Learning About the World from Language
and
Learning Language by Observing the World

Aida Nematzadeh

DeepMind

Learning About the World from Language

I am going to make pancakes for you. I have all the ingredients and now need to mix them together.

When does mixing begin and end?

Learning Language by Observing the World

Je casse les oeufs.

I need to mix the eggs with the flour.

Can we leverage this interaction between vision and language?

Obtaining real-world multimodal datasets of natural language and human activities is really hard.

YouTube Instructional Videos

Task: **Make Pancakes**

Video:

Narration: *Making pancakes is easy… first, you'll need to mix sifted flour, sugar …*

Why Are Instructional Videos Interesting?

Some correspondence between language & video.

Naturalistic language and videos.

- Talking about things irrelevant to the task.
- Filmed in a variety of settings (not in a lab).

The State of Instructional Videos

2014

Instructional Videos for Unsupervised Harvesting and Learning of Action Examples

Shoou-I Yu   Lu Jiang   Alexander Hauptmann

...

2018

Towards Automatic Learning of Procedures from Web Instructional Videos

Luowei Zhou, Robotics Institute   Chenliang Xu, Department of CS   Jason J. Corso, Department of EECS

2019

Cross-task weakly supervised learning from instructional videos

Dimitri Zhukov[1,2]   Jean-Baptiste Alayrac[1,3]   Ramazan Gokberk Cinbis[4]   David Fouhey[5]
Ivan Laptev[1,2]   Josef Sivic[1,2,6]

2019

HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips

Antoine Miech[1,2*]   Dimitri Zhukov[1,2*]   Jean-Baptiste Alayrac[2+]
Makarand Tapaswi[2]   Ivan Laptev[1,2]   Josef Sivic[1,2,3]

Many more domains (not just cooking).

Much larger datasets.

Multiple languages.

to Segment Actions

Learning ~~About the World~~ from Language

Why Segmenting Actions?

Crucial to understanding the world, remembering things, and planning.

Actions: background | pour batter | background | remove pancake

Video:

Narration: *hey folks here welcome to my kitchen ... pour a nice-sized amount ... change the angle to show ... and take it out*

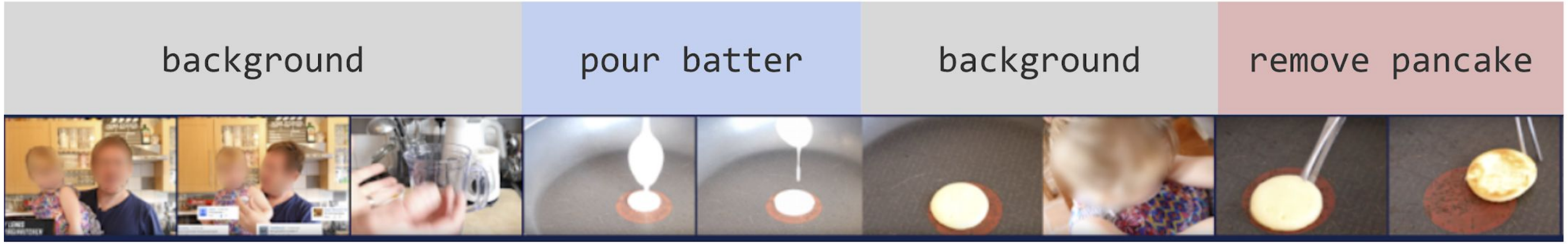
The CrossTask Dataset [Zhukov *et al.*, 2019]

Task: **Make Pancakes**

Steps: {add flour, add egg, whisk mixture, pour mixture ...}

Supervision:
task ordering
or set of steps

Video:



Narration: *hey folks here welcome to my kitchen . . . pour a nice-sized amount . . . change the angle to show . . . and take it out*

~70 percent of the videos is background regions.

supervision

to Segment Actions

Learning ~~About the World~~ from Language

Do **task orderings** & **narrations** help unsupervised action segmentation?

Training Without Segment Labels

`make pancakes:` `{add egg, add flour, …, pour batter, remove pancake}`

Actions a:

background	pour batter

Video features x:

hey folks here welcome to my kitchen … pour a nice-sized amount…

Generative: $$\max_{\theta} \sum_{a} p_\theta(a, x)$$ [Richard et al. 2018, Sener and Yao 2018]

Discriminative: $$\max_{\theta, a} p_\theta(a|x)$$ [Alayrac et al. 2016, Zhukov et al. 2019]

Weak-supervision for a:

