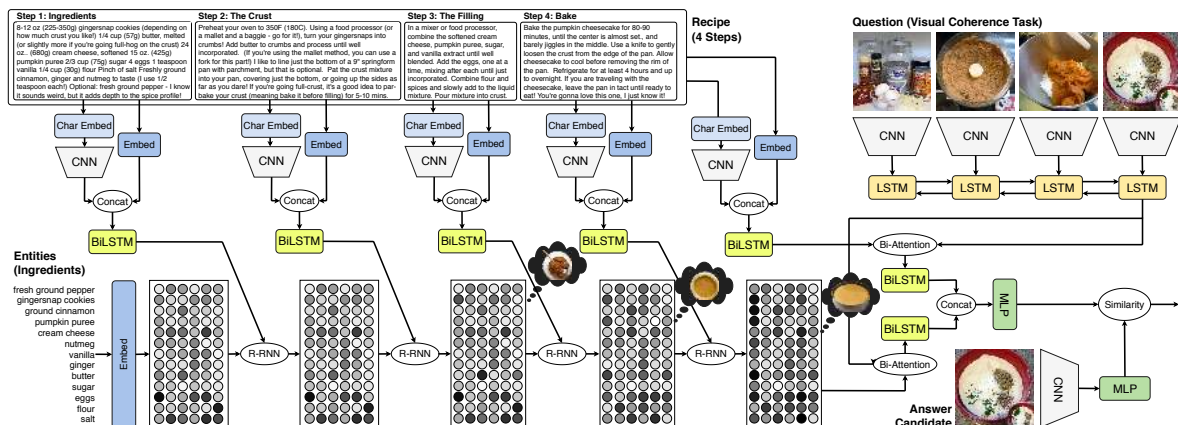


Figure 5.2 An illustration of our Procedural Reasoning Networks (PRN). For a sample question from the visual coherence task in RecipeQA, while reading the cooking recipe, the model constantly performs updates on the representations of the entities (ingredients) after each step and makes use of their representations along with the whole recipe when it scores a candidate answer. Please refer to the main text for more details.

should select the correct one from a multiple-choice list of four answer candidates to fill in the missing piece. In that regard, the task inherently requires aligning visual and textual information and understanding temporal relationships between the cooking actions and the entities.

Visual Coherence. The visual coherence task tests the ability to identify the image within a sequence of four images that is inconsistent with the text instructions of a cooking recipe. To succeed in this task, a model should have a clear understanding of the procedure described in the recipe and at the same time connect language and vision.

Visual Ordering. The visual ordering task is about grasping the temporal flow of visual events with the help of the given recipe text. The questions show a set of four images from the recipe and the task is to sort jumbled images into the correct order. Here, a model needs to infer the temporal relations between the images and align them with the recipe steps.

5.2. Procedural Reasoning Networks

In the following, we explain our Procedural Reasoning Networks model [156]. Its architecture is based on a bi-directional attention flow (BiDAF) model [157]⁷ but also equipped with an explicit reasoning module that acts on entity-specific relational memory units. Fig. 5.2 shows an overview of the network architecture. It consists of five main modules: An input module, an attention module, a reasoning module, a modeling module, and an output module. Note that the question-answering tasks we consider here are multimodal in that while the context is a procedural text, the question, and the multiple-choice answers are composed of images.

1. **Input Module** extracts vector representations of inputs at different levels of granularity by using several different encoders.
2. **Reasoning Module** scans the procedural text and tracks the states of the entities and their relations through a recurrent relational memory core unit [155].
3. **Attention Module** computes context-aware query vectors and query-aware context vectors as well as query-aware memory vectors.
4. **Modeling Module** employs two multi-layered RNNs to encode previous layers' outputs.
5. **Output Module** scores a candidate answer from the given multiple-choice list.

At a high level, as the model is reading the cooking recipe, it continually updates the internal memory representations of the entities (ingredients) based on the content of each step – it keeps track of changes in the states of the entities, providing an entity-centric summary of the recipe. The response to a question and a possible answer depends on the representation of the recipe text as well as the last states of the entities. All this happens in a series of implicit relational reasoning steps and there is no need for explicitly encoding the state in terms of a predefined vocabulary.

⁷Our implementation is based on the implementation publicly available in AllenNLP [157].

5.2.1. Input Module

Let the triple $(\mathbf{R}, \mathbf{Q}, \mathbf{A})$ be a sample input. Here, \mathbf{R} denotes the input recipe which contains textual instructions composed of N words in total. \mathbf{Q} represents the question that consists of a sequence of M images. \mathbf{A} denotes an answer that is either a single image or a series of L images depending on the reasoning task. In particular, for the visual cloze and the visual coherence type questions, the answer contains a single image ($L = 1$) and for the visual ordering task, it includes a sequence.

We encode the input recipe \mathbf{R} at character, word, and step levels. Character-level embedding layer uses a convolutional neural network, namely the CharCNN model by [158], which outputs character-level embeddings for each word and alleviates the issue of out-of-vocabulary (OOV) words. In the word embedding layer, we use a pre-trained GloVe model [159] and extract word-level embeddings (We also consider pre-trained ELMo embeddings [160] in our experiments but found out that the performance gain does not justify the computational overhead.). The concatenation of the character and the word embeddings are then fed to a two-layer highway network [161] to obtain a contextual embedding for each word in the recipe. This results in the matrix $\mathbf{R}' \in \mathbb{R}^{2d \times N}$.

On top of these layers, we have another layer that encodes the steps of the recipe in an individual manner. Specifically, we obtain a step-level contextual embedding of the input recipe containing T steps as $\mathcal{S} = (\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_T)$ where \mathbf{s}_i represents the final state of a BiLSTM encoding the i -th step of the recipe obtained from the character and word-level embeddings of the tokens exist in the corresponding step.

We represent both the question \mathbf{Q} and the answer \mathbf{A} in terms of visual embeddings. Here, we employ a pre-trained ResNet-50 model [162] trained on ImageNet dataset [163] and represent each image as a real-valued 2048-d vector using features from the penultimate average-pool layer. Then these embeddings are passed first to a multilayer perceptron (MLP) and then its outputs are fed to a BiLSTM. We then form a matrix $\mathbf{Q}' \in \mathbb{R}^{2d \times M}$ for the question by concatenating the cell states of the BiLSTM. For the visual ordering task, to represent the

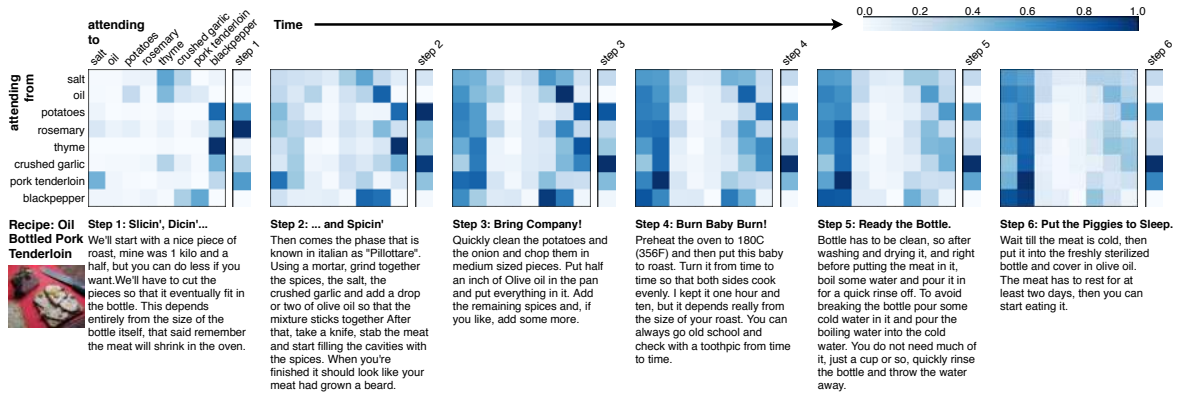


Figure 5.3 Sample visualizations of the self-attention weights demonstrating both the interactions among the ingredients and between the ingredients and the textual instructions throughout the steps of a sample cooking recipe from RecipeQA (darker colors imply higher attention weights). The attention maps do not change much after the third step as the steps after that mostly provide some redundant information about the completed recipe.

sequence of images in the answer with a single vector, we additionally use a BiLSTM and define the answering embedding by the summation of the cell states of the BiLSTM. Finally, for all tasks, these computations produce answer embeddings denoted by $\mathbf{a} \in \mathbb{R}^{2d \times 1}$.

5.2.2. Reasoning Module

As mentioned before, comprehending a cooking recipe is mostly about entities (basic ingredients) and actions (cooking activities) described in the recipe instructions. Each action leads to changes in the states of the entities, which usually affects their visual characteristics. A change rarely occurs in isolation; in most cases, the action affects multiple entities at once. Hence, in our reasoning module, we have an explicit memory component implemented with relational memory units [155]. This helps us to keep track of the entities, their state changes, and their relations in relation to each other over the course of the recipe (see Fig. 5.3). As we will examine in more detail in Section 5.3., it also greatly improves the interpretability of model outputs.

Specifically, we set up the memory with a memory matrix $\mathbf{E} \in \mathbb{R}^{d_E \times K}$ by extracting K entities (ingredients) from the first step of the recipe (The first steps of the recipes

in RecipeQA commonly contain a list of ingredients.). We initialize each memory cell \mathbf{e}_i representing a specific entity by its CharCNN and pre-trained GloVe embeddings (Multi-word entities (*e.g. minced garlic*) are represented by the average embedding vector of the words that they contain, and OOV words are expressed with the average word vector of all the words.). From now on, we will use the terms memory cells and entities interchangeably throughout the paper. Since the input recipe is given in the form of a procedural text decomposed into a number of steps, we update the memory cells after each step, reflecting the state changes that happened on the entities. This update procedure is modeled via a relational recurrent neural network (R-RNN), recently proposed by [155]. It is built on a 2-dimensional LSTM model whose matrix of cell states represents our memory matrix \mathbf{E} . Here, each row i of the matrix \mathbf{E} refers to a specific entity \mathbf{e}_i and is updated after each recipe step t as follows:

$$\phi_{i,t} = \text{R-RNN}(\phi_{i,t-1}, \mathbf{s}_t) \quad (2)$$

where \mathbf{s}_t denotes the embedding of recipe step t and $\phi_{i,t} = (\mathbf{h}_{i,t}, \mathbf{e}_{i,t})$ is the cell state of the R-RNN at step t with $\mathbf{h}_{i,t}$ and $\mathbf{e}_{i,t}$ being the i -th row of the hidden state of the R-RNN and the dynamic representation of entity \mathbf{e}_i at the step t , respectively. The R-RNN model exploits a multi-headed self-attention mechanism [164] that allows memory cells to interact with each other and attend multiple locations simultaneously during the update phase.

In Fig. 5.3, we illustrate how this interaction takes place in our relational memory module by considering a sample cooking recipe and by presenting how the attention matrix changes throughout the recipe. In particular, the attention matrix at a specific time shows the attention flow from one entity (memory cell) to another along with the attention weights to the corresponding recipe step (offset column). The color intensity shows the magnitude of the attention weights. As can be seen from the figure, the internal representations of the entities are actively updated at each step. Moreover, as argued in [155], this can be interpreted as a form of relational reasoning as each update on a specific memory cell is operated in relation to others. Here, we should note that it is often difficult to make sense of these attention weights. However, we observe that the attention matrix changes very gradually near the completion of the recipe.

5.2.3. Attention Module

The attention module is in charge of linking the question with the recipe text and the entities present in the recipe. It takes the matrices \mathbf{Q}' and \mathbf{R}' from the input module, and \mathbf{E} from the reasoning module and constructs the question-aware recipe representation \mathbf{G} and the question-aware entity representation \mathbf{Y} . Following the attention flow mechanism described in [165], we specifically calculate attention in four different directions: (1) from question to recipe, (2) from recipe to question, (3) from question to entities, and (4) from entities to question.

The first two of these attentions require computing a shared affinity matrix $\mathbf{S}^R \in \mathbb{R}^{N \times M}$ with $\mathbf{S}_{i,j}^R$ indicating the similarity between i -th recipe word and j -th image in the question estimated by

$$\mathbf{S}_{i,j}^R = \mathbf{w}_R^\top [\mathbf{R}'_i; \mathbf{Q}'_j; \mathbf{R}'_i \circ \mathbf{Q}'_j] \quad (3)$$

where \mathbf{w}_R^\top is a trainable weight vector, \circ and $[\cdot]$ denote elementwise multiplication and concatenation operations, respectively.

Recipe-to-question attention determines the images within the question that is most relevant to each word of the recipe. Let $\tilde{\mathbf{Q}} \in \mathbb{R}^{2d \times N}$ represent the recipe-to-question attention matrix with its i -th column being given by $\tilde{\mathbf{Q}}_i = \sum_j \mathbf{a}_{ij} \mathbf{Q}'_j$ where the attention weight is computed by $\mathbf{a}_i = \text{softmax}(\mathbf{S}_i^R) \in \mathbb{R}^M$.

Question-to-recipe attention signifies the words within the recipe that have the closest similarity to each image in the question, and construct an attended recipe vector given by $\tilde{\mathbf{r}} = \sum_i \mathbf{b}_i \mathbf{R}'_i$ with the attention weight is calculated by $\mathbf{b} = \text{softmax}(\max_{col}(\mathbf{S}^R)) \in \mathbb{R}^N$ where \max_{col} denotes the maximum function across the column. The question-to-recipe matrix is then obtained by replicating $\tilde{\mathbf{r}}$ N times across the column, giving $\tilde{\mathbf{R}} \in \mathbb{R}^{2d \times N}$.

Then, we construct the question-aware representation of the input recipe, \mathbf{G} , with its i -th column $\mathbf{G}_i \in \mathbb{R}^{8d \times N}$ denoting the final embedding of i -th word given by

$$\mathbf{G}_i = [\mathbf{R}'_i; \tilde{\mathbf{Q}}_i; \mathbf{R}'_i \circ \tilde{\mathbf{Q}}_i; \mathbf{R}'_i \circ \tilde{\mathbf{R}}_i;] \quad (4)$$

Attention from question to entities, and from entities to question is computed in a way similar to the ones described above. The only difference is that it uses a different shared affinity matrix to be computed between the memory encoding entities \mathbf{E} and the question \mathbf{Q}' . These attentions are then used to construct the question-aware representation of entities, denoted by \mathbf{Y} , that links and integrates the images in the question and the entities in the input recipe.

