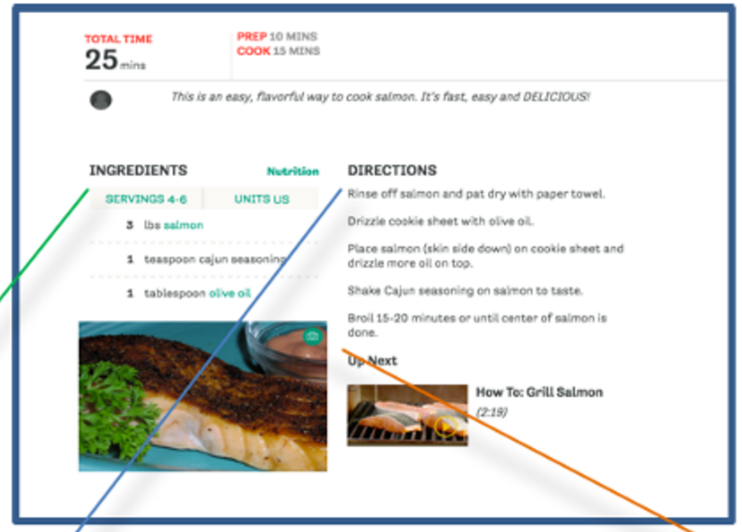
Abstract

- Recipe1M+: Largest public dataset with 1M+ recipes & 13M food images.
- Multimodal Learning: Trains AI models on aligned recipe-image data.
- Joint Embedding: Neural network links recipes & images for accurate retrieval.
- Better Performance: High level of Classification & Regularization boosts retrieval to human-level accuracy.
- Open Source: Code, data & models available.

INTRODUCTION

1. Cultural & Emotional Significance – Food is vital to health, emotions, and culture.
2. Computer Vision Challenges – Recognizing ingredients despite transformations (e.g., cutting, cooking) poses difficulties.
3. AI & Food Analysis – Machines can now learn from online recipes and food images, aiding cooking, public health, and cultural studies.
4. Need for Datasets – Developing these tools requires large, structured datasets.

Recent advances in AI, powered by large labeled datasets and deep learning, have greatly improved object recognition and scene understanding. These techniques also helped in tasks like detailed image labeling and segmentation. Introducing a large-scale food dataset could push the field even further, especially since food poses unique challenges—like identifying ingredient states (sliced, grilled, boiled, etc.), which current datasets don’t handle well. Unlike rigid classification tasks, food requires a more flexible approach to account for variations in recipes and preparation methods.