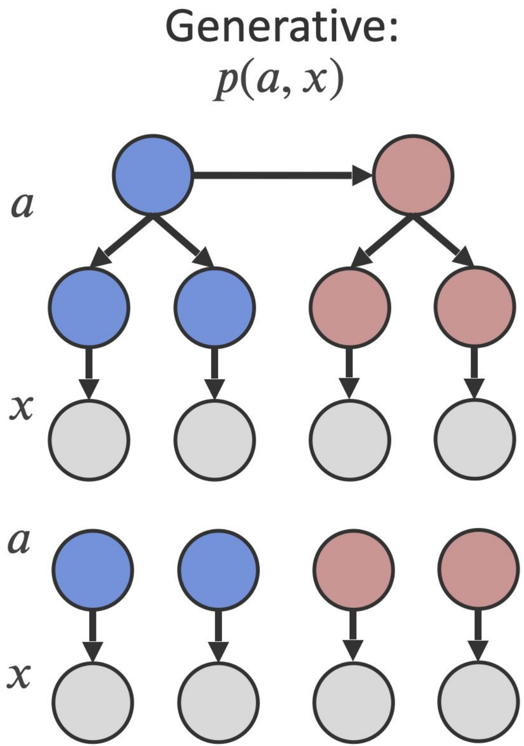
- Likely ordering of the actions
- Time-aligned narration

Structured:

Semi-Markov model

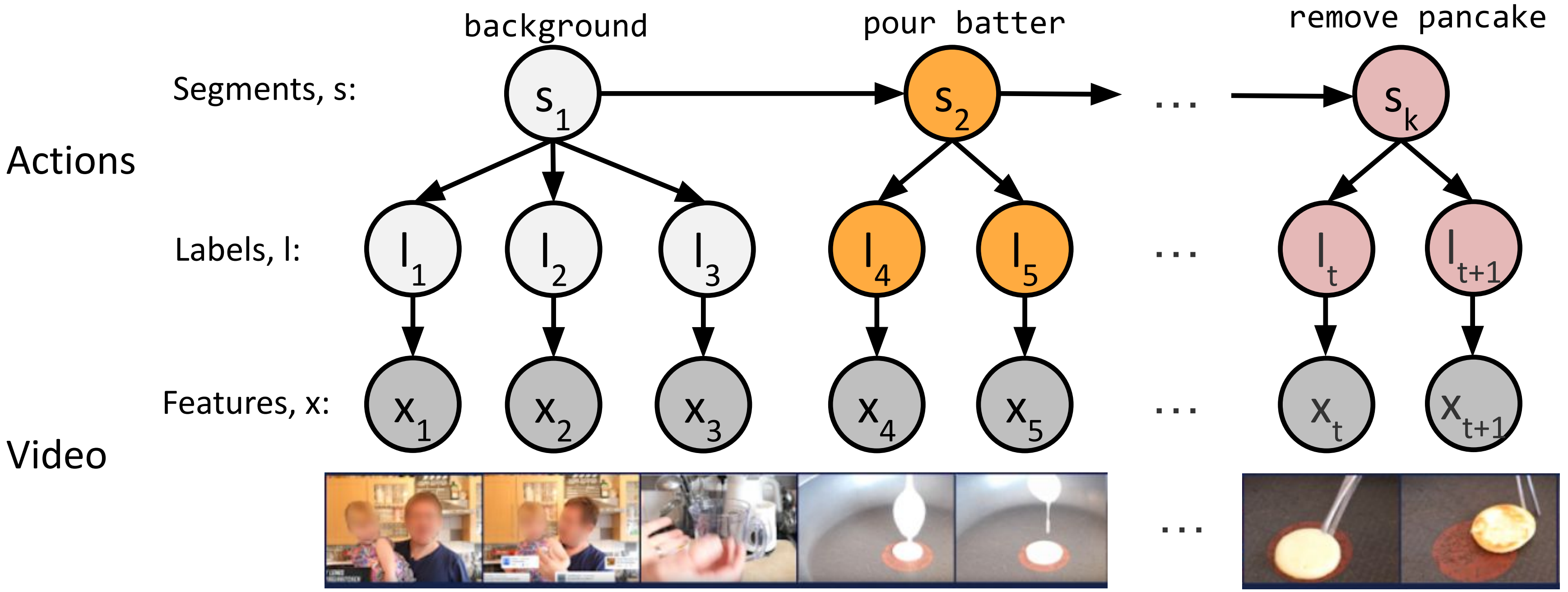
Unstructured:

Independent classifier
at each time-step



Semi-Markov Model

$$p(s,l,x) = \prod_{k} \underbrace{p(s_k|s_{k-1})}_{\text{tabular}} \underbrace{p(\mathrm{len}(s_k)|s_k)}_{\text{Poisson}} \prod_{t} \underbrace{p(x_t|l_t)}_{\text{Gaussian}}$$



