

# Recipe1M

January 2019

# Recipe1M: A Dataset for Learning Cross-Modal Embeddings for Cooking Recipes and Food Images

Javier Marín<sup>1\*</sup>, Aritro Biswas<sup>1\*</sup>, Ferda Ofli<sup>2</sup>, Nicholas Hynes<sup>1</sup>, Amaia Salvador<sup>3</sup>, Yusuf Aytar<sup>1</sup>, Ingmar Weber<sup>2</sup>, Antonio Torralba<sup>1</sup>

<sup>1</sup>Massachusetts Institute of Technology <sup>2</sup>Qatar Computing Research Institute, HBKU

<sup>3</sup>Universitat Politècnica de Catalunya

{abiswas,nhynes}@mit.edu, {jmarin,yusuf,torralba}@csail.mit.edu, amaia.salvador@upc.edu, {fofli,iweber}@hbku.edu.qa

**Abstract**—In this paper, we introduce Recipe1M, a new large-scale, structured corpus of over one million cooking recipes and 13 million food images. As the largest publicly available collection of recipe data, Recipe1M affords the ability to train high-capacity models on aligned, multi-modal data. Using these data, we train a neural network to learn a joint embedding of recipes and images that yields impressive results on an image-recipe retrieval task. Moreover, we demonstrate that regularization via the addition of a high-level classification objective both improves retrieval performance to rival that of humans and enables semantic vector arithmetic. We postulate that these embeddings will provide a basis for further exploration of the Recipe1M dataset and food and cooking in general. Code, data and models are publicly available.

**Index Terms**—Cross-modal, deep learning, cooking recipes, food images



# Contributions

- Address data limitation for food
- Multimodal neural model

# Related Work

- Herranz et al
- Min et al
- Carvalho et al
- Chen et al

**Data Input** 

Cuisine	American	Indian	Chinese
Course	Main Dishes	Side Dishes	Appetizers, Lunch
Flavors	piquant: 0.6667 sour: 0.3333 salty: 0.8333 sweet: 0.1667 bitter: 0.5000 meaty: 0.6667	piquant: 0.5 sour: 0.5 salty: 0.8333 sweet: 0.1667 bitter: 0.8333 meaty: 0.1667	piquant: 0.1667 sour: 0.1667 salty: 0.1667 sweet: 0.0 bitter: 0.1667 meaty: 0.8333
Ingredients	"1 large onion, diced", "1 tablespoon olive oil", "1 large green bell pepper, diced", .....	"oil", "onions", "green chilies", "cumin seed", "cooked rice", "turmeric", "chili powder", "water", "salt", "cilantro"	"chicken wings", "paprika", "crushed red pepper flakes", "ground white pepper", "szechwan peppercorns",



# Data Collection

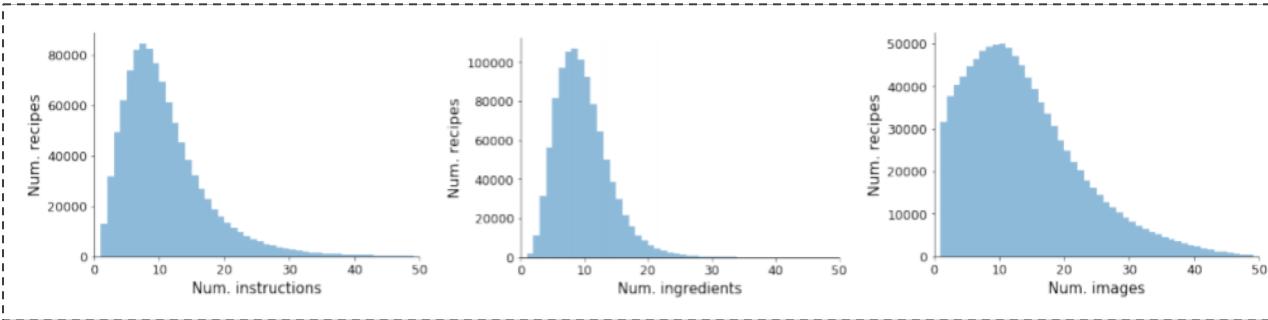
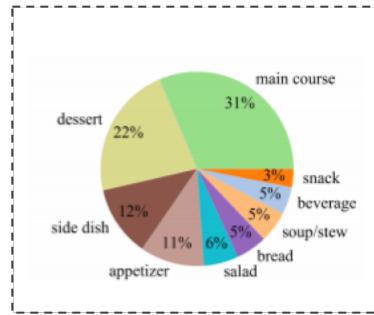
- Recipe1M
- Recipe1M+

TABLE 1  
**Recipe1M dataset.** Number of recipes and images in training, validation and test sets.