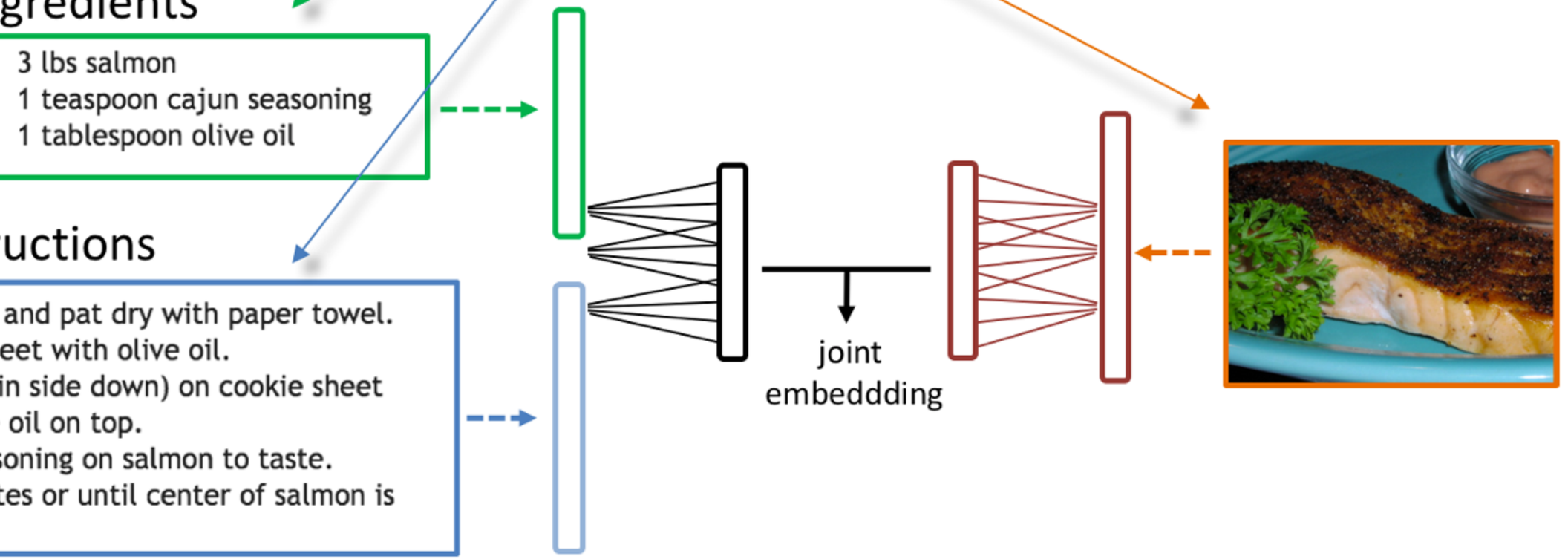


Ingredients

- 3 lbs salmon
- 1 teaspoon cajun seasoning
- 1 tablespoon olive oil

Cooking instructions

1. Rinse off salmon and pat dry with paper towel.
2. Drizzle cookie sheet with olive oil.
3. Place salmon (skin side down) on cookie sheet and drizzle more oil on top.
4. Shake Cajun seasoning on salmon to taste.
5. Broil 15-20 minutes or until center of salmon is done.



1. Cross-Modal Learning
    - Figure 1 demonstrates that pairing recipe text (ingredients + instructions) with food images helps AI understand cooking processes better.
2. Limitations of Current Datasets
    - Most research uses small/medium-sized food image datasets (e.g., Food-101).
    - Initial accuracy was 50.8%, later improved to ~80%, showing dataset size is a bottleneck.
3. Unaddressed Challenges
    - Some studies (e.g., Myers et al.) estimated meal calories but did not share critical data (like segmentation maps), limiting further research.
4. Need for Larger, Richer Datasets
    - Progress in food AI requires bigger datasets with detailed annotations (e.g., ingredient states, cooking methods).

1. Solution to Data Limitations
    - Introduces Recipe1M+, a large-scale dataset with:
    - 1 million+ structured recipes
    - Paired food images
2. Innovative Task: im2recipe Retrieval
    - Goal: Find recipes from food images (even hard-to-describe dishes).
    - Uses both text (ingredients/instructions) and images together.
3. Advanced AI Model
    - Multimodal neural network embeds recipes & images in a shared space.
    - Boosted by semantic regularization (high-level food categories).
4. Strong Performance
    - Outperforms baseline methods significantly.
    - Matches human-level accuracy in some tasks.
5. Future Impact
    - Enables new research in food AI (beyond retrieval).
    - Open dataset to solve unimagined challenges in cooking/health.

- Recipe1M+ = Big dataset (1M recipes + images).
- im2recipe task = Find recipes from photos.
- AI model combines text + images intelligently.
- Works better than old methods, close to humans.
- Opens doors for future food-tech research.

Related Work Overview

- Several studies have explored food-related AI and multimodal learning since 2017.
- Focus areas include image-recipe retrieval, food attribute analysis, and recipe structuring.

Multimodal Frameworks

- Herranz et al.: Proposed an advanced model using food images, recipes, nutrition, location, restaurant menus, and food styles.
- Min et al.: Developed the MATM approach to analyze cuisine type, meal course, and flavors using Yummly data.

Recipe Comparison & Structuring

- Chang et al.: Studied variations of the same dish (e.g., "chocolate chip cookies") by clustering recipe similarities.
- Their model manually selects features, while Recipe1M+ learns them automatically using AI.

2 DATASET

1. Challenges in Food AI
    - Food data is complex due to:
    - Multiple ingredient variations
    - Different cooking styles
    - Diverse presentations
2. Limitations of Existing Datasets
    - Most datasets have only images (e.g., Food-101) or only text (no images).
    - Few multimodal datasets exist but are:
    - Small (e.g., Wang et al.: 101K images + raw HTML recipes).
    - Limited to specific cuisines (e.g., Chen & Ngo: Chinese food only).

3.Recipe1M+ Breakthrough

- 1M+ structured recipes + 13M images (largest public collection).
- Created in two phases:
  i. Scraped recipes + images from cooking sites (1M recipes, 800K images).
  ii. Expanded with web-sourced images (total: 13M after deduplication).
- 130× more images and 2× more recipes than previous largest datasets.

4.Impact

- Enables training powerful AI models for:
- Cross-modal retrieval (image ↔ recipe).
- Fine-grained analysis (ingredient states, cooking methods).

Data Collection from Recipe Websites

Data Collection Overview

- Scraped recipes from over two dozen popular cooking websites.
- Processed data by extracting text from HTML, downloading linked images, and organizing it into a JSON schema.