5.2.4. Modeling Module

The modeling module takes the question-aware representations of the recipe \mathbf{G} and the entities \mathbf{Y} and forms their combined vector representation. For this purpose, we first use a two-layer BiLSTM to read the question-aware recipe \mathbf{G} and to encode the interactions among the words conditioned on the question. For each direction of BiLSTM, we use its hidden state after reading the last token as its output. In the end, we obtain a vector embedding $\mathbf{c} \in \mathbb{R}^{2d \times 1}$. Similarly, we employ a second BiLSTM, this time, over the entities \mathbf{Y} , which results in another vector embedding $\mathbf{f} \in \mathbb{R}^{2d_E \times 1}$. Finally, these vector representations are concatenated and then projected to a fixed size representation using $\mathbf{o} = \varphi_o([\mathbf{c}; \mathbf{f}]) \in \mathbb{R}^{2d \times 1}$ where φ_o is a multilayer perceptron with tanh activation function.

5.2.5. Output Module

The output module takes the output of the modeling module, encoding vector embeddings of the question-aware recipe and the entities \mathbf{Y} , and the embedding of the answer \mathbf{A} , and returns a similarity score which is used while determining the correct answer. Among all the candidate answers, the one having the highest similarity score is chosen as the correct answer. To train our proposed procedural reasoning network, we employ a hinge ranking loss [136], similar to the one used in [128], given below.

$$L = \max\{0, \gamma - \cos(\mathbf{o}, \mathbf{a}_+) + \cos(\mathbf{o}, \mathbf{a}_-)\} \quad (5)$$

where γ is the margin parameter, a_+ and a_- are the correct and the incorrect answers, respectively.

5.3. Experiments

In this section, we describe our experimental setup and then analyze the results of the proposed Procedural Reasoning Networks (PRN) model.

5.3.1. Entity Extraction

Given a recipe, we automatically extract the entities from the initial step of a recipe by using a dictionary of ingredients. While determining the ingredients, we exploit Recipe1M [166] and Kaggle What’s Cooking Recipes [167] datasets, and form our dictionary using the most commonly used ingredients in the training set of RecipeQA. For the cases when no entity can be extracted from the recipe automatically (20 recipes in total), we manually annotate those recipes with the related entities.

5.3.2. Training Details

In our experiments, we separately trained models on each task, as well as we investigated multi-task learning where a single model is trained to solve all these tasks at once. In total, the PRN architecture consists of ~ 12 M trainable parameters. We implemented our models in PyTorch [168] using AllenNLP library [157]. We used the Adam optimizer with a learning rate of $1e - 4$ and an early stopping criteria with the patience set to 10 indicating that the training procedure ends after 10 iterations if the performance would not improve. We considered a batch size of 32 due to our hardware constraints. In the multi-task setting, batches are sampled round-robin from all tasks, where each batch is solely composed of examples from one task. We performed our experiments on a system containing four NVIDIA GTX-1080Ti GPUs, and training a single model took around 2 hours. We employed

the same hyperparameters for all the baseline systems. We plan to share our code and model implementation after the review process.

5.3.3. Baselines

We compare our model with several baseline models as described below. We note that the results of the first two are previously reported in [128].

Hasty Student [128] is a heuristics-based simple model which ignores the recipe and gives an answer by examining only the question and the answer set using distances in the visual feature space.

Impatient Reader [28] is a simple neural model that takes its name from the fact that it repeatedly computes attention over the recipe after observing each image in the query.

BiDAF [165] is a strong reading comprehension model that employs a bi-directional attention flow mechanism to obtain a question-aware representation and bases its predictions on this representation. Originally, it is a span-selection model from the input context. Here, we adapt it to work in a multimodal setting and answer multiple-choice questions instead.

BiDAF w/ static memory is an extended version of the BiDAF model which resembles our proposed PRN model in that it includes a memory unit for the entities. However, it does not make any updates on the memory cells. That is, it uses the static entity embeddings initialized with GloVe word vectors. We propose this baseline to test the significance of the use of relational memory updates.

5.3.4. Results

Table 5.1 presents the quantitative results for the visual reasoning tasks in RecipeQA. In the single-task training setting, PRN gives state-of-the-art results compared to other neural models. Moreover, it achieves the best performance on average. These results demonstrate the importance of having a dynamic memory and keeping track of entities extracted from

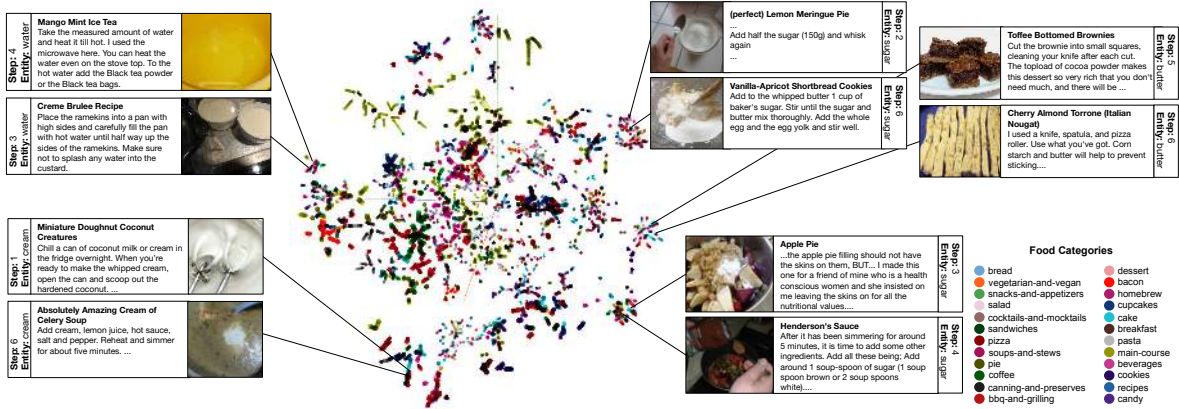


Figure 5.4 t-SNE visualizations of learned embeddings from each memory snapshot mapping to each entity and their corresponding states from each step for visual cloze task.

Table 5.1 Quantitative comparison of the proposed PRN model against the baselines.

Model	Single-task Training				Multi-task Training			
	Cloze	Coherence	Ordering	Average	Cloze	Coherence	Ordering	All
Human*	77.60	81.60	64.00	74.40	—	—	—	—
Hasty Student	27.35	65.80	40.88	44.68	—	—	—	—
Impatient Reader	27.36	28.08	26.74	27.39	—	—	—	—
BIDAF	53.95	48.82	62.42	55.06	44.62	36.00	63.93	48.67
BIDAF w/ static memory	51.82	45.88	60.90	52.87	47.81	40.23	62.94	50.59
PRN	56.31	53.64	62.77	57.57	46.45	40.58	62.67	50.17

* Taken from the RecipeQA project website, based on 100 questions sampled randomly from the validation set.

the recipe. In the multi-task training setting where a single model is trained to solve all the tasks at once, PRN and BIDAF w/ static memory perform comparably and give much better results than BIDAF. Note that the model performances in the multi-task training setting are worse than single-task performances. We believe that this is due to the nature of the tasks that some are more difficult than the others. We think that the performance could be improved by employing a carefully selected curriculum strategy [169].

In Fig. 5.4, we illustrate the entity embeddings space by projecting the learned embeddings from the step-by-step memory snapshots through time with t-SNE to 3-d space from 200-d vector space. Color codes denote the categories of the cooking recipes. As can be seen, these step-aware embeddings show clear clustering of these categories. Moreover, within each cluster, the entities are grouped together in terms of their state characteristics. For instance, in the zoomed parts of the figure, chopped and sliced, or stirred and whisked entities are

placed close to each other.




onions (Flowerpot Chicken) Step 1: This is a cheap and easy method of an ancient cooking technique known as clay pot cooking using a common terra cotta flowerpot and saucer. You can spend over \$100 on a clay cooker at a gourmet kitchen gadget store, or about \$20 at a garden supply. You choose. Some of you may already have the pot lying in your yard, garage or shed. Once you try this you will probably be cooking all kinds of things in it!	onions (Flowerpot Chicken) Step 3: Prepare Vegetables. <u>Chop</u> your vegetables while the pot is soaking. You can use whatever you like for this, root vegetables mixed with onions are always a nice base. This time I used leeks, bell peppers, garlic and red onions.	tomatoes (Flowerpot Chicken) Step 1: This is a cheap and easy method of an ancient cooking technique known as clay pot cooking using a common terra cotta flowerpot and saucer. You can spend over \$100 on a clay cooker at a gourmet kitchen gadget store, or about \$20 at a garden supply. You choose. Some of you may already have the pot lying in your yard, garage or shed. Once you try this you will probably be cooking all kinds of things in it!	
tomatoes (Chilli Con Carne) Step 1: Prepping the Vegetables. The first step is to have all the Vegetables prepped and ready to go in the pan, so finely <u>dice</u> the Garlic, onions and Peppers. Don't worry about <u>mixing</u> them up in the bowl, all of these items are going to be <u>sautéed</u> in a small amount of oil at the next stage. Picture 1. Finely <u>dice up</u> the Garlic, you want it to be almost puree consistency. Picture 2. Finely <u>dice up</u> the Onions, this doesn't need to be as fine as the garlic but you should ensure that they are all roughly the same size. Picture 3. Lastly <u>dice up</u> the bell pepper, I show you how I <u>cut</u> this in the video, but I will go over it quickly. Firstly I take off the four walls of the pepper, <u>flatten</u> them then cut them in to strips, then simply <u>cut</u> the other way so I have them diced.	tomatoes (Seven Layer Seven Grain Bread) Step 1: Ingredients ... pepperoni (I used what was left in a package which was enough for one layer) 1/2 onion 2 roma tomatoes <u>dried</u> rosemary <u>shredded</u> mozzarella and parmesan fresh savory, basil, tarragon, and thyme 2 or 3 cloves of garlic salt (sea or kosher salt are best) and pepper <u>Slice</u> the tomatoes and onion as thin as is reasonable, <u>slice</u> the garlic as thin as possible. Thoroughly <u>wash</u> the fresh herbs and <u>pull</u> the leaves from the stems. Discard the stems.	tomatoes (How to Make Chicken Cacciatore) Step 1: Gather Your Ingredients... ... 1 teaspoon <u>dried</u> oregano, 1/8 teaspoon red pepper flakes (see step five for a bit of humor on this note), 3/4 to 1 cup wine - Honestly, folks, don't be too particular about the wine. Red or white is fine. (you may substitute chicken broth, or even <u>add</u> broth in addition to the wine. Be creative!) 1 - 28 ounce can <u>diced</u> tomatoes (save the juice!) 1/2 teaspoon <u>dried</u> Porcino mushrooms (Optional, see step #2)	
water (Caribbean Curried Goat) Step 1: This is absolutely mind-blowingly good. Goat basically tastes like lamb, but is far leaner. (Lamb is the fattest of the red meats.) It's very popular in a variety of different countries' cuisines, but for some reason has yet to gain a real following in the US. This recipe is inspired by the curried goat roti from Penny's Caribbean Cafe. While Penny doesn't share her secrets, this tastes awfully similar. Go get yourself some goat (or lamb if you must) and try it out!	water (Caribbean Curried Goat) Step 4: Add Everything Else. <u>Add</u> the rest of the curry powder and stir things about. When it starts to <u>stick</u> again <u>add</u> the water and deglaze again. <u>Pour</u> in just enough water to cover the meat, and leave a cup full of water near the pot to refill as it boils off. You want the meat to stay wet during the entire cooking process. In the picture below I've dropped in another bouillon cube because they didn't all make it in with the onions. The details really don't matter too much in this dish - it <u>cooks</u> long enough that you've got LOTS of leeway to taste and modify.	milk (Birdcage-BQ) Step 1: All that sounded logic to me, and instead of looking on the net how others did it I started thinking how Bricobart would build such a device - I mean a bbq, not an anti-trill gun. And since I didn't want to spend any money I decided to build it from scratch. The project failed in the first trial, but ran like a small dog chased by a beeswarm in the second. Enjoy my poor men's vertical birdcage-based bbq!	
milk (Potato Soup for One) Step 3: Cooking. <u>Melt</u> the butter and add 1/3 cup <u>chopped</u> onions. When the onions are <u>cooked</u> <u>add</u> the bacon bits. Now <u>add</u> the potatoes back to the pot and <u>mash</u> the potato mixture. I use a potato masher or you can just use a fork. You still want it lumpy but the potatoes will help <u>thicken</u> the soup. <u>Pour</u> the milk and mix well. <u>Add</u> salt and pepper and heat until it is a slow <u>boil</u> . <u>Remove</u> from the stove and <u>add</u> the cheese and <u>stir</u> until melted. If you <u>add</u> the cheese too early it will go to the bottom and burn	milk (Family Size Lasagne) Step 2: Meat Sauce <u>Preheat</u> oven to 180 degrees celsius. <u>Brown off</u> the mince in a large pan, depending on the fat content of the meat, you may or may not need a little oil. <u>Drain</u> the mince onto some paper towel to <u>remove</u> any oil and then <u>place</u> back in the pan. <u>Add</u> 4 slices of chopped prosciutto (or bacon/panetta) and fry for a few minutes. <u>Add</u> beef stock, tomato sauce, nutmeg, bayleaf and oregano. <u>Simmer</u> for at least 30 minutes.	milk (Potato Soup) Step 1: Potato Prep + Seasonings Make sure all potatoes are <u>peeled</u> and <u>cut</u> into chunks. In a saucepan over medium heat, <u>dice</u> in the tablespoon of butter, the red pepper flakes and Italian seasoning. Let the butter <u>melt</u> and <u>stir</u> the seasonings around until they start smelling nice. :)	

Figure 5.5 Step-aware entity representations can be used to discover the changes that occurred in the states of the ingredients between two different recipe steps. The difference vector between two entities can then be added to other entities to find their next states. For instance, in the first example, the difference vector encodes the chopping action done on onions. In the second example, it encodes the pouring action done on the water. When these vectors are added to the representations of raw tomatoes and milk, the three most likely next states capture the semantics of state changes in an accurate manner.

Fig. 5.5 demonstrates the entity arithmetics using the learned embeddings from each entity step. Here, we show that the learned embedding from the memory snapshots can effectively capture the contextual information about the entities at each time point in the corresponding step while taking into account of the recipe data. This basic arithmetic operation suggests that the proposed model can successfully capture the semantics of each entity's state in the corresponding step⁸.

⁸We used Gensim for calculating entity arithmetics using cosine distances between entity embeddings.

5.4. Further Analysis

In the following, we provide implementation details and further analysis we did for the proposed Procedural Reasoning Networks model.

Entity Extraction. While initializing memory cells we used ingredients as entities and during each read we updated the memory cells, resulting in an implicit update for entity states. In that regard, detecting ingredients has been an important part of our experiments.