✓ **Generative:** $p(a, x)$

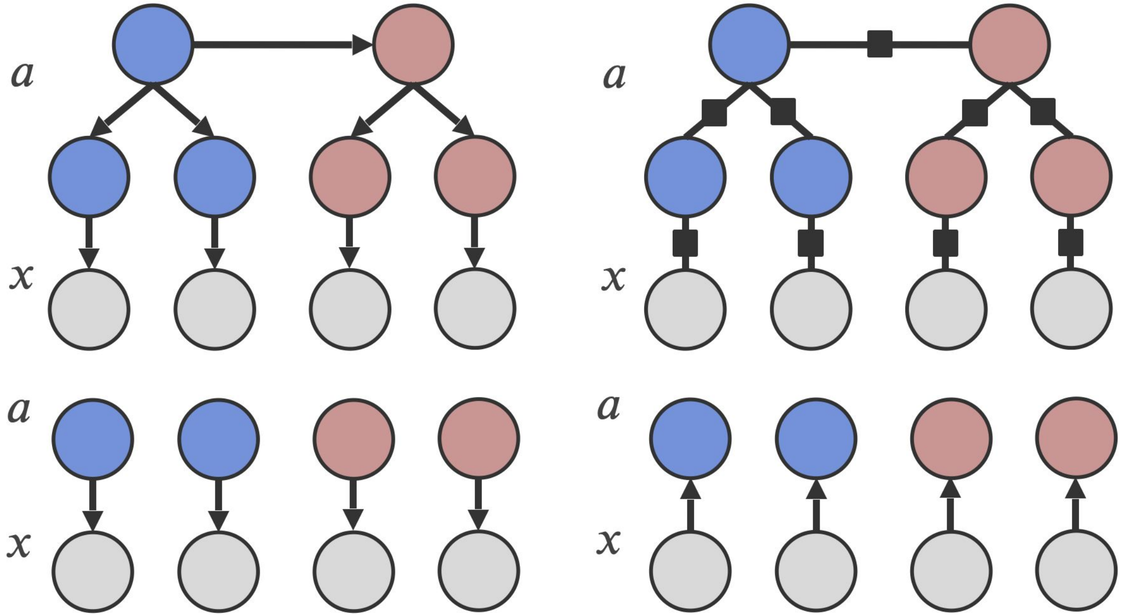
Discriminative: $p(a \mid x)$

✓ **Structured:**

Semi-Markov model

Unstructured:

Independent classifier
at each time-step



Evaluation

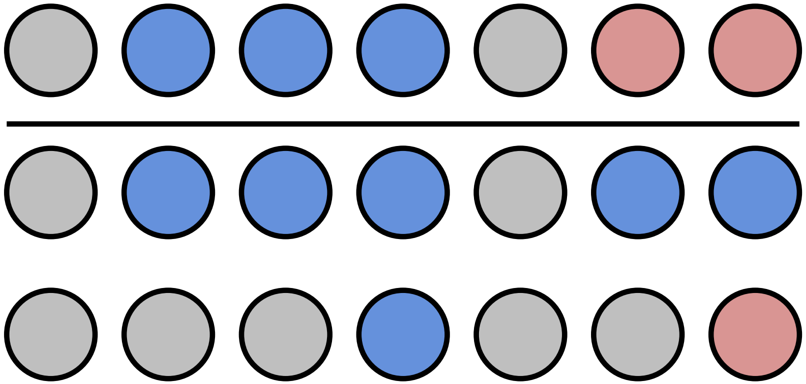
Two main metrics from past work:

- Timestep accuracy (1-second intervals) [Sener and Yao 2018, Richard et al. 2018, inter alia]
- Action recovery (with one timestep per action) [Alayrac et al. 2016, Zhukov et al. 2019]

Ground Truth:

Timestep Accuracy: 5/7
Actions Recovered: 1/2

Timestep Accuracy: 4/7
Actions Recovered: 2/2



Baselines

	Richard *et al.* 2018	Zhukov *et al.* 2019	Ours
generative model	✓		✓
step reordering	✓		✓
step repetitions	✓		✓
step duration	✓		✓
language		✓	✓

Picked models with non-overlapping strengths.

Supervised Results

		timestep accuracy	actions recovered	predicted bg %
Baselines	Ordered uniform	8.1	12.2	73.0
Unstructured	Discriminative linear	36.0	31.6	73.3
	✓Gaussian mixture	40.6	31.5	68.9
Structured	Zhukov *et al.* (2019)	18.1	**45.4**	*90.7*
	SMM, discriminative	37.3	24.1	65.9
	✓ SMM, generative	**49.4**	28.7	52.4

Un- & Weakly-Supervised Results

		timestep accuracy	actions recovered	predicted bg %
Baseline	Ordered uniform	8.1	12.2	73.0
Unsupervised	HSMM	28.8	10.6	*31.1*

Task orderings & narrations prevent over/underpredicting background.