Data Cleaning & Deduplication

- Removed excessive whitespace, HTML entities, and non-ASCII characters.
- Removed duplicates (2% of data).
- Final dataset includes 1M+ recipes and 800K images.

Dataset Scale & Expansion

Larger than previous datasets in the domain:

- Twice as many recipes.
- Eight times more images.
- Plan to extend image collection through an image search engine.

2.2 Data Extension using Image Search Engine

1. Google Image Search & Data Collection (Fig. 2 & Text)

- Query Example: Searched "chicken wings" on Google Images (Fig. 2 shows high-quality results).
- Scale: Targeted 50M images (50 per recipe) using automated tools (Python + Aria2).
- Challenges:
  - Duplicates: Removed 32M+ identical images (Euclidean distance = 0).
  - Near-Duplicates: Filtered resized/cropped/text-added variants with strict thresholds.
  - Noise: Eliminated non-recipe images (e.g., faces, nutrition labels) using face detection and feature analysis.

2. Dataset Composition (Table 1)

- Recipe1M+: 1M+ recipes + 13.7M images (largest public food dataset).
  - Training Set: 720K recipes, 9.7M images.
  - Validation/Test Sets: ~155K recipes each, ~2M images each.
- Key Improvement:
  - 130× more images than prior datasets (e.g., Chen & Ngo's 110K images).
  - Balanced Partitions: Ensured no overlap between training/validation/test sets.

3. Data Cleaning Innovations

- Deduplication: Used ResNet18 to detect duplicates via feature distances.
- Uniform Distribution: Assigned images evenly for recipes with identical titles.
- Efficiency: Used C++ for fast distance calculations.

4. Impact

- Enables Robust AI Models: Clean, large-scale data improves tasks like:
  - Image-to-recipe retrieval.
  - Cross-modal embedding (text + visuals).
- Open Resource: Public release accelerates food AI research.