We followed a simple scheme for extracting entities from each recipe. We used Recipe1M [166] and Kaggle What’s Cooking Recipes [167] datasets and formed a dictionary using the ingredients provided in those datasets from the food domain. Next, we scanned each recipe for existing ingredients using this dictionary.

In RecipeQA often the first steps contain ingredients of that recipe, but in some cases, ingredients might not be explicitly provided the first step. In order to address this we calculated the maximum number of unique ingredients in each step of a recipe and selected that step to extract ingredients. In cases ingredients can not be automatically extracted we manually annotated those recipes. In particular, we annotated 20 recipes manually.

Training Details. For training our model architecture in multitask mode, we used random shuffling for each batch. This is as well true for training in single mode, but by doing so we tried to reduce the task bias that sequential training might lead to.

Analyses. For t-SNE embedding analyses, we used Embedding Projector⁹ project. We used the embeddings learned by the model extracted from memory cells of each time step during each read. Each embedding corresponds to an entity state. In particular, we used entity embeddings and trained a t-SNE model with the following parameters perplexity=30, learning rate=10 for 1000 iterations. Due to a limitation in the embedding projector, the project downsamples data points to 10K, hence we used only 10K entity embeddings in our t-SNE analyses.

⁹<https://projector.tensorflow.org>

Hyperparameters. The hyperparameter we used to train our model is listed in Table 5.2.

Table 5.2 Training hyperparameter selection for our experiments.

Hyperparameter	Value
Optimizer	Adam
Batch Size	32
Early Stopping	Yes
Learning Rate	1-e4
Patience	10
Memory Cell Size	61

5.5. Discussions

In this chapter, we presented a new neural architecture called Procedural Reasoning Networks (PRN) [156] for the multimodal understanding of step-by-step instructions. Our proposed model is based on the successful BiDAF framework but is also equipped with an explicit memory unit that provides an implicit mechanism to keep track of the changes in the states of the entities over the course of the procedure. Our experimental analysis on visual reasoning tasks in the RecipeQA dataset shows that the model significantly improves the results of the previous models, indicating that it better understands the procedural text and the accompanying images. Additionally, we carefully analyze our results and find that our approach learns meaningful dynamic representations of entities without any entity-level supervision. Although we achieved state-of-the-art results on RecipeQA, clearly there is still room for improvement compared to human performance. We also believe that the PRN architecture will be of value to other visual and textual sequential reasoning tasks.

Much recently, the RecipeQA dataset attracted a lot of attention from both computer vision and natural language processing communities. There has been a body of research inspired by the RecipeQA dataset in various setups such as multimodal comprehension, procedural understanding, and commonsense reasoning, information retrieval, image-text coherence and cross-modal representation learning. In the following, we discuss the related follow-up

studies that report results on RecipeQA as well as utilize the RecipeQA dataset for solving various research problems.

[170] proposed CITE, based on the RecipeQA dataset, and investigated multimodal discourse relations between image and text pairs. Authors leveraged recipes that have a one-to-one correspondence between each instruction and image to study the relationship between instructions and images with a broader goal of better understanding natural communication and commonsense reasoning. [171] proposed a method to discover image-sentence relationships without relying on explicit multimodal annotation in training for unsupervised discovery of multimodal links in multi-image, multi-sentence documents using the RecipeQA dataset. [172] explored the RecipeQA dataset and studied the inferential relations between images and text from the perspective of captioning by looking at the verb usages in text and the coherence of images and text co-occurring with images. [173] proposed a method to match each activity in a sequence to the corresponding instruction in a given document of instruction sequences. The proposed model regularizes the partial order of activities to match the order of instructions. The authors evaluate the performance of the proposed model on the RecipeQA textual cloze task and propose two other ordering tasks using RecipeQA. [174] proposed a Multi-Level Multi-Modal Transformer (MLMM-Trans) framework for reasoning tasks on RecipeQA by modeling interactions between different modalities at multiple steps. [175] proposed a dataset for sequence-to-sequence retrieval task. The authors used the RecipeQA dataset to augment the proposed dataset with the main objective to retrieve sequence of images which best describes and aligns with the given textual query. [176] proposed an alignment mechanism for procedural reasoning on RecipeQA. The authors investigate latent alignment space and the positional encoding of questions and answers to constrain the output space of the latent alignment. They leverage cross-modal representation based on cross-attention and analyze the benefits of information flow between images and instructions. [177] explore uncertainty measures for image-caption embedding-and-retrieval task on the RecipeQA dataset to assess sample reliability. The authors proposed a Bayesian neural network to quantify feature uncertainty and posterior uncertainty with a broader goal to improve retrieval performance by rejecting uncertain

queries. [178] proposed a heuristic to reduce the inherent bias present in the RecipeQA dataset. In particular, authors were inspired by the Hasty Student baseline model which answers questions by only looking at the answers while ignoring the question completely. The authors discuss that the question generation scheme in the RecipeQA dataset introduces a distributional bias, such that the beginning and final step titles often contain ‘Ingredients’ and ‘Enjoy’ therefore might lead models to only choose answers that contain these words without looking at the context, hence proposed a method to overcome this shortcoming. [179] proposed a method to model temporal structures for recipes in RecipeQA. In particular, the authors propose a knowledge-based deep heterogeneous graph matching model to reduce the divergence between recipe representations of entities across temporal changes and guarantee neighbourhood consensus through a graph based approach. [180] proposed a cross-modal coherence model for text-to-image retrieval task for a joint understanding of co-occurring images and text and proposed CITE++ by extending the CITE dataset which is based on the RecipeQA dataset. [181] proposed Meta-RecipeQA to reduce the inherent biases in the RecipeQA dataset and proposed a model to solve comprehension and reasoning tasks on the RecipeQA dataset. The authors analyzed distance distributions between correct answers and incorrect answers and proposed methods to reduce the bias in question-answering tasks. [182] demonstrated a heuristic to construct rich recipe representations in the form of plans leveraging the recipes in the RecipeQA dataset. The authors enriched the recipes in RecipeQA by adding allergens, activity, objects, background knowledge such as tools and possible failures for each step of the recipes with a broader goal to improve retrieval efficiency. [183] proposed a method for sequencing unordered multimodal procedural instructions and utilized RecipeQA for multimodal event sequencing task. In particular, the authors proposed different techniques to align images and text to improve sequencing of unordered steps in multimodal instructions where images and text complement each other. [184] proposed a method to model entities in both their temporal and cross-modal relations and leveraged the temporal nature of the RecipeQA dataset as well as its multimodal nature. The authors examine entities both in textual modality and visual modality and encode entity relations through temporal relations as well as model their cross-modal relations in the recipes.

In summary, the RecipeQA dataset has been inspiring several researchers to investigate various aspect of the dataset and has become a traditional multimodal dataset in the literature. We hope that our work continues to attract more attention from different fields and more researchers can exploit RecipeQA for procedural understanding of multimodal how-to instructions.

6. COMPOSITIONAL GROUNDED UNDERSTANDING OF ACTIONS

In this chapter, we explore the multimodal understanding of actions in a challenging compositional generalization setup which we surveyed thoroughly in Ch. 3.. In particular, we review the problem of compositional understanding of actions and propose models to show the contribution of multimodality towards models’ compositional generalization abilities, based on our work¹⁰. We only made minor changes to fit the text in the narrative of the thesis and made small corrections in the text. In the following, we introduce the problem of compositional grounded understanding of actions and propose methods that can leverage multimodal signals to significantly improve models’ compositional generalization abilities which we believe will be a step towards solving the compositional generalization problem which is an open research problem.

Humans can rapidly understand new concepts, relying upon and combining context information with basic concepts from their existing knowledge. Compared to humans, neural network models trained over increasingly large datasets perform impressively well on a wide range of tasks, but they often fail to compositionally generalize to unseen concepts. In this study, we investigate compositionality and systematic generalization in a perceptually grounded setting by using a dataset of everyday household activities. This dataset depicts sequences of activities in pursuit of a wide variety of goals, *e.g. preparing celery, washing plates*. Each activity is represented with crowd-sourced utterances that describe different steps of the activity alongside the egocentric video frames and audio features. We evaluate several unimodal and multimodal baselines on future utterance prediction and action anticipation tasks, that respectively aim at describing and predicting an activity involving novel compositions of seen concepts. The models that exploit visual and audio signals do indeed improve over text-only model when they are evaluated on the long tail of rare complex concepts.

¹⁰In submission

As a long-standing problem in language, compositionality has been widely studied for many years. It deals with describing the relationship between an unbounded number of sentences and a vast set of meanings from a finite set of rules [61]. Therefore, compositionality aims to address the problem of finding ways to define the meaning of an entire sentence as a function of the meaning of its constituents and the rules that are used to put those constituents together. In that regard, compositionality and systematic generalization have been used to characterize symbolic computation and human cognition [53, 63]. Humans demonstrate compositional ability in different domains such as natural language understanding, and visual scene understanding. As put by [74], “*Once a person learns the meaning of a new verb ‘dax’, he or she can immediately understand the meaning of ‘dax twice’ and ‘sing and dax’.*” Similarly, it has been shown that humans can learn a new object shape and understand its compositions using previously learned colors or concepts [59, 185].

Neural networks have been recently shown to perform well on many different tasks that require effective generalization abilities [52]. The compositionality and systematicity of neural networks have been long debated whether – and to what extent – neural networks display compositional generalization [53, 64, 66, 69, 70]. Moreover, deep neural networks have been commonly criticized for requiring a very large number of training examples to succeed and argued to lack compositional abilities [186]. Discussions around neural networks’ inability to capture the compositional structure of the underlying problem, thus failing to generalize compositionally have recently sparked a lot of interest in the machine learning community [51, 57, 59, 77, 79, 91, 187, 188].

Although recent studies towards understanding and improving the compositional generalization abilities of neural networks have garnered a lot of interest in the research community, little work has been done on the role of multimodal and grounded language processing. [74] investigated picking up new concepts and applying them in test time by coupling previously learned concepts with new concepts. [78] investigated language modeling for acquiring new words and predicting new compositions by learning text representations from visual context for understanding instructional videos. Other existing works are centered around designing conceptual benchmark datasets specifically constructed

for testing compositionality, *e.g.* [77, 91, 189]. These studies have demonstrated that many state-of-the-art deep models fail to capture the compositional structures in the underlying tasks and cannot generalize well even on really simple textual data.

Our motivation in this study is to test linguistic compositionality and systematic generalization in a perceptually grounded setting and to understand whether leveraging visual and auditory cues can contribute to systematic generalization capabilities of deep models. Towards this goal, we turn our attention to multimodal how-to instructions as they provide a good test bed for our needs. More concretely, we make the following contributions:

- We curate a new benchmark: Epic-Kitchens-100-Systematicity (EK-100-SYS) for future utterance prediction and action anticipation tasks, which can be used to analyze compositional generalization in a grounded setup.
- We implement several neural models, and through them, we analyze whether multimodality helps linguistic compositionality in the context of the proposed tasks.

6.1. Problem Formulation

6.1.1. Future Utterance Prediction Task

Predicting what comes next plays a central role in cognition [190, 191] and also has been attributed as an interesting training scheme from cognitive perspective [76]. We study this task on a multimodal dataset of people performing everyday household activities, *e.g.* *preparing celery*. Each video in our dataset consists of short clips (microsegments) that define sub-tasks of an activity, and each sub-task is described by a textual description, *e.g.* “pick up plate”, “put plate in sink”, “turn on water”, and “wash plate”, annotated by the actors after the recording. In the following, we formulate the future utterance prediction task as a language generation problem. Let $\mathcal{S} = (\mathbf{X}, \mathbf{V}, \mathbf{A})$ denote a triplet representing a short video clip with $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^K$ being a sequence of K utterances, which describe a household activity and grounded with visual and audio signals, denoted by $\mathbf{V} = \{\mathbf{v}_i\}_{i=1}^K$ and $\mathbf{A} = \{\mathbf{a}_i\}_{i=1}^K$, respectively. Our proposed future utterance prediction task involves generating
















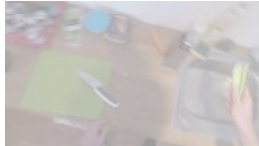
	Inputs (key frames and utterances)			Targets (future utterance)
Training				
	take celery	throw things into garbage bin	open fridge	put celery back into fridge
				
	pour sesame oil	close sesame oil	pick up onion	cut onion in half
				
	pick up containers	wash colander	put down colander	wash container
Evaluation				
	wash celery	close tap	put down celery	cut celery

Figure 6.1 Overview of our systematicity setup for future utterance prediction task using microsegments from our EK-100-SYS dataset. During training time, the model has already been exposed to the primitives ‘wash’, ‘close’, ‘put down’, and ‘celery’ but not the “cut celery” composition and the goal is to be able to generalize to novel compositions of primitive elements in test time.

the $(K + 1)^{th}$ utterance, $\mathbf{y} = \mathbf{x}_{K+1}$, following the preceding K utterances and multimodal cues. The training data consists of a set of sequences of microsegments, $\{(\mathcal{S}, \mathbf{y})\}$.

During training, the goal is to minimize the negative log-likelihood of generating the next utterance, where the multimodal models are conditioned on additional modalities such as image, or audio. Given the microsegment \mathcal{S} and the model parameters θ , our objective is to minimize the negative log-likelihood of all next utterance tokens $\mathbf{y} = \{y_i\}_{i=1}^m$:

$$\log p(\mathbf{y}|\mathcal{S}; \theta) = - \sum_{i=1}^m \log p(y_i|\mathcal{S}; \theta) \quad (6)$$

6.1.2. Action Anticipation Task

We formulate action anticipation as a classification task, where we study the problem of predicting the next action with the target verbs and nouns. The main difference between this task and the future utterance prediction task is that utterance prediction is a natural language generation task. In particular, in the action anticipation task, the goal is to predict the next action by leveraging the previously observed actions. Differing from other action anticipation tasks [7, 192, 193], our setup allows us to formulate action anticipation in a compositional manner by predicting the verb and noun separately.

More formally, let $\mathcal{S} = (\mathbf{X}, \mathbf{V}, \mathbf{A})$ denote a triplet representing a video clip with $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^K$ representing a sequence of K utterances, which describe a household activity and grounded with visual and audio signals, denoted by $\mathbf{V} = \{\mathbf{v}_i\}_{i=1}^K$ and $\mathbf{A} = \{\mathbf{a}_i\}_{i=1}^K$, respectively. Our action anticipation task involves predicting the verb/noun in the $(K + 1)^{th}$ utterance, $\mathbf{y} = \mathbf{x}_{K+1}^C$, following the preceding K utterances and multimodal cues where C denotes the verb or noun class.