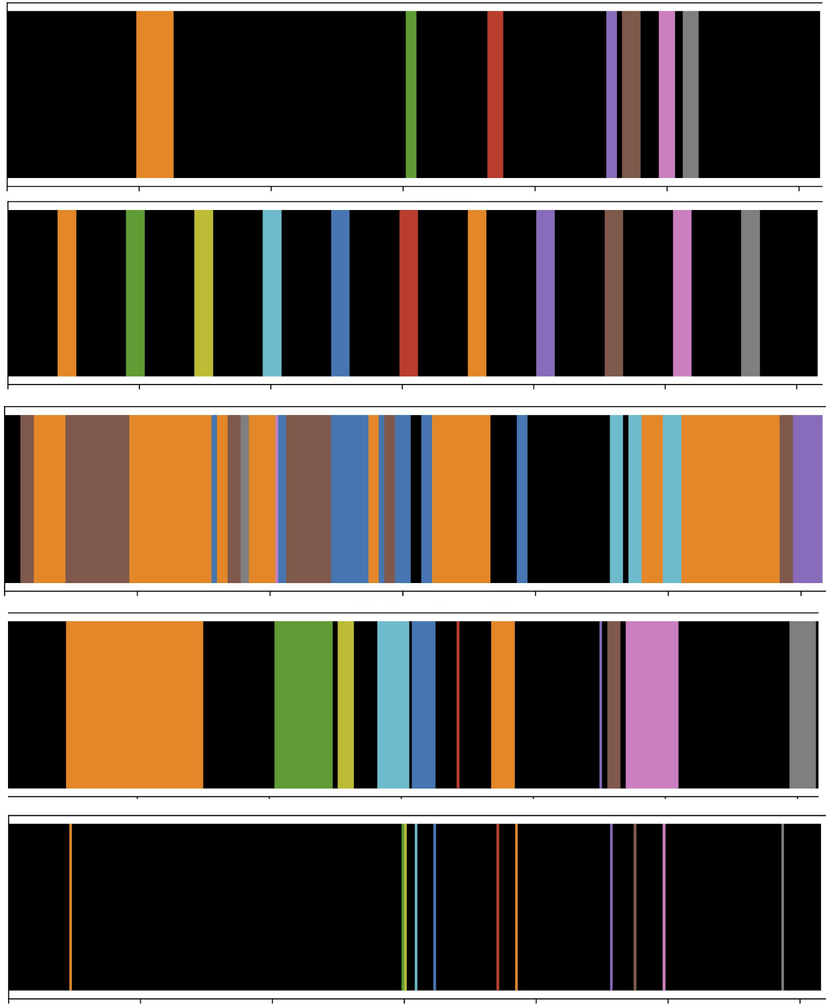
How Should We Model Action Segmentation?

Generative or discriminative models? Generative; discriminative ones overpredict bg.

Explicit structured modeling? Yes, especially for sequence-level metrics.

Do task ordering and narration help unsupervised models? Yes; prevent over/underpredicting bg.