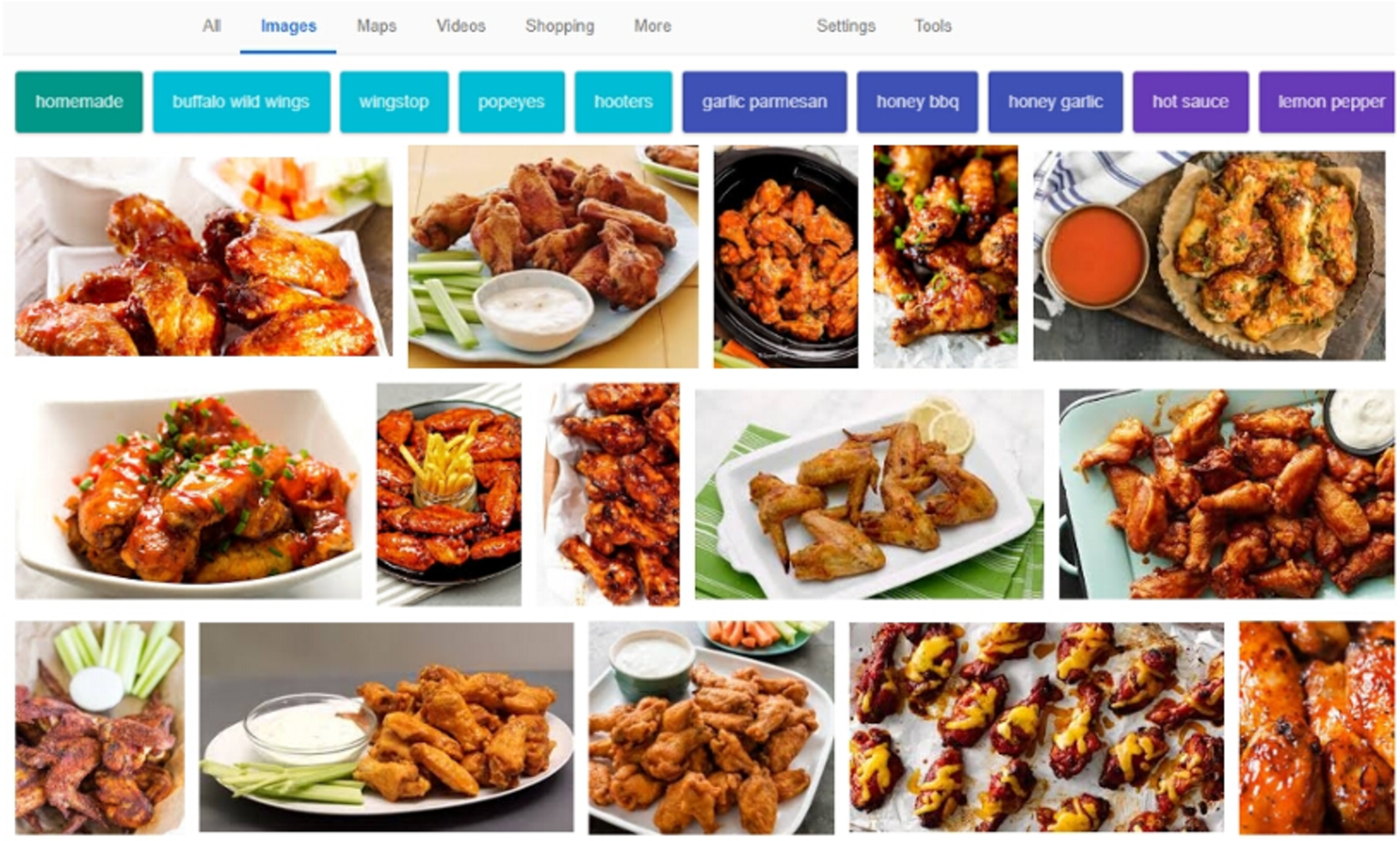


Fig. 2. **Google image search results.** The query used is *chicken wings*.

TABLE 1
Dataset sizes. Number of recipes and images in training, validation and test sets of each dataset.

		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

Nutritional Information

Ingredient Data Processing

- Ingredient lists from recipes were originally combined in single sentences (ingredient, quantity, and unit).

- Separated these three fields into distinct data points for easier nutritional analysis.

- Used Natural Language Processing (NLP) to tag words and identify the quantity-unit-ingredient pattern in the recipe sentences.

Data Cleaning & Ingredient Matching

- Identified measurable units for ingredient quantities (e.g., teaspoons, cups).
- Created a corpus of unique ingredients by matching the names with a USDA nutrient database.
- Filtered out non-measurable units and non-standard ingredient names to ensure accurate matching.

Nutritional Information & Visualization

- Matched 68,450 recipes with ingredients from the USDA database, yielding 50,637 recipes with complete nutritional data.
-
- Visualized recipe embeddings using t-SNE, categorizing recipes by semantic category and healthiness (sugar, fat, salt, etc.).

2.4 Data Structure

- **1. Data Structure (Figs. 3-4 & Text)**
- Layer 1 (Core Recipe Data):
- Textual Components:
- Recipe titles
- Ingredient lists (with units/quantities where possible)
- Step-by-step instructions
- Nutritional Data:

Calories, protein, sugar, fat, salt (for recipes with measurable units).

- Uses FSA traffic light system (Fig. 4) to label healthiness (red/amber/green).
- Layer 2 (Multimedia Extension):

13M+ food images (JPEG format) paired with recipes.

Course Labels: Some recipes tagged with meal types

2. Semantic Embeddings (Fig. 3)

- t-SNE Visualization:
- Clusters recipes by top 12 semantic categories (e.g., "chicken salad," "chocolate cake").
- Shows how recipes relate in the AI’s "understanding" space.

3. Health Analysis (Fig. 4)

- Embedding Maps Health Metrics:
- Color-coded by sugar/fat/salt content (FSA standards).
- Enables tasks like:
- Retrieving healthier recipe variants.
- Analyzing nutritional patterns across cuisines.