6.2. EK-100-SYS Dataset

We use the EPIC-Kitchens-100 dataset (EK-100) as the starting point for our experiments [7]. EK-100 contains first-person videos of unscripted daily kitchen activities in natural household environments. Each video \mathbf{V} is split into a sequence of shorter clips $\mathbf{v}_1, \dots, \mathbf{v}_k$, which have manually-annotated English narrations of the activities within clips denoted by $\mathbf{x}_1, \dots, \mathbf{x}_k$. The clips also have audio tracks $\mathbf{a}_1, \dots, \mathbf{a}_k$, which only contain sounds, *e.g.* a knife cutting an onion or a person opening the fridge. We denote a sequence of video clips – audio tracks – narrations as an *instance*.

Recall from Section 6.1.1. that our aim is to model sequences of video clips. Therefore, we select instances from the EK-100 dataset with a window of $K = 4$ clips: the first 3 clips are used for context, while the final clip is used for prediction. This results in 22,136 instances

that can be used for our experiments. In Fig. 6.1, we provide some examples, along with representative keyframes and their corresponding narrations.

6.2.1. Systematicity Splits (EK-100-SYS)

Given a dataset of video sequences, our main focus is to study how well models compositionally generalize to unseen combinations of concepts. In the same vein of compositional captioning [75] or novel compositions of object properties [60], we create a dataset where the distribution of the individual concepts is similar across the dataset, but the compositions of those concepts is different. Consider the example given in Fig. 6.1. The model has already seen the nouns CELERY, TAP and verbs WASH, CLOSE, PUT DOWN, CUT but it has not seen the combination of CUT CELERY during training. The model has to compositionally generalize to this new instance.

To obtain such splits of the dataset, we followed the *Maximum Compound Divergence* heuristic to create similar distributions of individual concepts (atoms) but different distributions of combinations of concepts [77]. We use the 97 verb classes and 300 noun classes in the EK-100 dataset as the atoms. In particular, we assign each sample to a split based on the atomic and compound divergence (similarity) based on weighted distributions using Chernoff coefficient $C_\alpha(P\|Q) = \sum_k p_k^\alpha q_k^{1-\alpha} \in [0, 1]$ [194]. To make atom distributions similar in train and test, we use $\alpha = 0.5$ for atom divergence. Here, we set $\alpha = 0.1$ to reflect that it is more important for a certain compound to be found in P (train) rather than the probabilities in P (train) and Q (test) match exactly. Following this logic, we define compound divergence, and atom divergence for a train set U and test set W as follows:

$$\mathcal{D}_C(U\|W) = 1 - C_{0.1}(\mathcal{F}_C(U) \parallel \mathcal{F}_C(W))$$

$$\mathcal{D}_A(U\|W) = 1 - C_{0.5}(\mathcal{F}_A(U) \parallel \mathcal{F}_A(W))$$

where $\mathcal{F}_A(T)$ denotes frequency distribution of atoms, and $\mathcal{F}_C(T)$ denotes the distribution of compounds for a given set T and \mathcal{D}_A and \mathcal{D}_C denote atom and compound divergences,

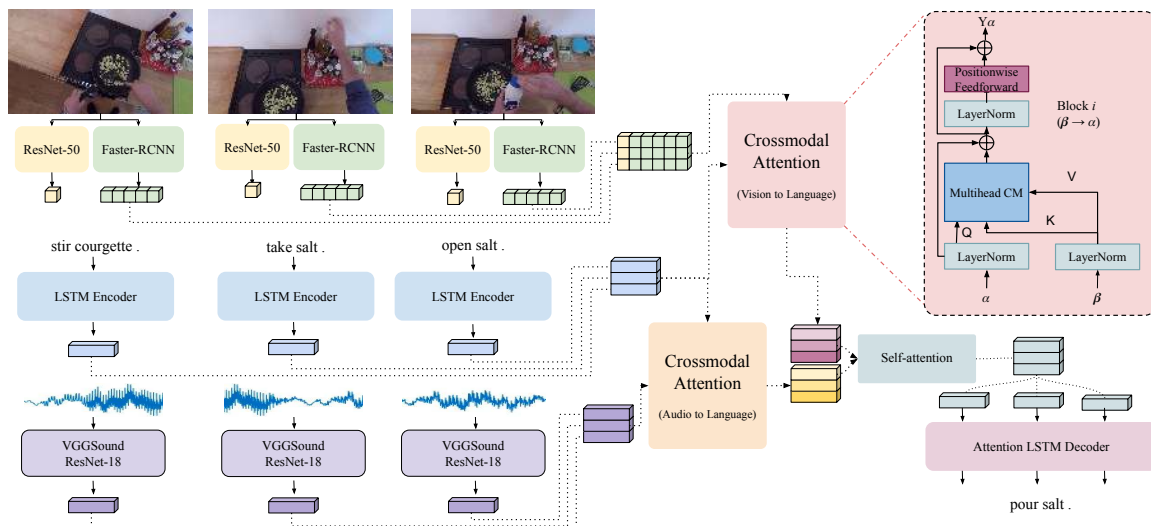


Figure 6.2 Overview of multimodal Audio, Vision and Language baseline model which incorporates global image features, object level image features, audio features as well as textual features using two crossmodal self-attention blocks with an LSTM decoder to predict next utterances.

respectively. We calculated divergence scores for each data sample until the atomic divergence of train and test set $D_A < 0.02$ and compound divergence of train and test set $D_C < 0.6$, which represents a sweet spot in terms of target distributions of atoms and compounds in the train and test sets (see Fig. 6.4). Finally, we randomly divide this test set into two sets with similar distributions, one for validation and the other for testing.

The resulting EK-100-SYS dataset has 8,766 instances, which are split into 4,407 training, 2,184 validation, and 2,175 test instances. We use these splits to train and evaluate our models on the future utterance prediction and action anticipation tasks.

6.3. Models for Future Utterance Prediction

We benchmark a text-only model, along with several multimodal models to assess the importance of different modalities in systematic generalization.

6.3.1. Text-only Unimodal Baseline (L)

Our first baseline is a text-only model to assess potential biases in the dataset [195]. This model is a 1-layer attention-based encoder-decoder model based on Long-Short Term Memory networks (LSTM) [196] with a hidden size of 256 units. The decoder learns a time-step dependent context over the encoder hidden states [197]. The model is trained using only the textual utterances $\mathbf{x}_{1:K}$ from the microsegment as the input, and the next utterance \mathbf{x}_{K+1} as the target, *i.e.* to predict $p(\mathbf{x}_{K+1}|\mathbf{x}_{1:K})$. The model uses 200D word embeddings, over a vocabulary of unique words in training samples where vocabularies are not shared between the encoder and the decoder. We use the same textual encoder and decoder in all multimodal baselines (See Appendix A for the number of trainable parameters and the vocabulary sizes).

6.3.2. Multimodal Baselines

We also evaluate several multimodal baselines that operate over combinations of the textual, visual, and audio modalities, as illustrated in Fig. 6.2.

6.3.3. Vision and Language (VL)

Our Vision and Language baseline encodes both textual and visual context for the future utterance prediction task. In particular, the model encodes the textual utterances $\mathbf{x}_{1:K}$ of each action from microsegments and the keyframe images $\mathbf{v}_{1:K}$ to predict the next utterance \mathbf{x}_{K+1} , *i.e.* $p(\mathbf{y} = \mathbf{x}_{K+1}|\mathbf{x}_{1:K}, \mathbf{v}_{1:K})$. This model is adapted from a model that parses a visual scene and learns cross-modal self-attention [198] over textual inputs and visual data.

The visual inputs are encoded using pre-trained CNN, and the textual inputs are encoded using an LSTM. More specifically, for the visual modality, we extracted two types of features: one type represents global visual features, and the other represents object-level features. For the global features, we used a pre-trained ResNet50 model [162] with ImageNet

weights [135]. Object-level features were extracted using a pre-trained Faster-RCNN object detector [199] with a ResNet-101 backbone [162] which is pre-trained on MSCOCO [87] and finetuned on EK-100. We extract visual features from 5 objects for each keyframe. The resulting representation of a visual keyframe is the concatenation of the global and the object-level features. This concatenated vector is projected into a lower-dimensional space using a 1D convolution. The textual inputs are encoded using an LSTM with 200D word embeddings and a 256D hidden layer.

The visual and textual modalities are then encoded by a cross-modal self-attention mechanism, CM. In this model, we consider two modalities α and β , sequences of each modalities are denoted as $X_\alpha \in \mathbb{R}^{T_\alpha \times d_\alpha}$ and $X_\beta \in \mathbb{R}^{T_\beta \times d_\beta}$, respectively and $T_{(\cdot)}$ denotes sequence length and $d_{(\cdot)}$ denotes feature dimension. In this model, α is the language modality, and β is the visual modality. In the cross-modal attention, the textual features are the *keys*, and the visual features are the *queries* and *values*, for aligning visual features to textual features. Let the Query be defined as $Q_\alpha = X_\alpha W_{Q_\alpha}$, the Keys as $K_\beta = X_\beta W_{K_\beta}$, and the Values as $V_\beta = X_\beta W_{V_\beta}$, where $W_{Q_\alpha} \in \mathbb{R}^{d_\alpha \times d_k}$, $W_{K_\beta} \in \mathbb{R}^{d_\beta \times d_k}$ and $W_{V_\beta} \in \mathbb{R}^{d_\beta \times d_v}$ are learnable weights. The cross-modal self-attention from β to α is formulated as a latent adaptation $Y_\alpha \in \mathbb{R}^{T_\alpha \times d_v}$:

$$\begin{aligned} Y_\alpha &= \text{CM}_{\beta \rightarrow \alpha}(X_\alpha, X_\beta) \\ &= \text{softmax} \left(\frac{Q_\alpha K_\beta^\top}{\sqrt{d_k}} \right) V_\beta \end{aligned} \quad (7)$$

The output Y_α has the same length as Q_α , but it is represented in the feature space of V_β . This enables the model to fuse different modalities, learning an alignment between the visual and textual features. Finally, there is a self-attention layer [103] over the aligned vision and language features, which are the input to an attention-based LSTM decoder that generates the next utterance.

6.3.4. Audio and Language (AL)

The Audio and Language baseline has the same structure as the Vision and Language baseline. The key difference is that we represent the additional context using audio features instead of visual features. In particular, the model encodes both the textual utterances $\mathbf{x}_{1:K}$ and the accompanying audio data $\mathbf{a}_{1:K}$ to predict the next utterance \mathbf{x}_{K+1} , *i.e.* $p(\mathbf{x}_{K+1}|\mathbf{x}_{1:K}, \mathbf{a}_{1:K})$. The audio features are 512D vectors extracted using VGGSound [200], which is pre-trained on 200K videos from YouTube videos totaling up to 550 hours of audio data. Here, the model learns a cross-modal attention block over the audio and the textual features, analogously to using the visual and textual features, as inputs to an LSTM-based decoder.

6.3.5. Object and Language (OL)

The Object and Language baseline uses the same architecture as Vision and Language baseline, but we represent the visual context using the tags of the detected objects, instead of the CNN visual features to explicitly encode the visual content. In this model, we embed object tags as a secondary set of textual features to our model along with the input utterances. Here, the object tags are represented as 292-dimensional one-hot encoded vectors (based on the number of unique tags) and projected to 256D with a simple linear layer. In this case, the cross-modal attention mechanism aligns object tag features with language features.

6.3.6. Audio, Vision and Language (AVL)

In the Audio, Vision, and Language (AVL) baseline, we leverage the audio, visual, and textual data using two cross-modal self-attention blocks. We use textual utterances $\mathbf{x}_{1:K}$ of each action along with the visual features $\mathbf{v}_{1:K}$ from the keyframes, and the VGGSound audio features $\mathbf{a}_{1:K}$ to guess the future utterance \mathbf{x}_{K+1} , *i.e.* $p(\mathbf{y} = \mathbf{x}_{K+1}|\mathbf{x}_{1:K}, \mathbf{v}_{1:K}, \mathbf{a}_{1:K})$. In this model, the input to the self-attention layer before the decoder is the concatenation

of the audio-aligned textual features from the audio-textual cross-modal block with the visual-aligned textual features from the visual-textual cross-modal block.

6.3.7. Object, Audio and Language (OAL)

We perform an extra experiment to determine whether adding an extra modality to the OL baseline model improves its performance by coupling the object tags with audio features. Here, we include the extracted audio features from each microsegment to the OL model and train accordingly.

6.3.8. Pretrained Vision and Language (PVL)

In order to understand the importance of pretraining on a large scale aligned audio, visual and linguistic data, we also evaluate Merlot Reserve [201]. Merlot Reserve learns multimodal video representations over video frames, text, and audio. We extract multimodal vision and language features through its pre-trained encoder while considering the same decoder network as the other baseline models. We report our experimental results with this pre-trained transformer baseline.

6.4. Models for Action Anticipation

Different than the future utterance prediction task which is formulated as a conditional language generation problem, the action anticipation task requires (grounded) language understanding. We adapt the models described in the previous section for this task by slightly modifying their architectures. In particular, we replace the final layer in these models with two new fully connected layers and finetune the pre-trained models by considering a classification objective that involves predicting either the VERB or the NOUN in the anticipated action.

6.5. Experimental Setup

6.5.1. Evaluation Metrics

For evaluation, we used BLEU [39], Exact Match (EM), and Categorical Accuracy (CA) metrics. For BLEU, we use the NLTK toolkit and report unigram BLEU scores. For EM, we calculate an accuracy score between the generated text sequence and the groundtruth. CA uses the verb and noun categories in `EK-100` and calculates the categorization accuracy based on noun category match between the predicted sequence and groundtruth, *e.g.* the verbs *slice*, *dice*, and *chop* fall into the same verb category *cut*, and the nouns *mozzarella*, *paneer* and *parmesan* are grouped into the same noun category *cheese*.

6.5.2. Training Details

We used SGD optimizer with 0.9 momentum with an initial learning rate of $1e - 1$ and a batch size of 128 and used ReduceOnPlateau learning rate scheduler to reduce the learning rate during training when validation loss metric plateaus. To train the models for future utterance prediction, we employed cross-entropy loss, initialized network weights via uniform distribution and used a dropout rate of 0.3 for both the encoder and the decoder. We used an early stopping strategy and stopped the training if validation BLEU did not improve after a certain threshold (patience = 50). We clipped gradients and set the gradient threshold to 0.1, and used a 4-head multihead attention mechanism in the crossmodal self-attention block in all our multimodal models. As a preprocessing step, we replace multiword tokens with a single word. For example, each occurrence of “olive oil” is replaced with “olive_oil”. While training the models for action anticipation, we again used early stopping, but stopped the training if validation loss did not improve after a certain threshold (patience = 5).