		Recipe1M	intersection	Recipe1M+
Partition	# Recipes	# Images	# Images	# Images
Training	720,639	619,508	493,339	9,727,961
Validation	155,036	133,860	107,708	1,918,890
Test	154,045	134,338	115,373	2,088,828
Total	1,029,720	887,706	716,480	13,735,679

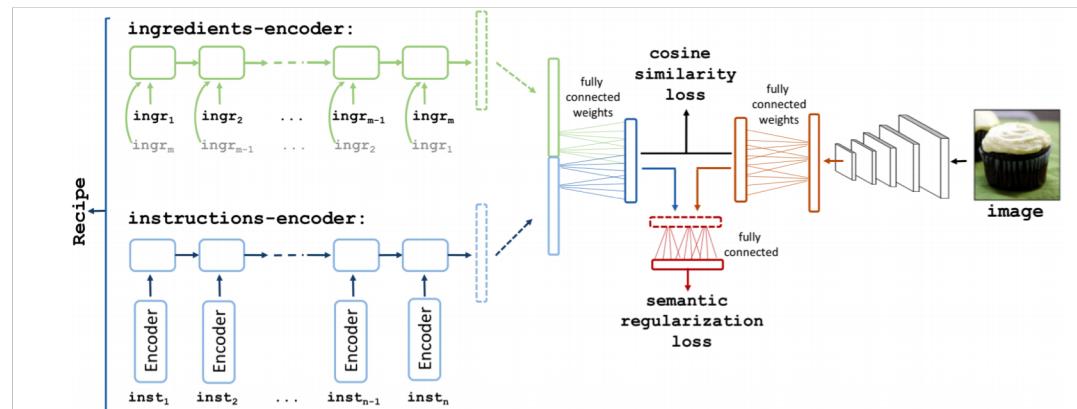
# Data Structure

- Recipe1M
- Recipe1M+



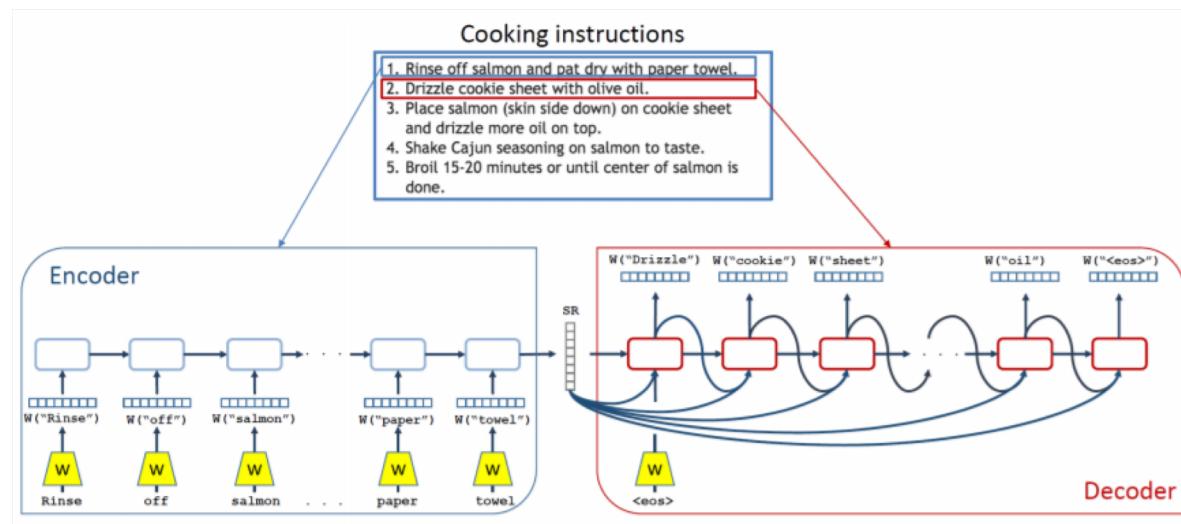
# Learning Embeddings

- Ingredients
- Cooking Instructions
- Images



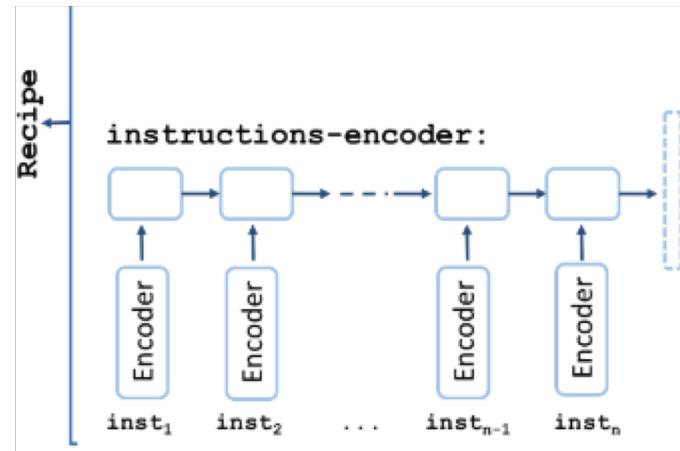
# Cooking Instructions Model

First Stage



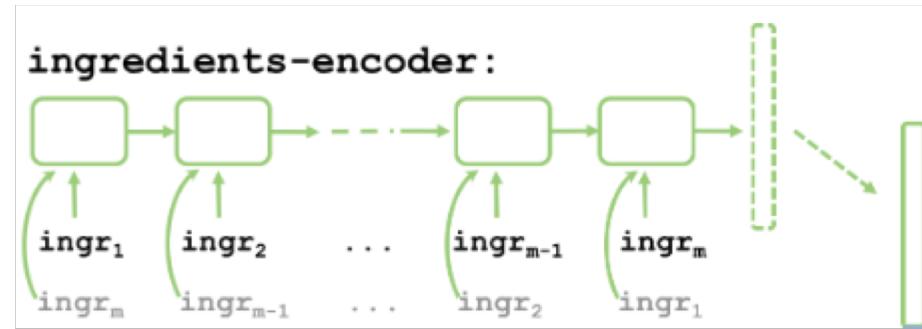
# Cooking Instructions Model

Second stage



# Ingredients Model

Bi-Directional LSTM



# Joint Neural Embedding

Loss Function

$$\phi^r = W^r h_k^r + b^r \text{ and } \phi^v = W^v h_k^v + b^v$$