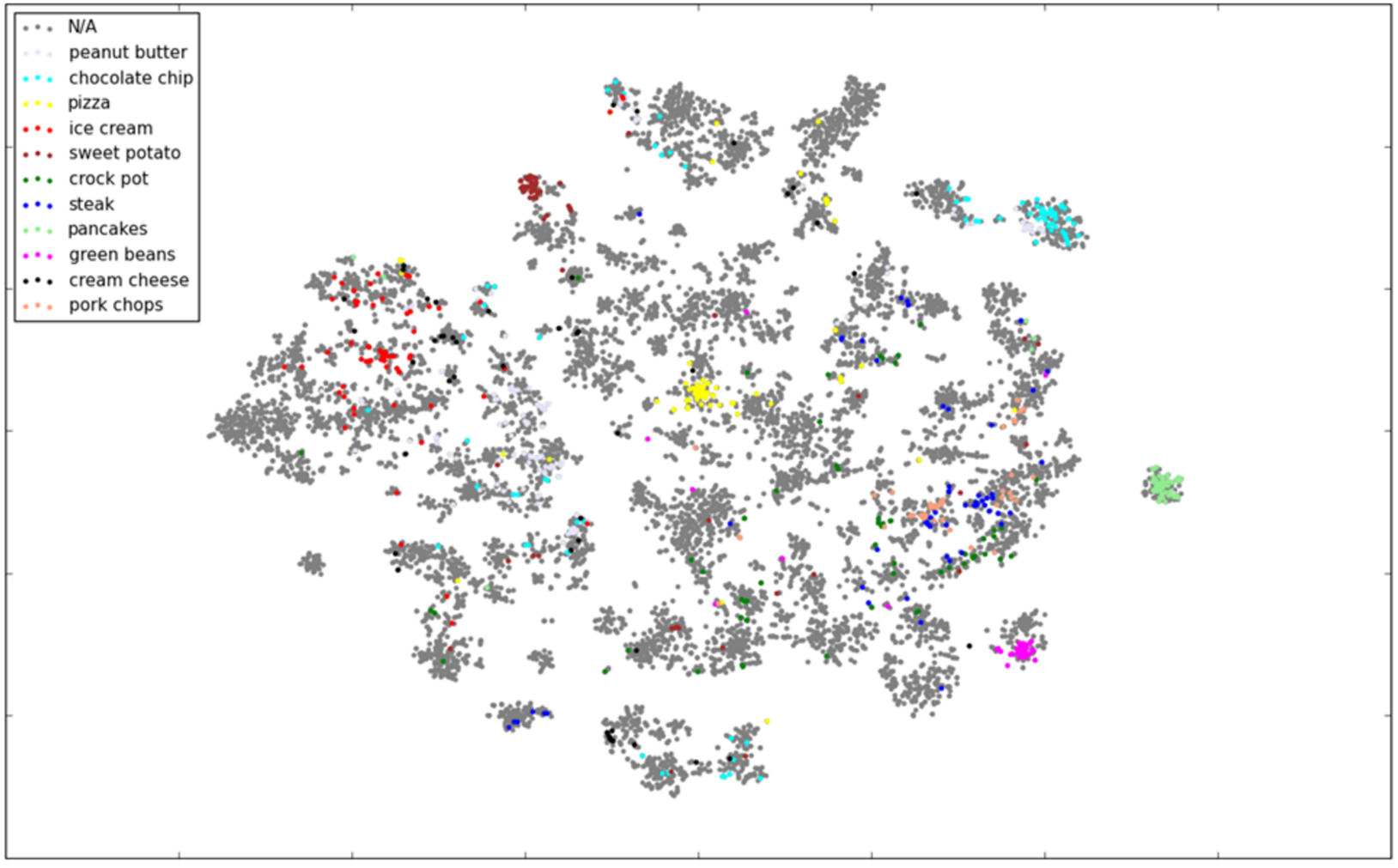


Fig. 3. **Embedding visualization using t-SNE**. Legend depicts the recipes that belong to the top 12 semantic categories used in our semantic regularization (see Section 5 for more details).