6.6. Results

6.6.1. Future Utterance Prediction

Table 6.1 Future utterance prediction results. Using audio, visual, or object features always improves performance compared to the language-only unimodal baseline. We report the mean and the standard deviation across 3 runs.

Inputs	BLEU val	BLEU test	EM val	EM test	CA val	CA test
L	10.88 ± 0.7	10.55 ± 1.0	0.91 ± 0.4	0.61 ± 0.4	2.18 ± 0.6	2.16 ± 0.9
VL	19.49 ± 1.2	19.43 ± 0.7	4.60 ± 0.8	4.62 ± 0.4	8.62 ± 0.6	8.38 ± 0.3
AL	24.38 ± 0.9	25.21 ± 1.3	6.54 ± 0.4	6.29 ± 0.5	12.46 ± 0.7	11.67 ± 1.1
AVL	19.53 ± 0.1	20.10 ± 0.1	4.70 ± 0.1	4.72 ± 0.2	8.77 ± 0.3	9.13 ± 0.5
OL	24.36 ± 0.8	25.60 ± 1.4	6.13 ± 0.4	6.17 ± 0.3	12.28 ± 0.7	12.52 ± 0.4
OAL	24.97 ± 0.4	25.55 ± 0.5	6.17 ± 0.4	5.72 ± 0.6	12.39 ± 0.7	12.06 ± 1.0
PVL	26.48 ± 0.7	27.78 ± 0.5	6.73 ± 0.7	7.23 ± 0.5	13.27 ± 0.5	14.28 ± 0.7

Table 6.1 shows the results of the future utterance prediction experiments. As can be seen, all of the multimodal models outperform the language-only baseline. Models that use visual features (VL) improve by 9 BLEU, 4 EM, and 10 CA points compared to the language-only model. Using visual features with a pre-trained model brings the largest overall improvement in performance; BLEU, EM, and CA increase by approximately 17, 6, and 19 points, respectively. However, using the combination of audio, visual, and language features (AVL) or using audio features in addition to object tags (OAL) do not bring further improvements, highlighting the difficulty of fusing multiple modalities. The performance of PVL shows the benefit of using pre-trained multimodal representations, as opposed to fusing separate unimodal encodings.

Table 6.2 Object-level attention accuracy at different layers of the crossmodal attention block in the VL model.

Layer1	Layer2	Layer3	Layer4	Layer5
8.6	9.8	10.4	8.8	9.8

To further investigate the behavior of the VL baseline, we examine the role of different layers in the crossmodal attention block in predicting which object regions correspond to the nouns in the target utterance. A model succeeds in this prediction task if the attention

weight to the expected object region is maximized. In Table 6.2, we report these object-level attention accuracies. We observe that the middle layers and final layer maximize attention to the expected object region. This indicates that to achieve better generalization, the VL model needs to use the semantic information encoded in the latter layers of the crossmodal attention block.







Inputs (utterances and auxiliary modalities)				Prediction (future utterance)	
Image				GT	: place bowl
				L	: put fridge
				OL	: close bin
Text	clean bowl .	open dishwasher .	open drawer .	VL	: wash bowl
				AL	: close bowl
				AVL	: place bowl
				OAL	: place bowl
				PVL	: close bowl
Image				GT	: rinse fork
				L	: put_down bowl
				OL	: rinse fork
Text	turn_on tap .	rinse chopsticks .	take fork .	VL	: put fork
				AL	: put fork
				AVL	: place bowl
				OAL	: rinse fork
				PVL	: rinse fork

Figure 6.3 Future Utterance Prediction qualitative results. Each model considers different combinations of input modality, as described in Section 6.3.. The targets, *rinse fork* and *place bowl* have not been observed by during the training phase; the multimodal inputs from Object, Audio, and Language are needed to predict them.

In Fig. 6.3, we provide a qualitative comparison of the baseline models. In training time, *place bowl*, or *rinse fork* compounds have never been observed by the models. In both of these examples, text-only unimodal model fails to generalize to novel compositions whereas OAL baseline predicts the target composition correctly for both examples.

6.6.2. Action Anticipation

We report the results in Table 6.3 on the action anticipation task. We present the results for the val and test sets to demonstrate that this task poses a more challenging case for the models. The trend in performance is in line with those for the future utterance prediction task results. Even though there is not one particular model that outperforms the remaining models, multimodality improves the overall compositional generalization on the action anticipation

task. Once again, the Object and Language and the Pretrained Vision and Language model performs strong in both measures.

Table 6.3 Quantitative comparison of baselines for action anticipation task for predicting compound action (nouns and verbs). We report mean across three runs. Best and second-best performing results are bolded and underlined, respectively.

	EM val	EM test	CA val	CA test
L	2.32 ± 0.6	1.99 ± 0.3	6.34 ± 0.7	6.02 ± 0.9
VL	2.16 ± 0.2	2.02 ± 0.0	7.45 ± 0.3	7.12 ± 0.6
AL	5.03 ± 0.5	4.45 ± 0.9	12.04 ± 1.4	11.36 ± 2.6
AVL	3.23 ± 0.3	2.84 ± 0.6	9.95 ± 0.5	8.76 ± 0.9
OL	4.74 ± 1.0	<u>5.31 ± 0.5</u>	13.30 ± 1.6	<u>14.50 ± 1.2</u>
OAL	6.09 ± 0.7	5.14 ± 1.1	13.26 ± 0.4	12.48 ± 1.3
PVL	6.39 ± 0.5	6.68 ± 0.3	14.06 ± 0.1	14.92 ± 0.3

6.7. Related Work

Compositionality. Much recently, the compositionality problem has been investigated in various settings. [76] analyzed the capacity of artificial neural networks in linguistic compositionality. [73] examined systematicity and compositionality with a human-like number of examples. [75] investigated compositional generalization, in terms of a model’s performance to composing unseen combinations of concepts when describing images. [202] explored compositionality in sentence embeddings for understanding how words combine, for generalizing to unencountered words and phrases. [203] examined compositionality in sentence vector representations by probing for compositional information in embeddings using a number of existing sentence composition models. [204] analyzed measuring compositionality in the aspect of representation learning *e.g.* a learned embedding. [72] investigated systematic generalization in a VQA-like setting. [74] examined compositionality in terms of systematic generalization for a meta-learning setting. [205] considered learning entire rule systems from examples instead of learning to predict the correct output given a novel input. [206] studied the emergence of systematic generalization in a situated agent setting where the agent learns to perform tasks based on textual instructions and visual observations. [60] proposed gSCAN dataset by extending the SCAN dataset to a 2d grid world setting for situated language understanding using

grounded instructions. [207] suggested a transformer-based method for analogical reasoning in language acquisition setup coupled with visual data to pick up novel words.

Among the prior work, the closest work to ours is [78] in which the authors investigated compositionality and generalization in a novel word acquisition setting from narrated videos. They suggested to train a model using a masking strategy with a reference set where the model learns to map the masked words in the target tokens using the reference set examples in the same episode which is a image-text pair sequence. Another work that is closely similar is [208], where the authors create a setup to examine the systematic generalization in unseen compound acquisition setting from paired image-caption streams. They concluded that continual learning methods show little systematic generalization when trained in a shifting compound distribution. In our study, we focus on systematic generalization to novel compositions where models need to generalize to unseen compositions and hence learn how to learn primitive elements and concepts in the training set to generalize to novel compositions (*e.g.* see Fig.6.1), whereas in [78], the models learn to map target words from the reference set examples consisting of image-text pairs and [208] investigate how well continual learning models systematically generalize under shifting compound distributions in a visually grounded setting.

Visually Grounded Reasoning. Neural reasoning models have shown good generalization across biased data splits, towards addressing this problem. [59] proposed the CLEVR-CoGenT dataset derived from the CLEVR dataset to test the systematic generalization of models on visual reasoning tasks. [209] investigated how models generalize to the previously unseen scene for visually-grounded natural language navigation in real buildings. [210] studied how models translate grounded instructions to robot actions to accomplish household tasks in novel environments. Related to our work [211] predicted future utterances from multimodal data using instructional videos and used the transcribed speech as text where the goal is to rank correct utterance among candidates. However, we focus on predicting future utterances by generating utterances to assess systematic generalization abilities of models trained on different and multiple multimodalities.

6.8. Experimental Setup

Systematicity Split (EK-100-SYS). Fig. 6.4 illustrates the atomic and compound distributions over the constructed training, validation, and test splits of our proposed systematicity setup. As can be seen, while these splits have similar distributions over atoms, training and val/test splits do differ in terms of compounds.

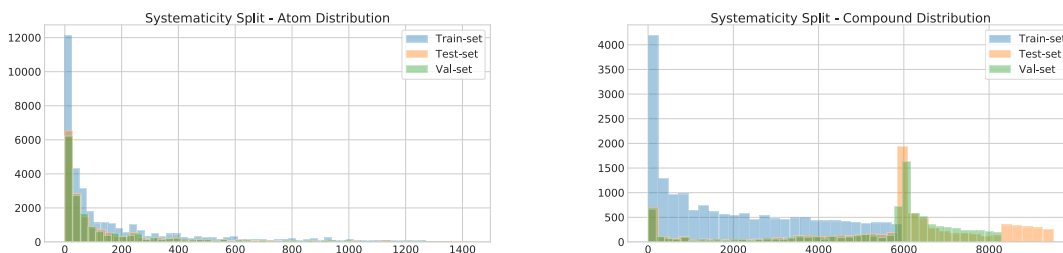


Figure 6.4 For train/val/test splits for systematicity split setup, the plot at the top demonstrates the distribution of atoms while the plot at the bottom shows the distribution of compounds.

Choosing Keyframes from Videos. In our experimental setup, we choose and use representative images from each microsegment. While choosing such a representative image for each video sequence, we follow a simple heuristics-based strategy. In particular, we run an object detector on the video frames and select the frames containing the maximum number of object proposals captured by the object detector as the representative frames.

Table 6.4 Model sizes and their training times along with the vocabulary sizes considered in our experiments.

Model	Parameters	Training Time
L	2,827,218	10 mins
OL	6,833,874	35 mins
VL	7,301,074	18 mins
AL	6,907,858	31 mins
AVL	15,331,026	37 mins
OAL	14,863,826	39 mins
PVL	6,973,394	31 mins

Model Sizes and Training Time. In Table 6.4, we present the number of trainable parameters and training time for all of our baseline models for the future utterance prediction task. All of our models are implemented with PyTorch and trained with Nvidia 1080Ti GPUs.

Further Analysis. In Table 6.5 we provide how models generalize for isolated verbs and in Table 6.6 we provide generalization performance for isolated nouns for the action anticipation task.

Table 6.5 Quantitative comparison of baselines for action anticipation task for predicting target verbs. We report mean across three runs. Best and second-best performing results are bolded and underlined, respectively.

	EM val	EM test	CA val	CA test
L	19.05 \pm 0.1	18.97 \pm 0.2	38.55 \pm 0.4	40.21 \pm 0.7
VL	21.15 \pm 0.3	19.69 \pm 0.5	38.64 \pm 1.0	37.67 \pm 1.2
AL	19.45 \pm 0.6	19.44 \pm 0.9	37.02 \pm 1.4	37.02 \pm 2.2
AVL	20.52 \pm 0.4	19.58 \pm 0.5	38.75 \pm 1.4	38.70 \pm 1.1
OL	18.94 \pm 0.6	18.79 \pm 0.2	38.51 \pm 1.1	<u>38.98</u> \pm 1.1
OAL	18.82 \pm 1.2	<u>19.59</u> \pm 0.7	36.09 \pm 1.1	37.17 \pm 1.8
PVL	18.88 \pm 0.6	19.36 \pm 0.2	35.00 \pm 0.9	36.23 \pm 1.1

Table 6.6 Quantitative comparison of baselines in the action anticipation task for predicting target nouns. We report mean across three runs. Best and second-best performing results are bolded and underlined, respectively.

	EM val	EM test	CA val	CA test
L	10.15 \pm 2.3	10.04 \pm 2.4	15.42 \pm 2.0	15.86 \pm 2.2
VL	11.87 \pm 0.5	12.2 \pm 1.1	19.94 \pm 0.2	21.17 \pm 0.8
AL	23.13 \pm 0.6	23.48 \pm 0.8	30.65 \pm 1.0	31.52 \pm 1.6
AVL	16.34 \pm 0.1	16.16 \pm 0.6	24.98 \pm 0.3	25.34 \pm 0.4
OL	27.08 \pm 0.8	28.12 \pm 0.6	35.49 \pm 0.5	36.49 \pm 0.1
OAL	28.81 \pm 0.4	<u>31.28</u> \pm 1.1	35.34 \pm 0.7	<u>38.03</u> \pm 1.0
PVL	31.64 \pm 2.8	33.33 \pm 2.0	37.45 \pm 3.4	39.06 \pm 1.7

6.9. Discussions

In this chapter, we presented an investigation of linguistic compositionality and systematic generalization in a perceptually grounded setting. We showed how a multimodal how-to instructions dataset can be utilized as a challenging test bed for this purpose. We designed the future utterance prediction and action anticipation tasks and followed a methodical approach in generating the training, validation, and test sets in our systematicity split. We

experimented with several baseline models and investigated models’ ability to generalize to novel compositions and showed how multimodal data can contribute towards solving systematic generalization problem. We found that a multimodal encoder pre-trained on video data gave the best generalization performance. Our findings indicate that the models that exploit visual and audio signals do indeed improve over text-only model when they are evaluated on the long tail of rare complex concepts. We hope our work will stimulate further research along these directions. That being said, the textual utterances that we consider in our work are simplistic and do not capture all of the complexities of natural language. Hence, extending this work to a more natural source of language data will be quite interesting. Furthermore, we highlight a few limitations of our work, which are listed below.

In our work, to analyze how visual and audio signals affect linguistic compositionality, we proposed a new dataset called `EK-100-SYS` that is curated from the EPIC-Kitchens-100 dataset [7]. As mentioned, this dataset involves videos including daily kitchen activities in natural household environments. Hence, it could be interesting to conduct future studies in an open-domain setting to alleviate limitations of the domain-specific nature of the EPIC-Kitchens-100 dataset.

We investigate several different multimodal models for both future utterance prediction and action anticipation tasks. However, it is important to note that for multimodal learning how to integrate different modalities is considered as an open research problem. In the literature, different strategies for multimodal data fusion have been proposed. Our experimental analysis could be further extended by considering some models that fuse the modalities in a way different than ours. More interestingly, from systematic generalization point of view, an analysis could be also carried out to explore the most effective fusion scheme.