segmentation examples



ground truth

uniform

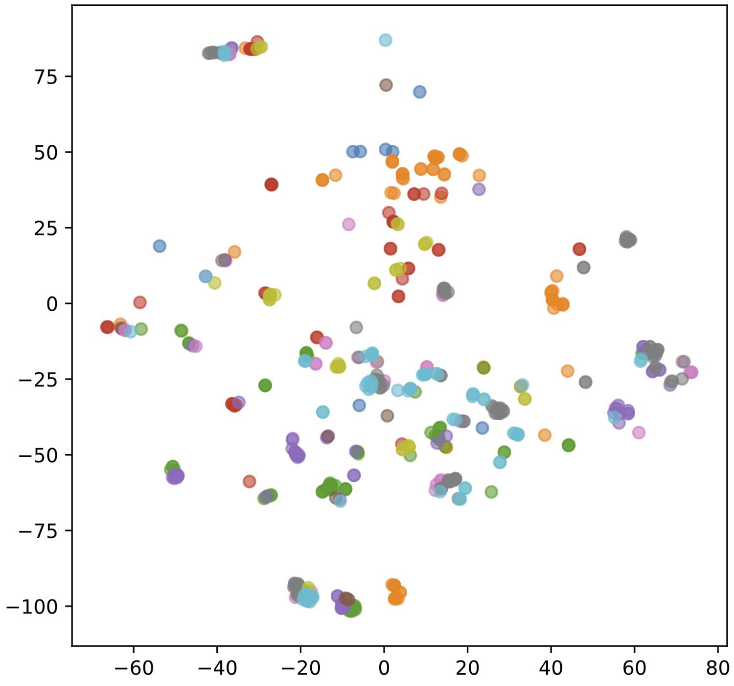
hsmm

hsmm + narration + ordering

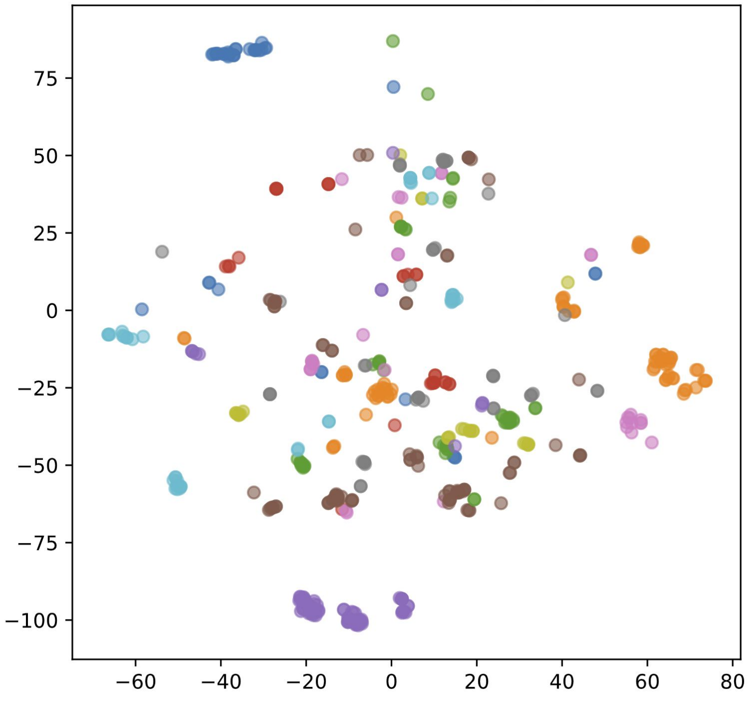
Zhukov *et al.* (2019)

Instructional Videos: Challenges

coloring by step



coloring by video



What Did We Learn? [Fried et. al, ACL 2020]

→ **Narrations** & **task orderings** are important inductive biases for action segmentation.

→ Need to report different metrics to evaluate action segmentation.

→ Deep visual features don't capture the finer-grained differences between actions.

to Translate Words

Learning ~~Language~~ by Observing the World

Translating Words Without Supervision

Je casse les oeufs.

different videos
in each language
(no paired data).

I need to mix the
eggs with the flour.

Two Ways to Translate Words Through Videos

Create a *paired* corpus using videos and then discard the videos,

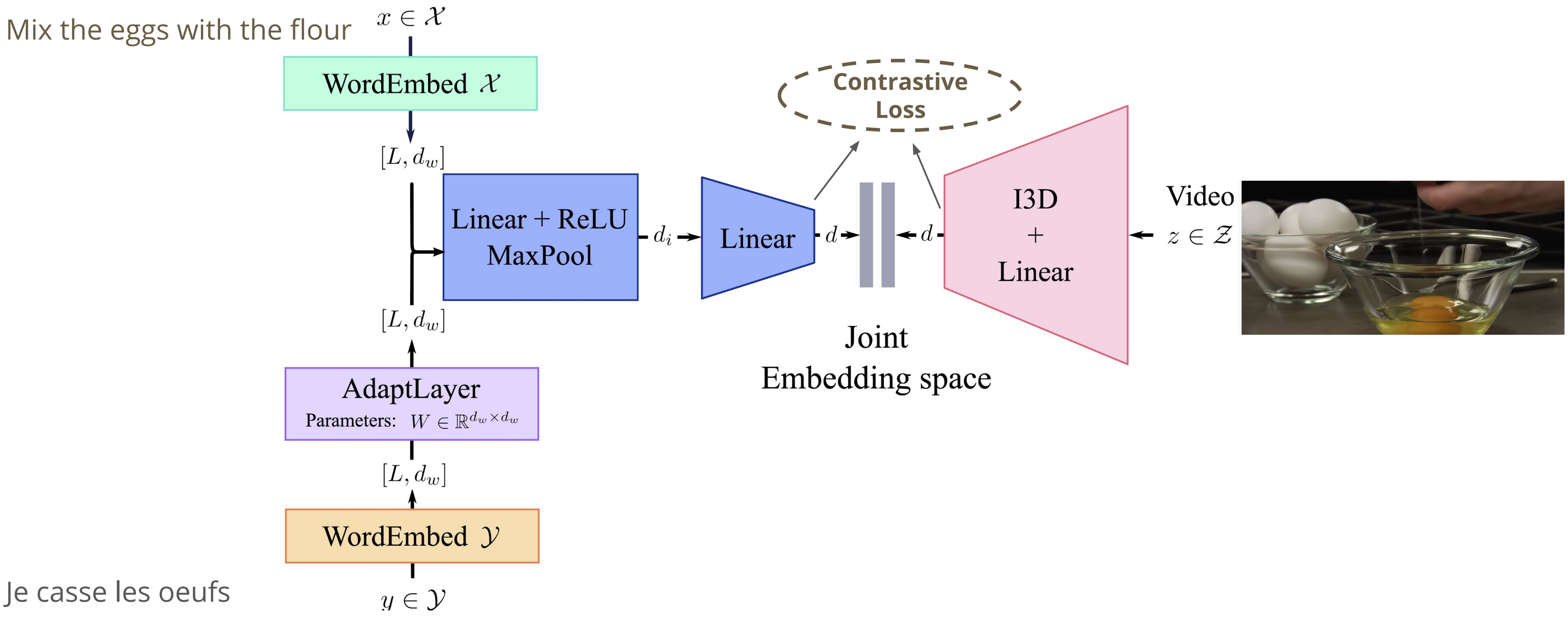
or,

learn a joint space between languages and videos, where the visual encoder is shared.