$$L_{cos}(\phi^r, \phi^v, y) = \begin{cases} 1 - \cos(\phi^r, \phi^v), & \text{if } y = 1 \\ \max(0, \cos(\phi^r, \phi^v) - \alpha), & \text{if } y = -1 \end{cases}$$

Semantic Regularization

$$L(\phi^r, \phi^v, c_r, c_v, y) = L_{cos}(\phi^r, \phi^v, y) + \lambda L_{reg}(\phi^r, \phi^v, c_r, c_v)$$

# Implementation and Optimization

- LMDB data format
- 3-stage training
- MedR as performance measure

# Experiments

## Retrieval

	im2recipe				recipe2im			
	medR	R@1	R@5	R@10	medR	R@1	R@5	R@10
random ranking	500	0.001	0.005	0.01	500	0.001	0.005	0.01
CCA w/ skip-thoughts + word2vec (GoogleNews) + image features	25.2	0.11	0.26	0.35	37.0	0.07	0.20	0.29
CCA w/ skip-instructions + ingredient word2vec + image features	15.7	0.14	0.32	0.43	24.8	0.09	0.24	0.35
joint emb. only	7.2	0.20	0.45	0.58	6.9	0.20	0.46	0.58
joint emb. + semantic	5.2	0.24	0.51	0.65	5.1	0.25	0.52	0.65