7. CONCLUSION

In the last few years, there has been tremendous progress in both computer vision and natural language processing fields which is paving the way towards a joint understanding of language and vision tasks.

In this thesis, we explore the problem of understanding multimodal how-to instructions with images and text and argue that joint understanding of multimodal how-to instructions as the next frontier in AI research. We broadly define “multimodal machine comprehension” and investigate the proposed research problem from the point of comprehension, reasoning, and systematic generalization.

Our emphasis is on understanding multimodal how-to instructions such as cooking recipes with step-by-step instructions and analyzing everyday household task videos consisting of multiple modalities, which are rich in context, objects, scenes, interactions, and temporal relations and procedures. We analyze the multimodal machine comprehension problem through three different lenses:

- We propose a multimodal machine comprehension dataset consisting of cooking recipes with images and text (Chapter 4.)
- We propose a novel neural model for understanding multimodal procedures and for tracking the state changes and interactions between entities in recipes (Chapter 5.)
- We propose a systematic generalization setup leveraging a natural dataset consisting of household actions and show the contribution of grounding towards models’ compositional generalization abilities (Chapter 6.).

In particular, throughout this thesis, we thoroughly analyzed the multimodal machine comprehension problem for how-to instructions with images and text. Towards understanding multimodal how-to instructions we presented RecipeQA, a dataset for multimodal comprehension of cooking recipes, which consists of cooking recipes with

context-question-answer triplets. To our knowledge, RecipeQA is the first machine comprehension dataset that deals with understanding procedural knowledge in a multimodal setting. Each one of the four question styles in our dataset is specifically tailored to evaluate a particular skill and requires connecting the dots between different modalities. The results of our baseline models demonstrate that RecipeQA is a challenging dataset for multimodal machine comprehension. Furthermore, we proposed a new neural architecture called Procedural Reasoning Networks (PRN) for the multimodal understanding of step-by-step instructions. We presented a model which is equipped with an explicit memory unit that provides an implicit mechanism to keep track of the changes in the states of the entities over the course of the procedure. Using the proposed method, we significantly improved upon the previous results on visual reasoning tasks in the RecipeQA dataset, indicating that it better understands the procedural text and the accompanying images. Additionally, we carefully analyzed our results and found that our approach learns meaningful dynamic representations of entities without any entity-level supervision. We explored the compositional generalization problem which stands as a serious challenge to the real success of deep learning models and reviewed the current state of compositional generalization research from the point of datasets, tasks, and modeling methodologies and discussed the open research challenges and limitations in the existing literature to lay the groundwork for our research problem. Finally, we presented an investigation of linguistic compositionality and systematic generalization in a perceptually grounded setting. We showed how a multimodal how-to instructions dataset can be utilized as a challenging test bed for this purpose. We designed the future utterance prediction and action anticipation tasks and followed a methodical approach in generating the training, validation, and test sets for benchmarking models' compositional generalization abilities in a multimodal how-to instructions dataset. We experimented with several baseline models and investigated models' ability to generalize to novel compositions and showed how grounding linguistic data with auxiliary multimodal data can contribute towards solving systematic generalization problem. We showed that multimodality can indeed improve models' compositional generalization performance even if models are not exposed to certain compositions during training time and can better generalize to novel compositions thanks to grounding textual data with auxiliary

modalities such as images and audio.

In conclusion, throughout this thesis, we investigated the multimodal understanding of how-to instructions from various perspectives, reviewed the current state of the art in the literature, and proposed methods and models to show how multimodality helps models generalization performance in various reasoning and comprehension tasks on different datasets and showed how linguistic components grounded with different modalities can help models to systematically generalize to never seen compositions even models were not exposed to these compositions during training time.

7.1. Future work and open directions

In this thesis, we proposed two new datasets, namely the `RecipeQA` dataset, curated from online cooking recipes, and `EK-100-SYS`, curated from the EPIC-Kitchens-100 dataset consisting of videos including daily kitchen activities. An interesting extension to our work would be to explore different domains beyond cooking recipes, and daily household tasks, such as building a larger-scale how-to instructions benchmark dataset to include broader categories and different domains *e.g.* auto repair or do-it-yourself type of how-to guides and conduct future studies in an open-domain setting.

Throughout the thesis, we made thorough analyses with different reasoning and comprehension tasks on the proposed datasets and investigated the contribution of multimodality and grounding in different setups. An obvious future research direction would be exploring the contribution of multimodality in different downstream tasks such as navigation, entailment, word acquisition *etc.*

Visually grounded text processing has become an increasingly popular research field both computer vision and natural language processing communities have been studying. Nevertheless, in the existence of additional auxiliary modalities such as audio, object tags, or speech, joint modeling of such modalities yet remains an underexplored direction in the literature. Therefore building such models that can incorporate the contextual signals coming

from different modalities in a joint fashion, would be an interesting research direction to explore.

Furthermore, we explored how different modalities could improve models' generalization abilities through different setups and experiments. However, integrating different modalities yet remains an open research problem. An interesting research direction would be analyzing different fusion techniques, both early and late fusion strategies in the existence of additional modalities accompanying visual and textual data such as audio, *etc.*

An obvious extension of our work on compositional generalization is to design specific architectures for the joint modeling of multiple modalities to further improve the systematic generalization performance of neural models.

A fruitful research direction would be to develop a large-scale comprehensive multimodal benchmark dataset and evaluation methods to test models' compositional generalization across different comprehension and reasoning tasks.

Another interesting research direction would be exploring the embodiment of multimodal how-to instructions setup towards developing real-world applications such that in robotics. In particular, given the cooking recipe instructions, designing systems that can prepare the actual food by interacting with real-world objects and ingredients, which remains an open research problem and we hope our work could further stimulate research towards that direction.

Finally, the RecipeQA dataset introduces a challenging benchmark for understanding step-by-step how-to instructions. We hope our work would motivate further research towards understanding multimodal how-to instructions and facilitate research in this direction both in vision and language communities.

REFERENCES

- [1] David Marr and A Vision. A computational investigation into the human representation and processing of visual information. *WH San Francisco: Freeman and Company*, 1(2), **1982**.
- [2] Maggie Tallerman and Kathleen R Gibson. *The Oxford handbook of language evolution*. Oxford University Press, **2012**.
- [3] Stephen R. Anderson. How many languages are there in the world? <http://www.linguisticsociety.org/content/how-many-languages-are-there-world>, **2010**. Accessed: 2017-05-10.
- [4] Stephen R. Anderson. When nettles were dish of the day. <http://www.linguisticsociety.org/content/how-many-languages-are-there-world>, **2010**. Accessed: 2017-05-10.
- [5] Peter RAULWING and H Meyer. The kikkuli text. *Hittite Training Instructions for Chariot Horses in the Second Half of the 2nd Millennium BC and Their Interdisciplinary Context*, **2009**.
- [6] Cassie. Salad-in-a-jar 101: How to make mason jar salads + 4 fool-proof salad in a jar recipes. <https://wholefully.com/salad-in-a-jar-101/>, **2017**. Accessed: 2017-05-29.
- [7] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision. *CoRR*, abs/2006.13256, **2020**.

- [8] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, **2013**.
- [9] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, **2014**.
- [10] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433. **2015**.
- [11] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137. **2015**.
- [12] Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. Baby talk: Understanding and generating image descriptions. In *Proceedings of the 24th CVPR*. Citeseer, **2011**.
- [13] Rebecca Mason and Eugene Charniak. Nonparametric method for data-driven image captioning. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 592–598. **2014**.
- [14] Semih Yagcioglu, Erkut Erdem, Aykut Erdem, and Ruket Cakici. A distributed representation based query expansion approach for image captioning. In *ACL (2)*, pages 106–111. **2015**.
- [15] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Advances in Neural Information Processing Systems*, pages 1682–1690. **2014**.

- [16] Licheng Yu, Eunbyung Park, Alexander C Berg, and Tamara L Berg. Visual madlibs: Fill in the blank description generation and question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2461–2469. **2015**.
- [17] Lynette Hirschman, Marc Light, Eric Breck, and John D Burger. Deep read: A reading comprehension system. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 325–332. Association for Computational Linguistics, **1999**.
- [18] Lucy Vanderwende. Answering and questioning for machine reading. In *AAAI Spring Symposium: Machine Reading*, page 91. **2007**.
- [19] Ellen Riloff and Michael Thelen. A rule-based question answering system for reading comprehension tests. In *Proceedings of the 2000 ANLP/NAACL Workshop on Reading comprehension tests as evaluation for computer-based language understanding systems-Volume 6*, pages 13–19. Association for Computational Linguistics, **2000**.
- [20] Eugene Charniak, Yasemin Altun, Rodrigo de Salvo Braz, Benjamin Garrett, Margaret Kosmala, Tomer Moscovich, Lixin Pang, Changhee Pyo, Ye Sun, Wei Wy, et al. Reading comprehension programs in a statistical-language-processing class. In *Proceedings of the 2000 ANLP/NAACL Workshop on Reading comprehension tests as evaluation for computer-based language understanding systems-Volume 6*, pages 1–5. Association for Computational Linguistics, **2000**.
- [21] Hwee Tou Ng, Leong Hwee Teo, and Jennifer Lai Pheng Kwan. A machine learning approach to answering questions for reading comprehension tests. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 13*, pages 124–132. Association for Computational Linguistics, **2000**.

- [22] W Wang, J Auer, R Parasuraman, I Zubarev, D Brandyberry, and MP Harper. A question answering system developed as a project in a natural language processing course. In *Proceedings of the 2000 ANLP/NAACL Workshop on Reading comprehension tests as evaluation for computer-based language understanding systems-Volume 6*, pages 28–35. Association for Computational Linguistics, **2000**.
- [23] Mengqiu Wang, Noah A Smith, and Teruko Mitamura. What is the jeopardy model? a quasi-synchronous grammar for qa. In *EMNLP-CoNLL*, volume 7, pages 22–32. **2007**.
- [24] Matthew Richardson, Christopher JC Burges, and Erin Renshaw. Mctest: A challenge dataset for the open-domain machine comprehension of text. In *EMNLP*, volume 3, page 4. **2013**.
- [25] Yi Yang, Wen-tau Yih, and Christopher Meek. Wikiqa: A challenge dataset for open-domain question answering. In *EMNLP*, pages 2013–2018. **2015**.
- [26] Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*, **2015**.
- [27] Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. Newsqa: A machine comprehension dataset. *arXiv preprint arXiv:1611.09830*, **2016**.
- [28] Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1693–1701. **2015**.

- [29] Danqi Chen, Jason Bolton, and Christopher D Manning. A thorough examination of the cnn/daily mail reading comprehension task. *arXiv preprint arXiv:1606.02858*, **2016**.
- [30] Jason Weston, Antoine Bordes, Sumit Chopra, and Tomas Mikolov. Towards ai-complete question answering: A set of prerequisite toy tasks. *CoRR*, abs/1502.05698, **2015**.
- [31] Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. End-to-end memory networks. In *Advances in neural information processing systems*, pages 2440–2448. **2015**.
- [32] Michael Heilman and Noah A Smith. Tree edit models for recognizing textual entailments, paraphrases, and answers to questions. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1011–1019. Association for Computational Linguistics, **2010**.
- [33] Xuchen Yao, Benjamin Van Durme, Chris Callison-Burch, and Peter Clark. Answer extraction as sequence tagging with tree edit distance. In *HLT-NAACL*, pages 858–867. Citeseer, **2013**.
- [34] Aliaksei Severyn and Alessandro Moschitti. Automatic feature engineering for answer selection and extraction. In *EMNLP*, volume 13, pages 458–467. **2013**.
- [35] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*, **2016**.
- [36] Mohit Iyyer, Varun Manjunatha, Anupam Guha, Yogarshi Vyas, Jordan Boyd-Graber, Hal Daumé III, and Larry Davis. The amazing mysteries of the gutter: Drawing inferences between panels in comic book narratives. *arXiv preprint arXiv:1611.05118*, **2016**.

- [37] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4995–5004. **2016**.
- [38] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1–9. **2015**.
- [39] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, **2002**.
- [40] Michael Denkowski and Alon Lavie. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In *Proc. of EACL Workshop on Statistical Machine Translation*. **2014**.
- [41] Zhibiao Wu and Martha Palmer. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138. Association for Computational Linguistics, **1994**.
- [42] Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. Are you talking to a machine? dataset and methods for multilingual image question. In *Advances in Neural Information Processing Systems*, pages 2296–2304. **2015**.
- [43] Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. The goldilocks principle: Reading children’s books with explicit memory representations. *arXiv preprint arXiv:1511.02301*, **2015**.
- [44] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, **2016**.

- [45] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgb-d images. *Computer Vision–ECCV 2012*, pages 746–760, **2012**.
- [46] Stanislaw Antol, C Lawrence Zitnick, and Devi Parikh. Zero-shot learning via visual abstraction. In *European Conference on Computer Vision*, pages 401–416. Springer, **2014**.
- [47] C Lawrence Zitnick and Devi Parikh. Bringing semantics into focus using visual abstraction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3009–3016. **2013**.
- [48] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4631–4640. **2016**.
- [49] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. *arXiv preprint arXiv:1612.06890*, **2016**.
- [50] Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. **2017**.
- [51] Brenden M. Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 2879–2888. PMLR, **2018**.