Creating a Paired Corpus

1. Compare English and French video features.
2. Select the nearest neighbors for each video.
3. Create a *paired* corpus with the narrations associated to neighboring videos.
4. Calculate joint probability between all En-Fr word pairs.

Our Base Model Architecture



The HowTo100M Dataset [Miech *et al.*, 2019]

~100M video clips-narrations (ASR output).

~23k high-level tasks.

Does a Shared Visual Encoder Help?

English-French (reporting recall@1)	Dictionary (Conneau *et al.*, 2017)	
	All	Visual
Random Chance	0.1	0.2
Paired Corpus	1.6	2.4
Base Model	**9.1**	**15.2**

Yes, especially for the visual words.

Failures When Pairing Two Languages

video (En) | nearest video (Fr)

“...stich getting color sequence...”

“...le pompon va se placer...”
(...the pompom will be placed...)

Videos are semantically similar but not the narrations.

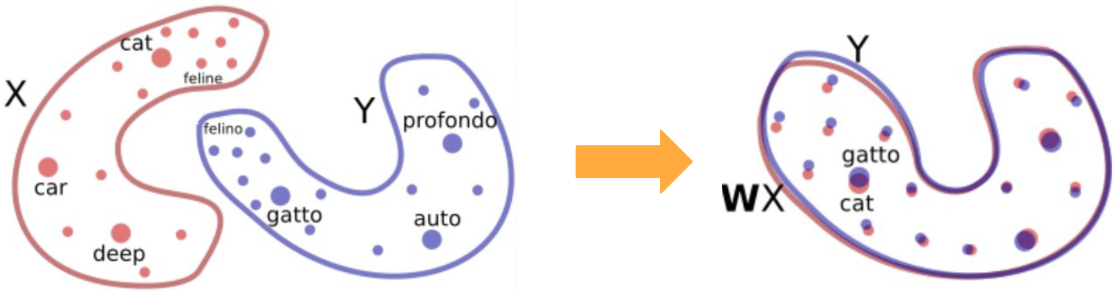
“...thank you for watching bye bye...”

“...j’ai besoin de curcuma et de clous...”
(...I need turmeric and cloves...)

Videos are semantically less similar.

Why Should We Use Videos at All?

Text-based methods align the space of the two languages.



Unsupervised approaches are possible:

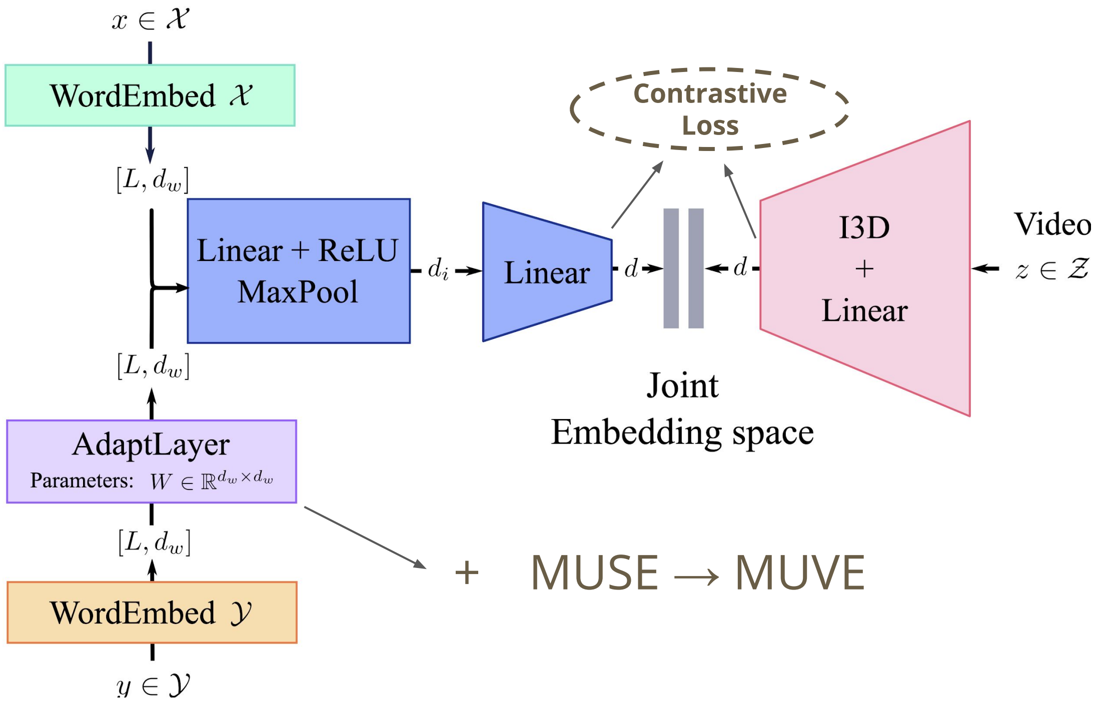
- MUSE (Conneau *et al.*, 2017)
- VecMap (Artetxe *et al.*, 2018)

Why Should We Use Videos at All?