# Experiments

## Retrieval

Query Image	True ingr.	Retrieved ingr.	Retrieved Image
	whole milk half - and - half cr white sugar lemon extract ground cinnamon frozen blueberries vanilla wafers ice cubes	berries strawberry yogurt banana milk white sugar	
	butter garlic cloves all - purpose flour kosher salt milk chicken broth mozzarella cheese parmesan cheese onion	1 box any pasta you ground beef 1 envelope taco seas water 1/2 packages cream c cheese	
	cooked white rice salt shrimp Broccolini mayonnaise nori	sushi rice salmon avocado cream cheese nori	
	mayonnaise onion cider vinegar sugar celery seeds green cabbage carrot salt & freshly groun ground chuck	yellow onion coarse salt ground pepper ground chuck buns eggs ketchup canned beets lettuce leaves	

Fig. 8. **Im2recipe retrieval examples.** From left to right: (1) the query image, (2) its associated ingredient list, (3) the retrieved ingredients and (4) the image associated to the retrieved recipe.

# Experiments

## Ablation Studies

- Changing test sizes

TABLE 4  
**Ablation studies.** Effect of the different model components to the median rank, medR (the lower is better).

	Joint emb. methods	im2recipe			recipe2im		
		medR-1K	medR-5K	medR-10K	medR-1K	medR-5K	medR-10K
VGG-16	fixed vision	15.3	71.8	143.6	16.4	76.8	152.8
	finetuning (ft)	12.1	56.1	111.4	10.5	51.0	101.4
	ft + semantic reg.	8.2	36.4	72.4	7.3	33.4	64.9
ResNet-50	fixed vision	7.9	35.7	71.2	9.3	41.9	83.1
	finetuning (ft)	7.2	31.5	62.8	6.9	29.8	58.8
	ft + semantic reg.	5.2	21.2	41.9	5.1	20.2	39.2

# Experiments

Human performance comparison

AMT worker

TABLE 5

**Comparison with human performance on im2recipe task.** The mean results are highlighted as bold for better visualization. Note that on average our method with semantic regularization performs better than average AMT worker.

	all recipes	course-specific recipes					dish-specific recipes									
		dessert	salad	bread	beverage	soup-stew	course-mean	pasta	pizza	steak	salmon	smoothie	hamburger	ravioli	sushi	dish-mean
human	<b>81.6 ± 8.9</b>	52.0	70.0	34.0	58.0	56.0	<b>54.0 ± 13.0</b>	54.0	48.0	58.0	52.0	48.0	46.0	54.0	58.0	<b>52.2 ± 04.6</b>
joint-emb. only	<b>83.6 ± 3.0</b>	76.0	68.0	38.0	24.0	62.0	<b>53.6 ± 21.8</b>	58.0	58.0	58.0	64.0	38.0	58.0	62.0	42.0	<b>54.8 ± 09.4</b>
joint-emb.+semantic	<b>84.8 ± 2.7</b>	74.0	82.0	56.0	30.0	62.0	<b>60.8 ± 20.0</b>	52.0	60.0	62.0	68.0	42.0	68.0	62.0	44.0	<b>57.2 ± 10.1</b>

# Experiments

## Recipe1M vs Recipe1M+ Comparison

TABLE 6

**Comparison between Recipe1M and Recipe1M+ trained models.** Median ranks and recall rate at top  $K$  are reported for both models. They have similar performance on the Recipe1M test set in terms of medR and R@K. However, when testing on the Recipe1M+ test set, the model trained on Recipe1M+ yields significantly better medR and better R@5 and R@10 scores.

	Recipe1M test set				Recipe1M+ test set			
	im2recipe							
	medR	R@1	R@5	R@10	medR	R@1	R@5	R@10
Recipe1M training set	5.1	0.24	0.52	0.64	13.6	0.15	0.35	0.46
Recipe1M+ training set	5.7	0.21	0.49	0.62	8.6	0.17	0.42	0.54
recipe2im								
	medR	R@1	R@5	R@10	medR	R@1	R@5	R@10
Recipe1M training set	4.8	0.27	0.54	0.65	11.9	0.17	0.38	0.48
Recipe1M+ training set	4.6	0.26	0.54	0.66	6.8	0.21	0.46	0.58

# Experiments

## Model Generalization Ability Comparison

TABLE 7  
**Im2recipe retrieval comparisons on Food-101 dataset.** Median ranks and recall rate at top  $K$  are reported for both models. Note that the model trained on Recipe1M+ performs better than the model trained on Recipe1M.