- [52] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, **2015**.
- [53] Jerry A Fodor and Zenon W Pylyshyn. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71, **1988**.
- [54] Gary F Marcus. Rethinking eliminative connectionism. *Cognitive psychology*, 37(3):243–282, **1998**.
- [55] Jerry A Fodor and Ernest Lepore. *The compositionality papers*. Oxford University Press, **2002**.
- [56] Paco Calvo and John Symons. *The architecture of cognition: Rethinking Fodor and Pylyshyn’s systematicity challenge*. MIT Press, **2014**.
- [57] Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. Compositionality decomposed: how do neural networks generalise? *Journal of Artificial Intelligence Research*, 67:757–795, **2020**.
- [58] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Learning to compose neural networks for question answering. *arXiv preprint arXiv:1601.01705*, **2016**.
- [59] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1988–1997. IEEE Computer Society, **2017**. doi:10.1109/CVPR.2017.215.
- [60] Laura Ruis, Jacob Andreas, Marco Baroni, Diane Bouchacourt, and Brenden M. Lake. A benchmark for systematic generalization in grounded language understanding. In Hugo Larochelle, Marc’Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural*

Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual. **2020.**

- [61] Gottlob Frege. Compound thoughts. *Mind*, 72(285):1–17, **1963**.
- [62] Barbara Partee et al. Lexical semantics and compositionality. *An invitation to cognitive science: Language*, 1:311–360, **1995**.
- [63] Paul Smolensky and Géraldine Legendre. *The harmonic mind: From neural computation to optimality-theoretic grammar (Cognitive architecture), Vol. 1*. MIT press, **2006**.
- [64] M Christiansen, Nick Chater, et al. Generalization and connectionist language learning. *Mind and Language*, 9(3), **1994**.
- [65] Franklin Chang. Symbolically speaking: A connectionist model of sentence production. *Cognitive science*, 26(5):609–651, **2002**.
- [66] Frank van der Velde, Gwendid T van der Voort van der Kleij, and Marc de Kamps. Lack of combinatorial productivity in language processing with simple recurrent networks. *Connection Science*, 16(1):21–46, **2004**.
- [67] Matthew M Botvinick and David C Plaut. Short-term memory for serial order: a recurrent neural network model. *Psychological review*, 113(2):201, **2006**.
- [68] Francis CK Wong and William SY Wang. Generalisation towards combinatorial productivity in language acquisition by simple recurrent networks. In *2007 International Conference on Integration of Knowledge Intensive Multi-Agent Systems*, pages 139–144. IEEE, **2007**.
- [69] Philémon Brakel and Stefan Frank. Strong systematicity in sentence processing by simple recurrent networks. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 31. **2009**.

- [70] Stefan L Frank, P Calvo, and J Symons. Getting real about systematicity. *The architecture of cognition: Rethinking Fodor and Pylyshyn’s systematicity challenge*, pages 147–164, **2014**.
- [71] Samuel R Bowman, Christopher D Manning, and Christopher Potts. Tree-structured composition in neural networks without tree-structured architectures. *arXiv preprint arXiv:1506.04834*, **2015**.
- [72] Dzmitry Bahdanau, Shikhar Murty, Michael Noukhovitch, Thien Huu Nguyen, Harm de Vries, and Aaron C. Courville. Systematic generalization: What is required and can it be learned? In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, **2019**.
- [73] Brenden M. Lake, Tal Linzen, and Marco Baroni. Human few-shot learning of compositional instructions. In Ashok K. Goel, Colleen M. Seifert, and Christian Freksa, editors, *Proceedings of the 41th Annual Meeting of the Cognitive Science Society, CogSci 2019: Creativity + Cognition + Computation, Montreal, Canada, July 24-27, 2019*, pages 611–617. cognitivesciencesociety.org, **2019**.
- [74] Brenden M. Lake. Compositional generalization through meta sequence-to-sequence learning. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d’Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 9788–9798. **2019**.
- [75] Mitja Nikolaus, Mostafa Abdou, Matthew Lamm, Rahul Aralikatte, and Desmond Elliott. Compositional generalization in image captioning. In Mohit Bansal and Aline Villavicencio, editors, *Proceedings of the 23rd Conference on Computational Natural Language Learning, CoNLL 2019, Hong Kong, China*,

- November 3-4, 2019, pages 87–98. Association for Computational Linguistics, **2019**. doi:10.18653/v1/K19-1009.
- [76] Marco Baroni. Linguistic generalization and compositionality in modern artificial neural networks. *Phil. Trans. R. Soc. B*, 375(1791):20190307, **2020**.
- [77] Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, et al. Measuring compositional generalization: A comprehensive method on realistic data. *arXiv preprint arXiv:1912.09713*, **2019**.
- [78] Dídac Surís, Dave Epstein, Heng Ji, Shih-Fu Chang, and Carl Vondrick. Learning to learn words from visual scenes. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXIX*, volume 12374 of *Lecture Notes in Computer Science*, pages 434–452. Springer, **2020**. doi:10.1007/978-3-030-58526-6_26.
- [79] Jasmijn Bastings, Marco Baroni, Jason Weston, Kyunghyun Cho, and Douwe Kiela. Jump to better conclusions: SCAN both left and right. In Tal Linzen, Grzegorz Chrupala, and Afra Alishahi, editors, *Proceedings of the Workshop: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP 2018, Brussels, Belgium, November 1, 2018*, pages 47–55. Association for Computational Linguistics, **2018**. doi:10.18653/v1/w18-5407.
- [80] Yafu Li, Yongjing Yin, Yulong Chen, and Yue Zhang. On compositional generalization of neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4767–4780. Association for Computational Linguistics, Online, **2021**. doi:10.18653/v1/2021.acl-long.368.

- [81] Verna Dankers, Elia Bruni, and Dieuwke Hupkes. The paradox of the compositionality of natural language: A neural machine translation case study. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4154–4175. Association for Computational Linguistics, Dublin, Ireland, **2022**. doi:10.18653/v1/2022.acl-long.286.
- [82] Adam Liška, Germán Kruszewski, and Marco Baroni. Memorize or generalize? searching for a compositional rnn in a haystack. *arXiv preprint arXiv:1802.06467*, **2018**.
- [83] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. **2015**.
- [84] Yuval Atzmon, Jonathan Berant, Vahid Kezami, Amir Globerson, and Gal Chechik. Learning to generalize to new compositions in image understanding. *arXiv preprint arXiv:1608.07639*, **2016**.
- [85] Emanuele Bugliarello and Desmond Elliott. The role of syntactic planning in compositional image captioning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 593–607. Association for Computational Linguistics, Online, **2021**. doi:10.18653/v1/2021.eacl-main.48.
- [86] George Pantazopoulos, Alessandro Suglia, and Arash Eshghi. Combine to describe: Evaluating compositional generalization in image captioning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 115–131. Association for Computational Linguistics, Dublin, Ireland, **2022**. doi:10.18653/v1/2022.acl-srw.11.

- [87] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, **2014**.
- [88] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, **2015**.
- [89] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *ArXiv*, abs/1504.00325, **2015**.
- [90] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086. **2018**.
- [91] Harm de Vries, Dzmitry Bahdanau, Shikhar Murty, Aaron C. Courville, and Philippe Beaudoin. CLOSURE: assessing systematic generalization of CLEVR models. In *Visually Grounded Interaction and Language (ViGIL), NeurIPS 2019 Workshop, Vancouver, Canada, December 13, 2019*. **2019**.
- [92] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don’t just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4971–4980. **2018**.
- [93] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, **2018**.

- [94] David Saxton, Edward Grefenstette, Felix Hill, and Pushmeet Kohli. Analysing mathematical reasoning abilities of neural models. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, **2019**.
- [95] Zhengxuan Wu, Elisa Kreiss, Desmond C. Ong, and Christopher Potts. ReaSCAN: Compositional reasoning in language grounding. *NeurIPS 2021 Datasets and Benchmarks Track*, **2021**.
- [96] Najoung Kim and Tal Linzen. COGS: A compositional generalization challenge based on semantic interpretation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9087–9105. Association for Computational Linguistics, Online, **2020**. doi:10.18653/v1/2020.emnlp-main.731.
- [97] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *European Conference on Computer Vision (ECCV)*. **2018**.
- [98] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, , Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)*, **2021**.
- [99] Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, **1990**.
- [100] Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751. Association for Computational Linguistics, Doha, Qatar, **2014**. doi:10.3115/v1/D14-1181.

- [101] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. Convolutional sequence to sequence learning. In *International Conference on Machine Learning*, pages 1243–1252. PMLR, **2017**.
- [102] Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. Learning to reason: End-to-end module networks for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 804–813. **2017**.
- [103] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008. **2017**.
- [104] Santiago Ontanon, Joshua Ainslie, Zachary Fisher, and Vaclav Cvicek. Making transformers solve compositional tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3591–3607. **2022**.
- [105] Arkil Patel, Satwik Bhattamishra, Phil Blunsom, and Navin Goyal. Revisiting the compositional generalization abilities of neural sequence models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 424–434. **2022**.
- [106] Ankur Sikarwar, Arkil Patel, and Navin Goyal. When can transformers ground and compose: Insights from compositional generalization benchmarks. *arXiv preprint arXiv:2210.12786*, **2022**.
- [107] Róbert Csordás, Kazuki Irie, and Juergen Schmidhuber. The devil is in the detail: Simple tricks improve systematic generalization of transformers. In

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 619–634. **2021**.

- [108] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, **2014**.
- [109] Juyong Kim, Pradeep Ravikumar, Joshua Ainslie, and Santiago Ontanon. Improving compositional generalization in classification tasks via structure annotations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 637–645. **2021**.
- [110] Jacob Andreas. Good-enough compositional data augmentation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7556–7566. **2020**.
- [111] Demi Guo, Yoon Kim, and Alexander M Rush. Sequence-level mixed sample data augmentation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5547–5552. **2020**.
- [112] Ekin Akyürek, Afra Feyza Akyürek, and Jacob Andreas. Learning to recombine and resample data for compositional generalization. In *International Conference on Learning Representations*. **2020**.
- [113] Andrew Drozdov, Nathanael Schärli, Ekin Akyürek, Nathan Scales, Xinying Song, Xinyun Chen, Olivier Bousquet, and Denny Zhou. Compositional semantic parsing with large language models. *arXiv preprint arXiv:2209.15003*, **2022**.
- [114] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop*

BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 353–355. Association for Computational Linguistics, Brussels, Belgium, **2018**. doi:10.18653/v1/W18-5446.

- [115] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., **2019**.
- [116] Sebastian Gehrmann, Tosin P. Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Aremu Anuoluwapo, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh D. Dhole, Wanyu Du, Esin Durmus, Ondrej Dusek, Chris Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Rubungo Andre Niyongabo, Salomey Osei, Ankur P. Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. The GEM benchmark: Natural language generation, its evaluation and metrics. *CoRR*, abs/2102.01672, **2021**.
- [117] Lynette Hirschman and Robert Gaizauskas. Natural language question answering: The view from here. *Natural Language Engineering*, 7(4):275–300, **2001**.

- [118] Christopher JC Burges. Towards the machine comprehension of text: An essay. Technical report, Technical report, Microsoft Research Technical Report MSR-TR-2013-125, **2013**.
- [119] Saku Sugawara, Hikaru Yokono, and Akiko Aizawa. Prerequisite skills for reading comprehension: Multi-perspective analysis of mctest datasets and systems. In *Association for the Advancement of Artificial Intelligence (AAAI)*, pages 3089–3096. **2017**.
- [120] Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, **2018**.
- [121] Daniel Hewlett, Alexandre Lacoste, Llion Jones, Illia Polosukhin, Andrew Fandrianto, Jay Han, Matthew Kelcey, and David Berthelot. WikiReading: A novel large-scale language understanding task over Wikipedia. In *Association for Computational Linguistics (ACL)*. **2016**.
- [122] Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Association for Computational Linguistics (ACL)*, pages 1601–1611. Association for Computational Linguistics, Vancouver, Canada, **2017**.
- [123] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. **2017**.
- [124] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European Conference on Computer Vision*, pages 235–251. Springer, **2016**.

- [125] Samira Ebrahimi Kahou, Adam Atkinson, Vincent Michalski, Akos Kadar, Adam Trischler, and Yoshua Bengio. FigureQA: An annotated figure dataset for visual reasoning. In *International Conference on Learning Representations (ICLR) Workshops*. **2018**.
- [126] Peter Norvig. A unified theory of inference for text understanding. Technical report, University of California at Berkeley, Berkeley, CA, USA, **1987**.
- [127] Léon Bottou. From machine learning to machine reasoning - an essay. *Machine Learning*, 94(2):133–149, **2014**.
- [128] Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis. RecipeQA: A Challenge Dataset for Multimodal Comprehension of Cooking Recipes. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. **2018**.
- [129] Marco Lui and Timothy Baldwin. `langid.py`: An off-the-shelf language identification tool. In *Association for Computational Linguistics (ACL) Demo Session*, pages 25–30. **2012**.
- [130] Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. From word embeddings to document distances. In *International Conference on Machine Learning (ICML)*, pages 957–966. **2015**.
- [131] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. **2016**.
- [132] Harsh Agrawal, Arjun Chandrasekaran, Dhruv Batra, Devi Parikh, and Mohit Bansal. Sort story: Sorting jumbled images and captions into stories. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 925–931. **2016**.

- [133] Amaia Salvador, Nicholas Hynes, Yusuf Aytar, Javier Marin, Ferda Ofli, Ingmar Weber, and Antonio Torralba. Learning cross-modal embeddings for cooking recipes and food images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. **2017**.
- [134] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International Conference on Machine Learning (ICML)*, pages 1188–1196. **2014**.
- [135] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet: Large scale visual recognition challenge. *International Journal of Computer Vision*, 113:211–252, **2015**.
- [136] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, pages 2493—2537, **2011**.
- [137] Jonathan Malmaud, Earl Wagner, Nancy Chang, and Kevin Murphy. Cooking with semantics. In *ACL 2014 Workshop on Semantic Parsing*, pages 33–38. **2014**.
- [138] Jermisak Jermisurawong and Nizar Habash. Predicting the structure of cooking recipes. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 781–786. **2015**.
- [139] Jonathan Malmaud, Jonathan Huang, Vivek Rathod, Nick Johnston, Andrew Rabinovich, and Kevin Murphy. What’s cookin’? interpreting cooking videos using text, speech and vision. *arXiv preprint arXiv:1503.01558*, **2015**.
- [140] Ozan Sener, Amir Zamir, Silvio Savarese, and Ashutosh Saxena. Unsupervised semantic parsing of video collections. In *IEEE International Conference on Computer Vision (ICCV)*. **2015**.

- [141] Chloé Kiddon, Luke Zettlemoyer, and Yejin Choi. Globally coherent text generation with neural checklist models. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 329—339. **2016**.
- [142] Antoine Bosselut, Corin Ennis, Omer Levy, Ari Holtzman, Dieter Fox, and Yejin Choi. Simulating action dynamics with neural process networks. In *International Conference on Learning Representations (ICLR)*. **2018**.
- [143] Antoine Bosselut, Corin Ennis, Omer Levy, Ari Holtzman, Dieter Fox, and Yejin Choi. Simulating Action Dynamics with Neural Process Networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*. **2018**.
- [144] Bhavana Dalvi, Lifu Huang, Niket Tandon, Wen-tau Yih, and Peter Clark. Tracking state changes in procedural text: a challenge dataset and models for process paragraph comprehension. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. **2018**.
- [145] Niket Tandon, Bhavana Dalvi, Joel Grus, Wen-tau Yih, Antoine Bosselut, and Peter Clark. Reasoning about actions and state changes by injecting commonsense knowledge. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. **2018**.
- [146] Xinya Du, Bhavana Dalvi Mishra, Niket Tandon, Antoine Bosselut, Wen-tau Yih, Peter Clark, and Claire Cardie. Be consistent! improving procedural text comprehension using label consistency. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. **2019**.
- [147] Mikael Henaff, Jason Weston, Arthur Szlam, Antoine Bordes, and Yann LeCun. Tracking The World State with Recurrent Entity Networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*. **2017**.