But these methods are sensitive to the similarity of languages and their training corpora.

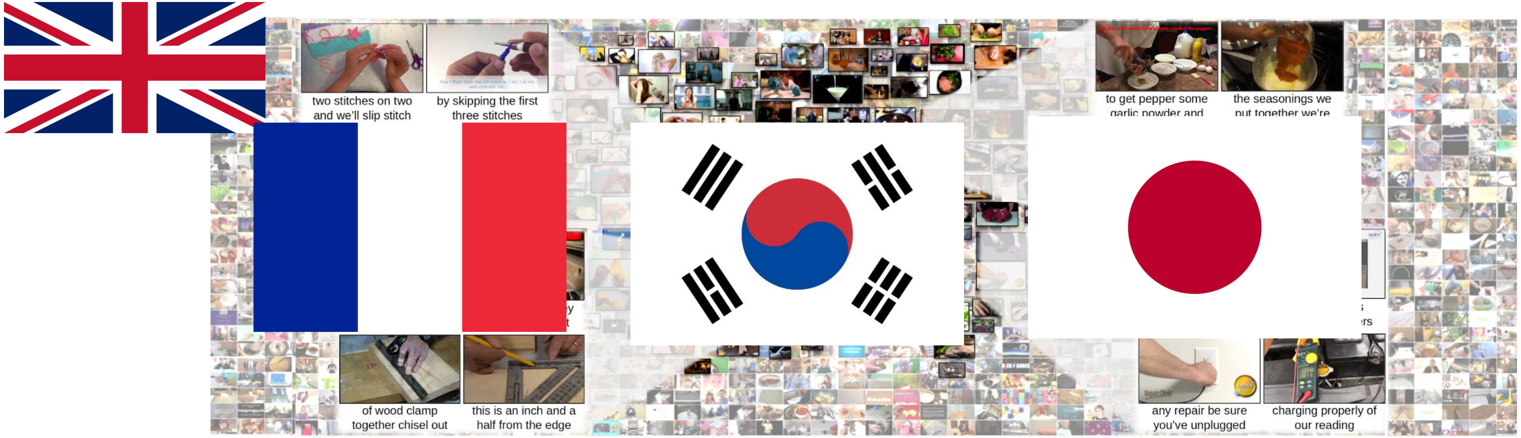
Can grounding improve unsupervised word translation -- make it more robust?

Multilingual Visual Embeddings (MUVE)



+ MUSE → MUVE

Extending the HowTo100M Dataset [Miech *et al*, 2019]



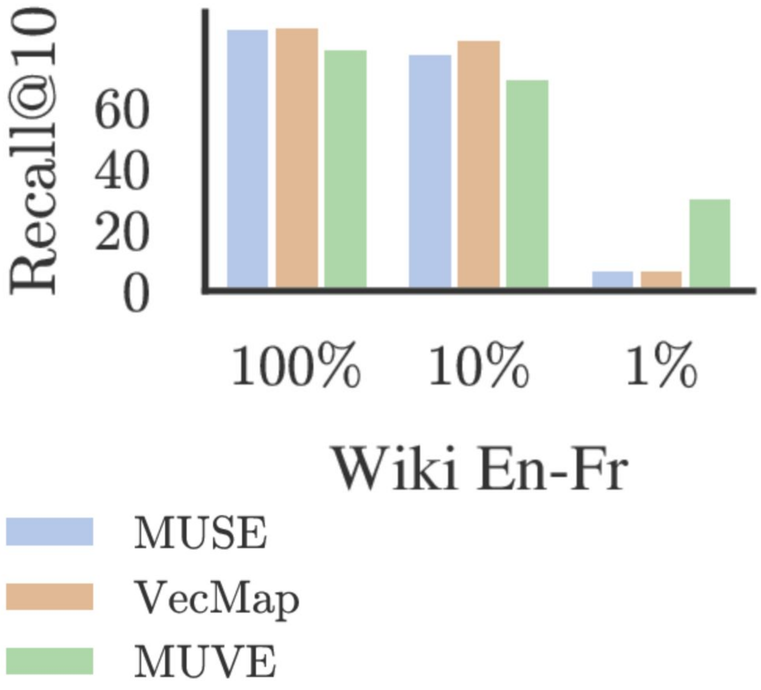
Performance of Models Across Language Pairs