FSA

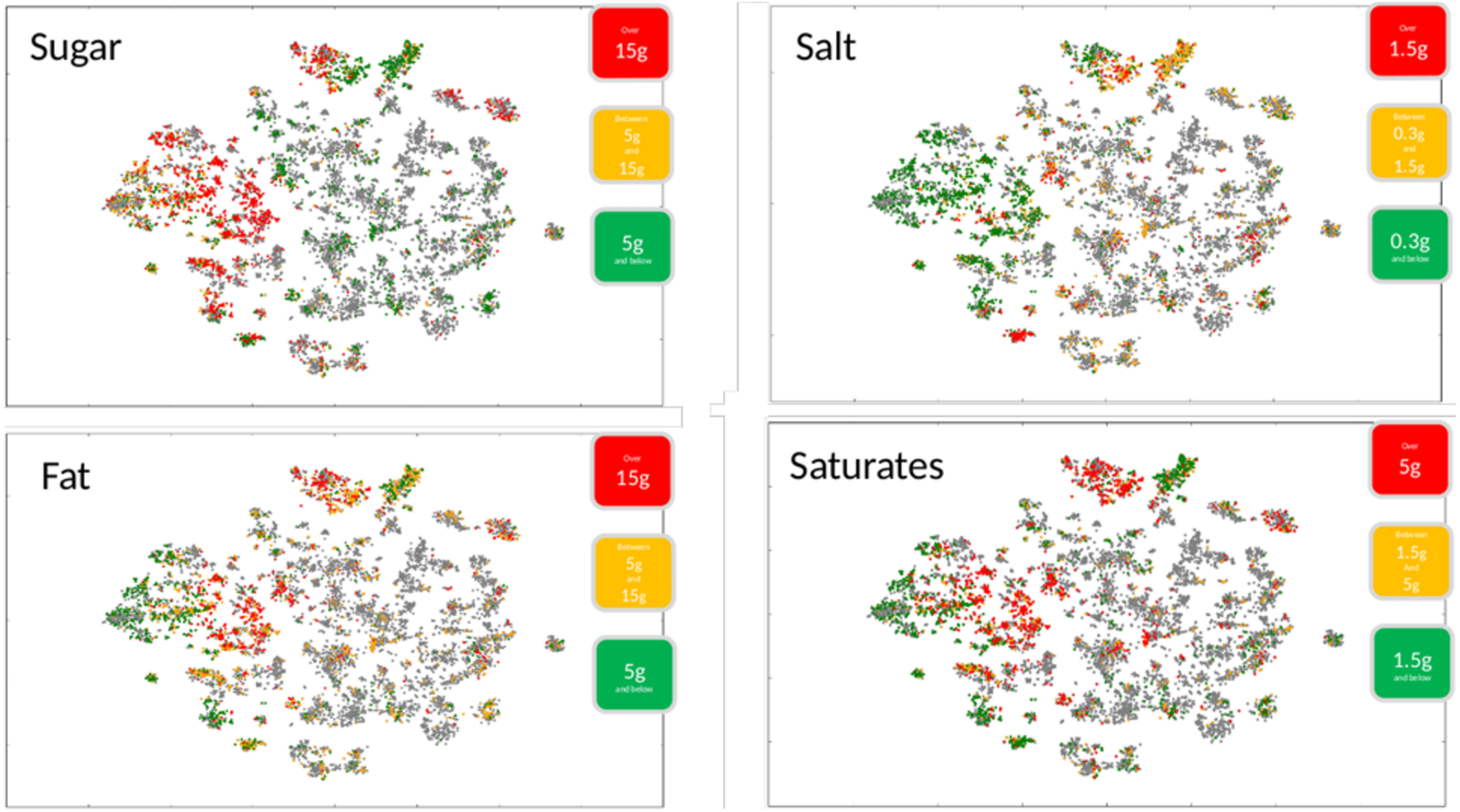


Fig. 4. **Healthiness within the embedding.** Recipe health is represented within the embedding visualization in terms of sugar, salt, saturates, and fat. We follow FSA traffic light system to determine how healthy a recipe is.

Analysis

Dataset Analysis

- Recipe1M+ contains minimal duplicates (~0.4%) and 20% of recipes have non-unique titles, differing by a median of 16 ingredients.
-
- 50% of recipes initially lacked images, but after data extension, only 2% of recipes are without images.
-
- Recipe1M+ consists of 16k unique ingredients; 4,000 account for 95% of occurrences.
-
- Recipe average: 9 ingredients, 10 instructions. Recipe images are heavily skewed with some popular recipes having significantly more images.

Image Matching Experiment

- During data extension, the number of recipes with images increased from 333K to over 1M.
- On average, Recipe1M+ has 13 images per recipe.

3 LEARNING EMBEDDINGS

1. Core Objective

- Learn a shared embedding space where:
- Recipes (text) and food images (visuals) map to similar points.
- Enables cross-modal tasks like image-to-recipe retrieval (Fig. 1).

2. Recipe Representation

- Ingredients:
  - Processed via bi-directional LSTM to handle unordered ingredient lists.
  - Ingredient names extracted with 99.5% accuracy using a custom NLP model.
- Cooking Instructions:
  - Encoded with "skip-instructions" (modified skip-thoughts model).
  - Captures context by predicting neighboring steps in the recipe.

3. Image Representation

- Uses pre-trained ResNet-50/VGG-16 (minus final classification layer).
- Extracts high-level visual features (e.g., texture, color, dish composition).

4. Joint Embedding Model

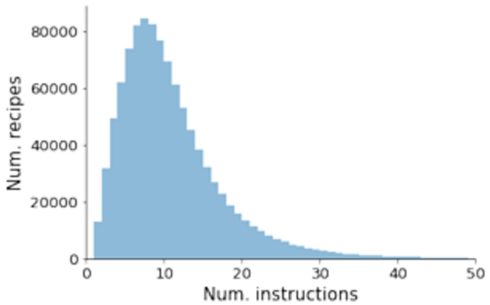
- Architecture:
  - Combines recipe (text) + image encoders into a unified space.
  - Trained with cosine similarity loss:
    - Maximizes similarity for matching recipe-image pairs.
    - Minimizes similarity for mismatched pairs (with margin $\alpha = 0.1$).
- Semantic Regularization:
  - Adds food-category classification (1,047 labels) to align embeddings semantically.