- [148] Trapit Bansal, Arvind Neelakantan, and Andrew McCallum. RelNet: End-to-End Modeling of Entities & Relations. In *NeurIPS Workshop on Automated Knowledge Base Construction (AKBC)*. **2017**.
- [149] Juan Pavez, Hector Allende, and Hector Allende-Cid. Working memory networks: Augmenting memory networks with a relational reasoning module. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1000–1009. **2018**.
- [150] Julien Perez and Fei Liu. Dialog state tracking, a machine reading approach using memory network. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 305–314. **2017**.
- [151] Hai Wang, Takeshi Onishi, Kevin Gimpel, and David McAllester. Emergent predication structure in hidden state vectors of neural readers. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 26–36. Association for Computational Linguistics, Vancouver, Canada, **2017**.
- [152] Bhuwan Dhingra, Qiao Jin, Zhilin Yang, William W Cohen, and Ruslan Salakhutdinov. Neural Models for Reasoning over Multiple Mentions using Coreference. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. **2018**.
- [153] Yangfeng Ji, Chenhao Tan, Sebastian Martschat, Yejin Choi, and Noah A Smith. Dynamic Entity Representations in Neural Language Models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. **2017**.
- [154] Rajarshi Das, Tsendsuren Munkhdalai, Xingdi Yuan, Adam Trischler, and Andrew McCallum. Building Dynamic Knowledge Graphs from Text using

- Machine Reading Comprehension. In *Proceedings of the International Conference on Learning Representations (ICLR)*. **2019**.
- [155] Adam Santoro, Ryan Faulkner, David Raposo, Jack Rae, Mike Chrzanowski, Theophane Weber, Daan Wierstra, Oriol Vinyals, Razvan Pascanu, and Timothy Lillicrap. Relational Recurrent Neural Networks. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 7299–7310. **2018**.
- [156] Mustafa Sercan Amac, Semih Yagcioglu, Aykut Erdem, and Erkut Erdem. Procedural reasoning networks for understanding multimodal procedures. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 441–451. **2019**.
- [157] Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6. Association for Computational Linguistics, Melbourne, Australia, **2018**.
- [158] Yoon Kim. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. **2014**.
- [159] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global Vectors for Word Representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. **2014**.
- [160] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the Conference of the North American Chapter of the*

Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), pages 2227–2237. **2018**.

- [161] R. K. Srivastava, K. Greff, and J. Schmidhuber. Highway networks. In *Proceedings of the International Conference on Machine Learning (ICML)*. **2015**.
- [162] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Rerning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778. **2016**.
- [163] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255. **2009**.
- [164] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 5998–6008. **2017**.
- [165] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. Bidirectional Attention Flow for Machine Comprehension. In *Proceedings of the International Conference on Learning Representations (ICLR)*. **2017**.
- [166] Javier Marin, Aritro Biswas, Ferda Ofli, Nicholas Hynes, Amaia Salvador, Yusuf Aytar, Ingmar Weber, and Antonio Torralba. Recipe1M: A Dataset for Learning Cross-Modal Embeddings for Cooking Recipes and Food Images. *arXiv preprint arXiv:1810.06553*, **2018**.
- [167] Yummly. Kaggle What’s Cooking? <https://www.kaggle.com/c/whats-cooking/data>, **2015**. [Accessed: 2018-05-31].

- [168] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic Differentiation in pytorch. In *NIPS-W*. **2017**.
- [169] Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*, **2018**.
- [170] Malihe Alikhani, Sreyasi Nag Chowdhury, Gerard de Melo, and Matthew Stone. Cite: A corpus of image-text discourse relations. *arXiv preprint arXiv:1904.06286*, **2019**.
- [171] Jack Hessel, Lillian Lee, and David Mimno. Unsupervised discovery of multimodal links in multi-image, multi-sentence documents. *arXiv preprint arXiv:1904.07826*, **2019**.
- [172] Malihe Alikhani and Matthew Stone. “caption” as a coherence relation: Evidence and implications. In *Proceedings of the Second Workshop on Shortcomings in Vision and Language*, pages 58–67. **2019**.
- [173] Ao Liu, Lizhen Qu, Junyu Lu, Chenbin Zhang, and Zenglin Xu. Machine reading comprehension: Matching and orders. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 2057–2060. **2019**.
- [174] Ao Liu, Shuai Yuan, Chenbin Zhang, Congjian Luo, Yaqing Liao, Kun Bai, and Zenglin Xu. Multi-level multimodal transformer network for multimodal recipe comprehension. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 1781–1784. **2020**.
- [175] Vishwash Batra, Aparajita Haldar, Yulan He, Hakan Ferhatosmanoglu, George Vogiatzis, and Tanaya Guha. Variational recurrent sequence-to-sequence retrieval for stepwise illustration. In *European Conference on Information Retrieval*, pages 50–64. Springer, **2020**.
- [176] Hossein Rajaby Faghihi, Roshanak Mirzaee, Sudarshan Paliwal, and Parisa Kordjamshidi. Latent alignment of procedural concepts in multimodal recipes. *arXiv preprint arXiv:2101.04727*, **2021**.

- [177] Kenta Hama, Takashi Matsubara, Kuniaki Uehara, and Jianfei Cai. Exploring uncertainty measures for image-caption embedding-and-retrieval task. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 17(2):1–19, **2021**.
- [178] Pritish Sahu, Karan Sikka, and Ajay Divakaran. Towards solving multimodal comprehension. *arXiv preprint arXiv:2104.10139*, **2021**.
- [179] Yunjie Wu, Sai Zhang, Xiaowang Zhang, Zhiyong Feng, and Liang Wan. A knowledge-based deep heterogeneous graph matching model for multimodal recipeqa. In *ISWC (Posters/Demos/Industry)*. **2021**.
- [180] Malihe Alikhani, Fangda Han, Hareesh Ravi, Mubbasir Kapadia, Vladimir Pavlovic, and Matthew Stone. Cross-modal coherence for text-to-image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10427–10435. **2022**.
- [181] Pritish Sahu, Karan Sikka, and Ajay Divakaran. Challenges in procedural multimodal machine comprehension: A novel way to benchmark. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3654–3663. **2022**.
- [182] Vishal Pallagani, Priyadharsini Ramamurthy, Vedant Khandelwal, Revathy Venkataramanan, Kausik Lakkaraju, Sathyanarayanan N Aakur, and Biplav Srivastava. A rich recipe representation as plan to support expressive multi modal queries on recipe content and preparation process. *arXiv preprint arXiv:2203.17109*, **2022**.
- [183] Te-Lin Wu, Alex Spangher, Pegah Alipoormolabashi, Marjorie Freedman, Ralph Weischedel, and Nanyun Peng. Understanding multimodal procedural knowledge by sequencing multimodal instructional manuals. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4525–4542. **2022**.
- [184] Huibin Zhang, Zhengkun Zhang, Yao Zhang, Jun Wang, Yufan Li, Zhenglu Yang, et al. Modeling temporal-modal entity graph for procedural multimodal machine comprehension. *arXiv preprint arXiv:2204.02566*, **2022**.

- [185] Irina Higgins, Nicolas Sonnerat, Loic Matthey, Arka Pal, Christopher P. Burgess, Matko Bosnjak, Murray Shanahan, Matthew Botvinick, Demis Hassabis, and Alexander Lerchner. SCAN: learning hierarchical compositional visual concepts. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, **2018**.
- [186] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40, **2017**.
- [187] João Loula, Marco Baroni, and Brenden M. Lake. Rearranging the familiar: Testing compositional generalization in recurrent networks. In Tal Linzen, Grzegorz Chrupala, and Afra Alishahi, editors, *Proceedings of the Workshop: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP 2018, Brussels, Belgium, November 1, 2018*, pages 108–114. Association for Computational Linguistics, **2018**. doi:10.18653/v1/w18-5413.
- [188] Jake Russin, Jason Jo, Randall C. O’Reilly, and Yoshua Bengio. Compositional generalization in a deep seq2seq model by separating syntax and semantics. *CoRR*, abs/1904.09708, **2019**.
- [189] Ankit Vani, Max Schwarzer, Yuchen Lu, Eeshan Dhekane, and Aaron Courville. Iterated learning for emergent systematicity in VQA. In *International Conference on Learning Representations*. **2021**.
- [190] Moshe Bar. The proactive brain: using analogies and associations to generate predictions. *Trends in cognitive sciences*, 11(7):280–289, **2007**.
- [191] Andy Clark. *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford University Press, **2015**.
- [192] Harshala Gammulle, Simon Denman, Sridha Sridharan, and Clinton Fookes. Predicting the future: A jointly learnt model for action anticipation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5562–5571. **2019**.

- [193] Qiuhong Ke, Mario Fritz, and Bernt Schiele. Time-conditioned action anticipation in one shot. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9925–9934. **2019**.
- [194] JK Chung, PL Kannappan, CT Ng, and PK Sahoo. Measures of distance between probability distributions. *Journal of mathematical analysis and applications*, 138(1):280–292, **1989**.
- [195] Jesse Thomason, Daniel Gordon, and Yonatan Bisk. Shifting the baseline: Single modality performance on visual navigation & QA. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1977–1983. Association for Computational Linguistics, Minneapolis, Minnesota, **2019**. doi:10.18653/v1/N19-1197.
- [196] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, **1997**.
- [197] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, **2014**.
- [198] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Florence, Italy, **2019**.
- [199] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6):1137–1149, **2017**. doi:10.1109/TPAMI.2016.2577031.
- [200] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE, **2020**.

- [201] Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. Merlot reserve: Neural script knowledge through vision and language and sound. *arXiv preprint arXiv:2201.02639*, **2022**.
- [202] Ishita Dasgupta, Demi Guo, Andreas Stuhlmüller, Samuel J. Gershman, and Noah D. Goodman. Evaluating compositionality in sentence embeddings. *CoRR*, abs/1802.04302, **2018**.
- [203] Allyson Ettinger, Ahmed Elgohary, Colin Phillips, and Philip Resnik. Assessing composition in sentence vector representations. In Emily M. Bender, Leon Derczynski, and Pierre Isabelle, editors, *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 1790–1801. Association for Computational Linguistics, **2018**.
- [204] Jacob Andreas. Measuring compositionality in representation learning. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, **2019**.
- [205] Maxwell Nye, Armando Solar-Lezama, Joshua Tenenbaum, and Brenden Lake. Learning compositional rules via neural program synthesis. *Proceedings of the NeurIPS*, **2019**.
- [206] Felix Hill, Andrew K. Lampinen, Rosalia Schneider, Stephen Clark, Matthew Botvinick, James L. McClelland, and Adam Santoro. Emergent systematic generalization in a situated agent. *CoRR*, abs/1910.00571, **2019**.
- [207] Bo Wu, Haoyu Qin, Alireza Zareian, Carl Vondrick, and Shih-Fu Chang. Analogical reasoning for visually grounded language acquisition. *CoRR*, abs/2007.11668, **2020**.
- [208] Xisen Jin, Junyi Du, and Xiang Ren. Visually grounded continual learning of compositional semantics. *CoRR*, abs/2005.00785, **2020**.
- [209] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In

Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3674–3683. **2018**.

- [210] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. ALFRED: A benchmark for interpreting grounded instructions for everyday tasks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10737–10746. IEEE, **2020**. doi:10.1109/CVPR42600.2020.01075.
- [211] Paul Hongsuck Seo, Arsha Nagrani, and Cordelia Schmid. Look before you speak: Visually contextualized utterances. *CoRR*, abs/2012.05710, **2020**.

CURRICULUM VITAE

Credentials

Name, Surname : Semih Yağcıoğlu
Place of Birth : Nazilli / TURKEY
Marital Status : Married
E-mail : semihyagcioglu@gmail.com
Address : Department of Computer Engineering, Hacettepe University
Ankara / TURKEY

Education

M.Sc. : Engineering Management, Middle East Technical University
B.Sc. : Computer Engineering, Eskişehir Osmangazi University

Foreign Languages : English

Work Experience : Director of AI, Apziva
Group Lead, STM
Software Engineer, MITS

Areas of Experiences : Natural Language Processing, Computer Vision

Publications

- A Survey on Compositional Generalization, S. Yagcioglu, O. B. Ince, A. Erdem, E. Erdem. (In submission)
- Grounded Compositional Understanding of Actions, S. Yagcioglu, O. B. Ince, S. Seker, D. Elliot, E. Erdem, A. Erdem, D. Yuret. (Submitted, Under Review)
- Neural Natural Language Generation: A Survey on Multilinguality, Multimodality, Controllability and Learning, E. Erdem, M. Kuyu, S. Yagcioglu, A. Frank, L. Parcalabescu, B. Plank, A. Babii, O. Turuta, A. Erdem, I. Calixto, E. Lloret, E. Apostol, C. Truică, B. Šandrih, S. Martinčić-Ipšić, G. Berend, A. Gatt, G. Korvel, JAIR 2022
- Procedural Reasoning Networks for Understanding Multimodal Procedures, M. S. Amac, S. Yagcioglu, A. Erdem, E. Erdem. CONLL 2019, Hong Kong
- Detecting Cybersecurity Events from Noisy Short-Text, S. Yagcioglu, M. S. Seyfioglu, B. Citamak, B. Bardak, S. Ozbek, A. Yuksel, E. I. Tatli. NAACL 2019, Minnesota, USA
- RecipeQA: A Dataset for Multimodal Comprehension of Cooking Recipes, S. Yagcioglu, A. Erdem, E. Erdem, N. Ikizler-Cinbis. EMNLP 2018, Brussels, Belgium, Oct. 2018 (Covered by the MIT Tech Review)
- TasvirEt: A Benchmark Dataset for Automatic Turkish Description Generation from Images, M. E. Unal, B. Citamak, S. Yagcioglu, A. Erdem, E. Erdem, N. Ikizler-Cinbis, R. Cakici. SIU 2016, Zonguldak, Turkey, May 2016. (Best Paper Award) also featured on national TV and press.
- A Distributed Representation Based Query Expansion Approach for Image Captioning, S. Yagcioglu, E. Erdem, A. Erdem, R. Cakici. ACL 2015, Beijing, China, July 2015.
- City Scale Image Geolocalization via Dense Scene Alignment, S. Yagcioglu, E. Erdem and A. Erdem. WACV 2015, Hawaii, USA, Jan 2015.