	im2recipe			
	medR	R@1	R@5	R@10
Recipe1M training set	17.35	16.13	33.68	42.53
Recipe1M+ training set	10.15	21.89	42.31	51.14
recipe2im				
Recipe1M training set	4.75	26.19	54.52	67.50
Recipe1M+ training set	2.60	37.38	65.00	76.31

# Experiments: (Analysis of the Learned Embedding)

## Neuron visualization

	Top 4 images	Top 2 ingredients	Top 2 instructions
unit 352		vanilla extract heavy cream sugar nutmeg creme fraiche all purpose flour potatoes garlic cloves chunks	Start with bowl and beaters cold! In a large bowl, whip cream until stiff peaks are ju... Beat In vanilla and sugar until stiff peaks form. Do not overbeat!
unit 386		tomatoes garlic fillets leaf vinegar tomato paste carrots cashews dates milk sugar	Cook pan with cooking oil and pan fry Mahi Mahi fill... To prepare sauce, saute garlic and shallots in pan. Stir in chicken stock and simmer until sauce thickens. Remove from heat and add basil. To Serve, top Mahi Mahi fillets with generous helping... Garnish with a pretty whole basil leaf or bunch of ...
unit 144		onion mung beans chedd_leaves chill_pepper vegetable_oil coconut_milk onion fresh spinach mushroom olive油 soy_sauce black_pepper	Fry bacon in a Dutch oven until almost done. Add onions and garlic and saute until the onions are... Cover the bacon, onions and garlic with 4 cups water... Add wine, soy sauce, salt, hot sauce and collards. Return to a boil and simmer for 1 hour.
unit 22		butter milk vanilla blend baking_powder sugar pudding almond_extract water yellow_cake_mix oil powdered_sugar	Preheat oven to 350F. Beat butter and sugar in large bowl with electric mi... Add eggs, one at a time, beating well after each add... Add cheese and sour cream; mix well. Bake 40 min. Cool completely.
unit 571		steaks garlic_powder brown_sugar onion_powder rasp black_pepper green_pepper swiss_cheese steak italian_dressing tomato_paste beef_broth	Heat grill to medium heat. Mix all ingredients except steaks; rub onto both sid... Grill 6 to 8 min. Remove from grill. Let stand 5 min. before serving.

Fig. 9. **Localized unit activations.** We find that ingredient detectors emerge in different units in our embeddings, which are aligned across modalities, (e.g., unit 352: "cream", unit 22: "sponge cake" or unit 571: "steak").