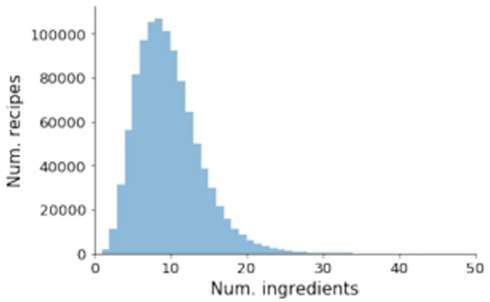
5. Training Strategy

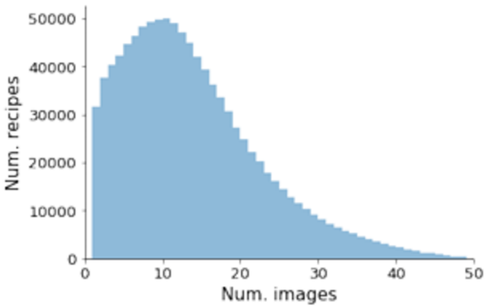
- Two-Phase Optimization:
  a. Freeze image network → Train recipe encoder.
  b. Freeze recipe encoder → Fine-tune image network.
- Tools: Implemented in PyTorch/Torch7, trained on 4 GPUs (3+ days).

6. Human Evaluation (via Amazon Mechanical Turk)

- Model performance rivals humans in recipe retrieval tasks (Table 5).
- Struggles with fine-grained cases (e.g., distinguishing sushi types).

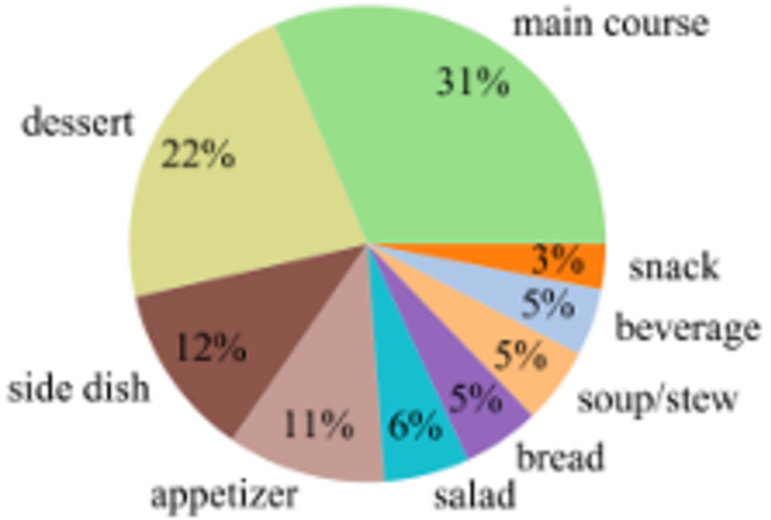

Num. recipes
Num. instructions
0
20000
40000
60000
80000
0
10
20
30
40
50


Num. recipes
Num. ingredients
0
20000
40000
60000
80000
100000
0
10
20
30
40
50


Num. recipes
Num. images
0
10000
20000
30000
40000
50000
0
10
20
30
40
50

de

sid


main course
31%
dessert
22%
3%
snack
5%
beverage
5%
soup/stew
5%
bread
6%
salad
11%
appetizer
12%
side dish

Representation of Recipes

Representation of Recipes

Ingredients:

- Each recipe contains a list of ingredients.
- Ingredient names are extracted from the text (e.g., "2 tbsp of olive oil" → "olive oil").
- Word2Vec is used to create a representation of each ingredient.
- Ingredient extraction is done using a Bi-directional LSTM with 99.5% accuracy.

Cooking Instructions:

- Instructions are lengthy (average of 208 words per recipe).
- A two-stage LSTM model is used to process instructions:
- First, each instruction is represented as a “skip-instructions” vector.
- Then, an LSTM encodes the sequence of vectors to represent all instructions.

Skip-Instructions and Embedding

Skip-Instructions:

- Based on Skip-Thoughts technique, which encodes a sentence and predicts previous/next sentences.
- Modifications include adding start and end-of-recipe tokens and using LSTM instead of GRU.
- Each instruction is encoded into a fixed-length vector.

Final Model:

- The instruction encoding is input into the joint embedding model for further processing.

JOINT NEURAL EMBEDDING

- Objective: Learn a shared representation for recipes and images.

Components:

- Ingredients Encoder: Uses a bidirectional LSTM to process unordered ingredient lists.
- Instructions Encoder: Uses a regular LSTM for cooking instructions represented as skip-instructions.

Recipe & Image Representation

Ingredients & Instructions:

- Ingredients are encoded as word2vec vectors.
- Cooking instructions are represented as skip-instructions vectors.

Joint Space:

- The recipe and image representations are concatenated and projected into a joint space.
- Image Encoder: Transforms images into fixed-length vectors for the shared space.