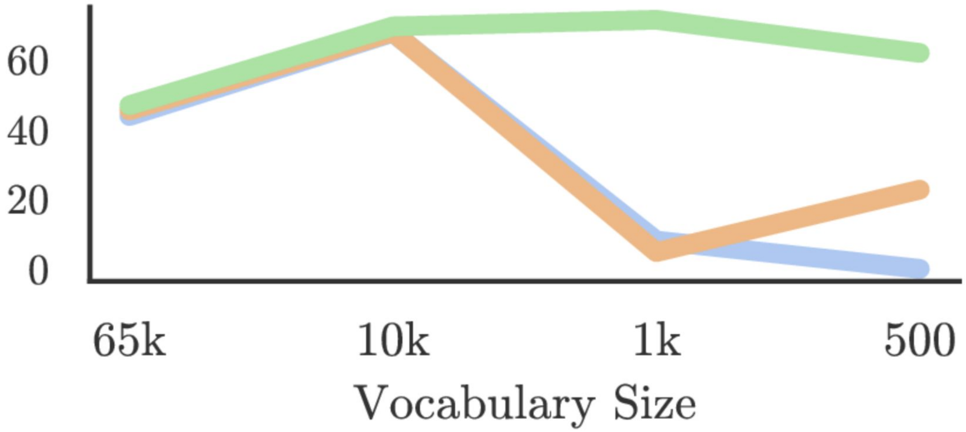
reporting recall@1	**En-Fr**
MUSE (Conneau *et al.*, 2017)	*26.3*
VecMap (Artetxe *et al.*, 2018)	*28.4*
MUVE (ours)	***28.9***
Supervised	57.9

Robustness to Dissimilarity of Corpora

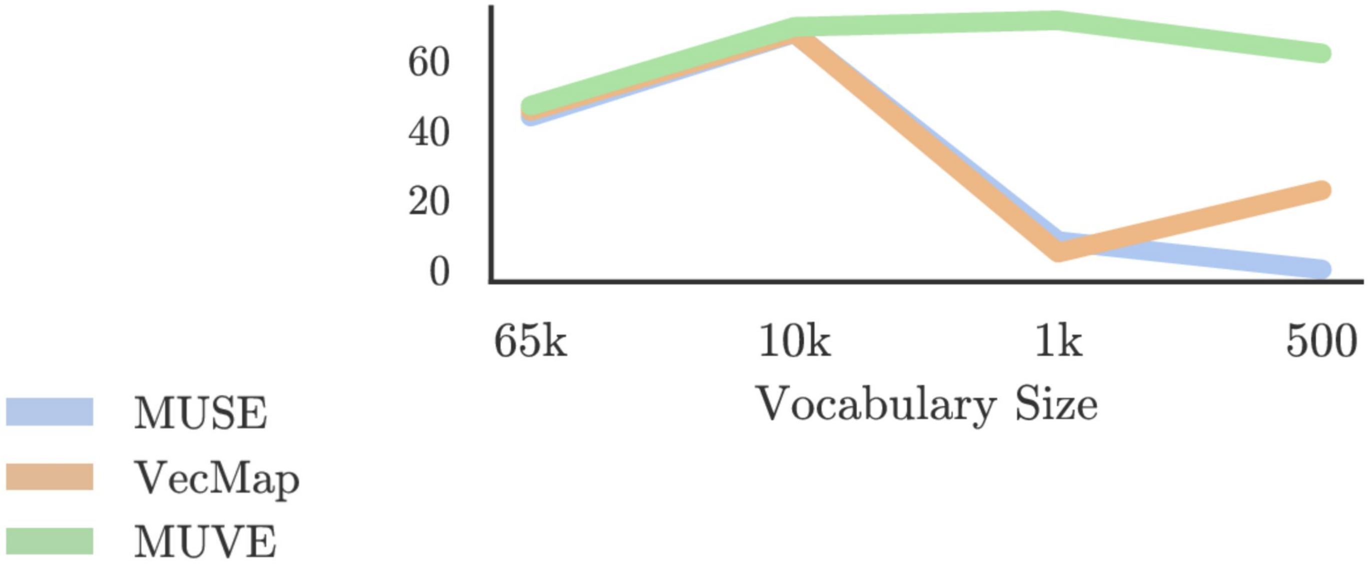
reporting recall@10	HowTo-Fr		
	MUSE **(Conneau *et al.*, 2017)**	**VecMap** **(Artetxe *et al.*, 2018)**	**MUVE** **(ours)**
HowTo-En	*45.8*	*45.4*	*47.3*

Robustness to Amount of Training Data





Robustness to Amount of Training Data



What Did We Learn? [Sigurdsson *et. al*, CVPR 2020]

Learning a joint multimodal space (sharing the visual encoder) improves word translations.

Grounding in videos improves text-based methods, especially in challenging situations.

Takeaways

Working with real-world data is an important first step for AI systems -- easy problems for humans are still challenging for our models.

Collecting less-noisy datasets is expensive. We need better ways of removing noise, *e.g.*, through better grounding.

Towards Better Multimodal Features

Recent multimodal transformers show impressive performance on a range of benchmarks.

Do these models provide features that better ground language to vision? [Hendricks *et. al*, TACL 2021]

Acknowledgments

Daniel Fried, Gunnar Atli Sigurdsson, Jean-Baptiste Alayrac, Lucas Smaira, Mateusz Malinowski, Joao Carreira, Chris Dyer, Stephen Clark, Phil Blunsom, and Andrew Zisserman.

Thanks!