# Experiments

## Semantic Vector Arithmetic

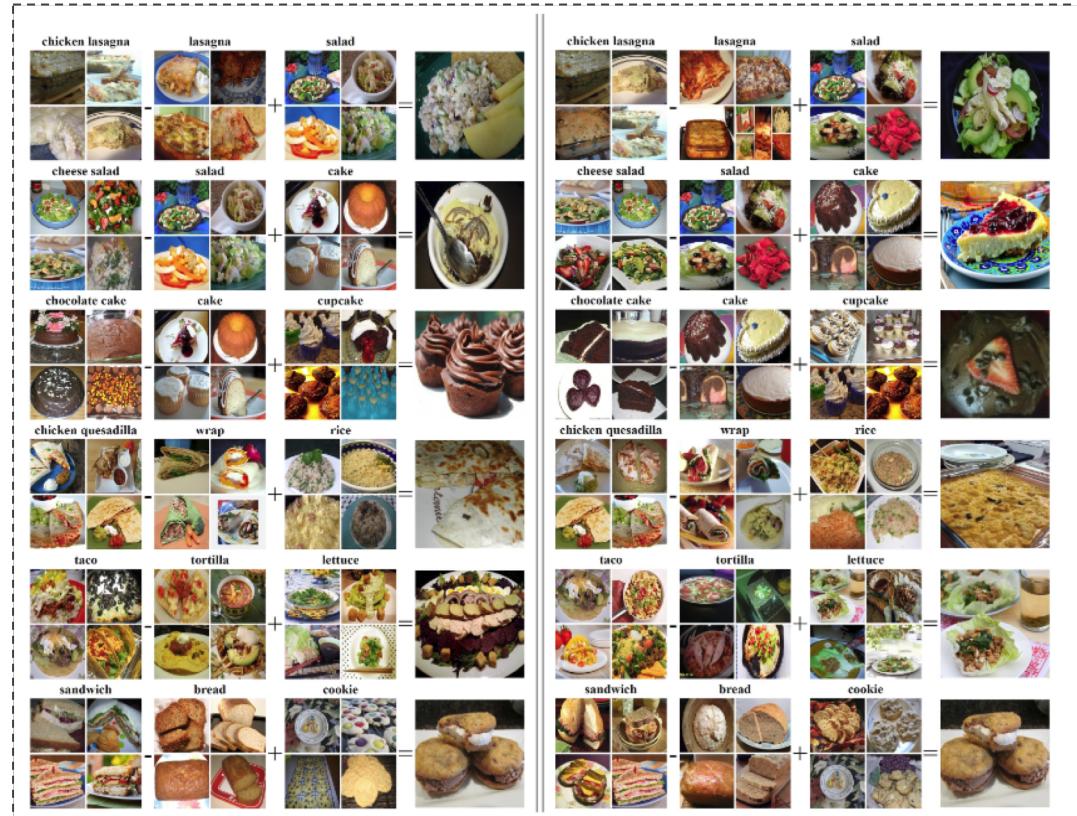
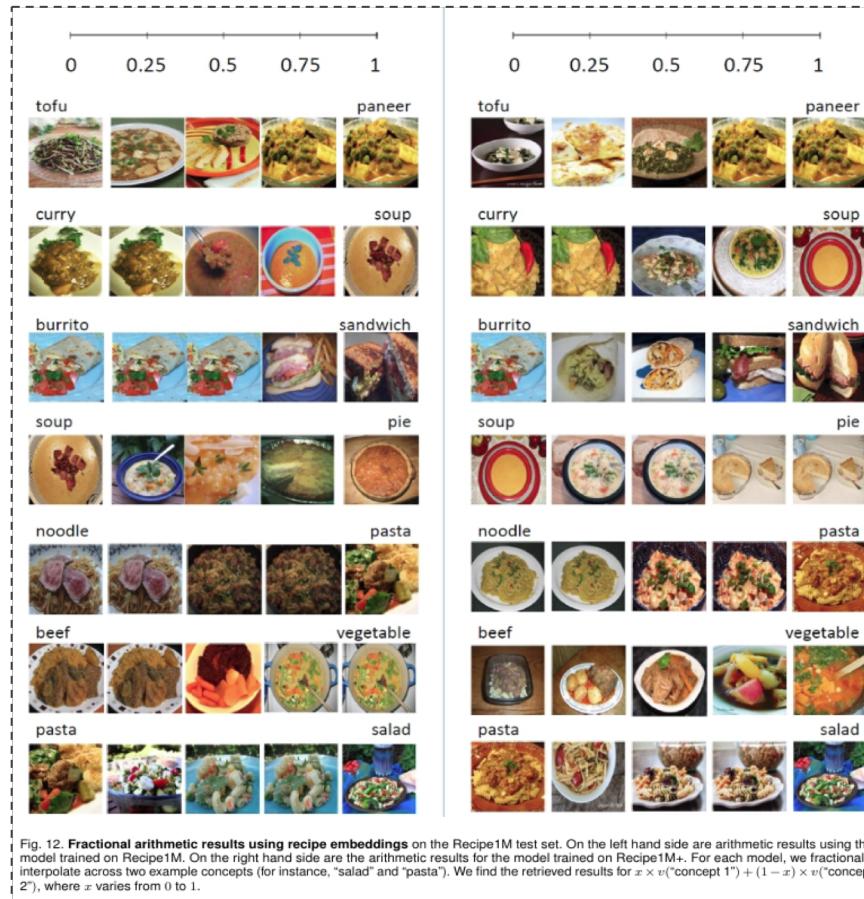


Fig. 10. **Analogy arithmetic results using recipe embeddings** on the Recipe1M test set. On the left hand side are analogy results using the model trained on Recipe1M. On the right hand side are the analogy results for the model trained on Recipe1M+. We represent the average vector of a query with the images from its 4 nearest neighbors. In the case of the arithmetic result, we show the nearest neighbor only.

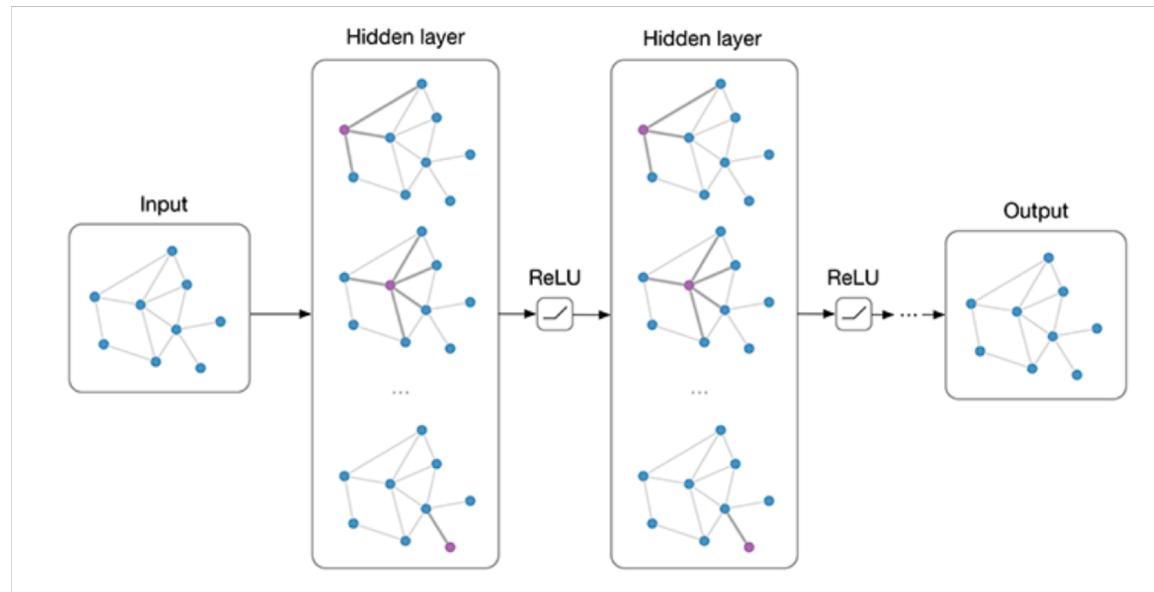
# Experiments

## Fractional Arithmetic



# Additional thoughts

Graph Model



# Ingredients Model

Node Initialization

Adjacency Matrix

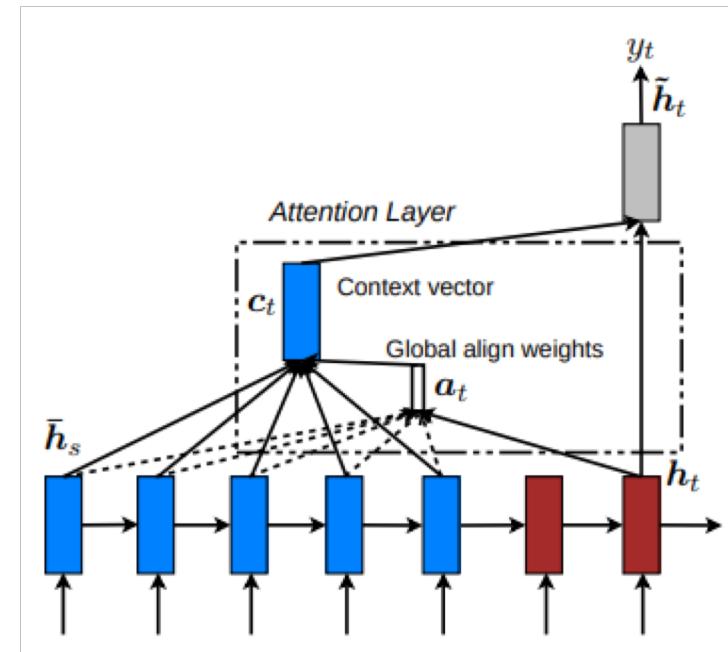
Edge Weights

$$a_v^{(k)} = \text{MEAN} \left( \left\{ \text{ReLU} \left( W \cdot h_u^{(k-1)} \right), \forall u \in \mathcal{N}(v) \right\} \right)$$

$$h_G = \text{READOUT}(\{h_v^{(K)} \mid v \in G\})$$

# Additional Thoughts

Attention



Questions?