Training & Objective

- Goal: Maximize the cosine similarity between matching recipe-image pairs.
- Loss Function:

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

- End-to-End Training: The model is trained using positive and negative recipe-image pairs

5 SEMANTIC REGULARIZATION

1. Core Idea

- Goal: Improve alignment between recipe and image embeddings by forcing both modalities to solve the same food-category classification task.
- Mechanism: Shared classification weights for images/recipes → embeddings must encode similar semantic features.

2. Semantic Categories

- Sources:
  - Food-101 labels (13% coverage).
  - Recipe title bigrams (e.g., "chicken salad", "grilled vegetable").
- Final Set: 1,047 categories covering 50% of recipes (remainder marked as "background").
- Filtering: Removed non-discriminative phrases (e.g., "super easy").

3. Implementation

- Architecture: Added a shared fully connected layer for classification (Fig. 6).
- Loss Function: Combined with joint embedding loss (weighted by $\lambda = 0.02$).
- Effect:
  - Improves retrieval accuracy (Table 3).
  - Enables "semantic arithmetic" in embedding space (Figs. 10–11).

4. Limitations

- Partial Coverage: Only 50% of recipes have category labels.
- Ambiguity: Some titles match multiple categories (resolved by picking the most frequent one).

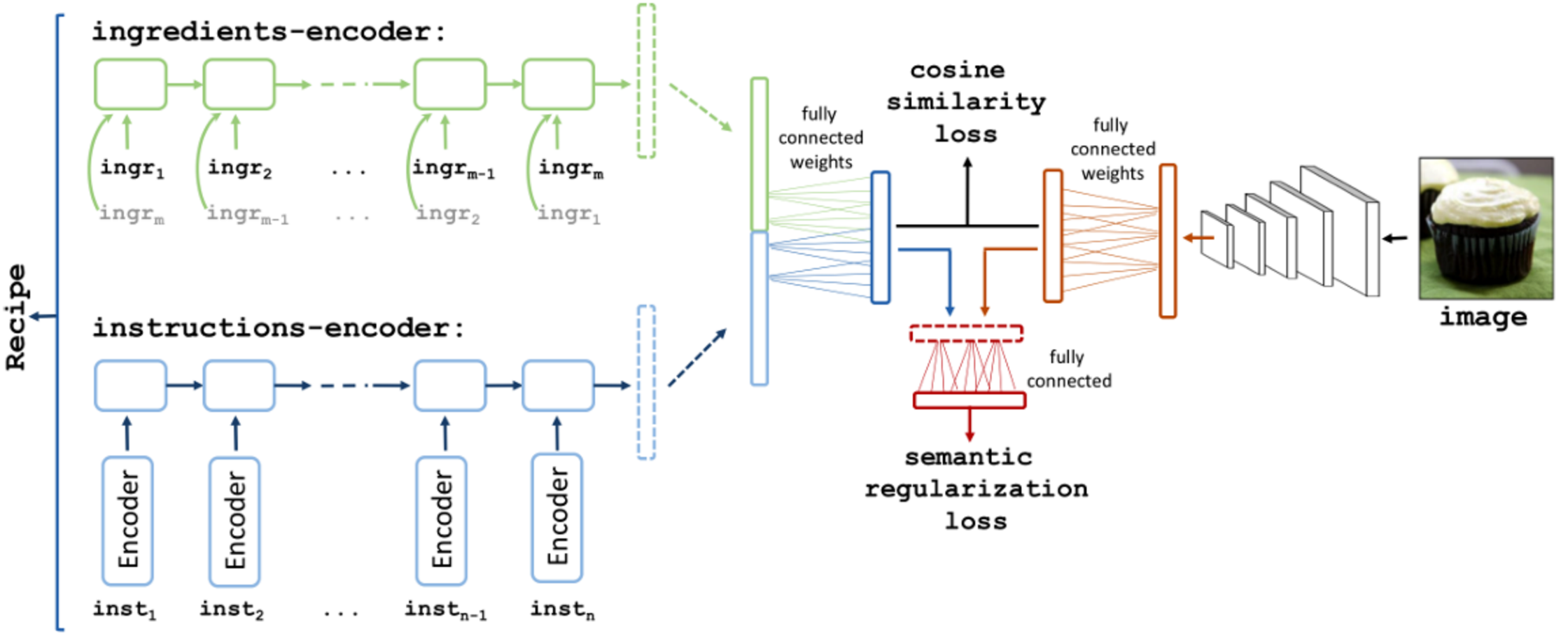


Fig. 6. **Joint neural embedding model with semantic regularization.** Our model learns a joint embedding space for food images and cooking recipes.

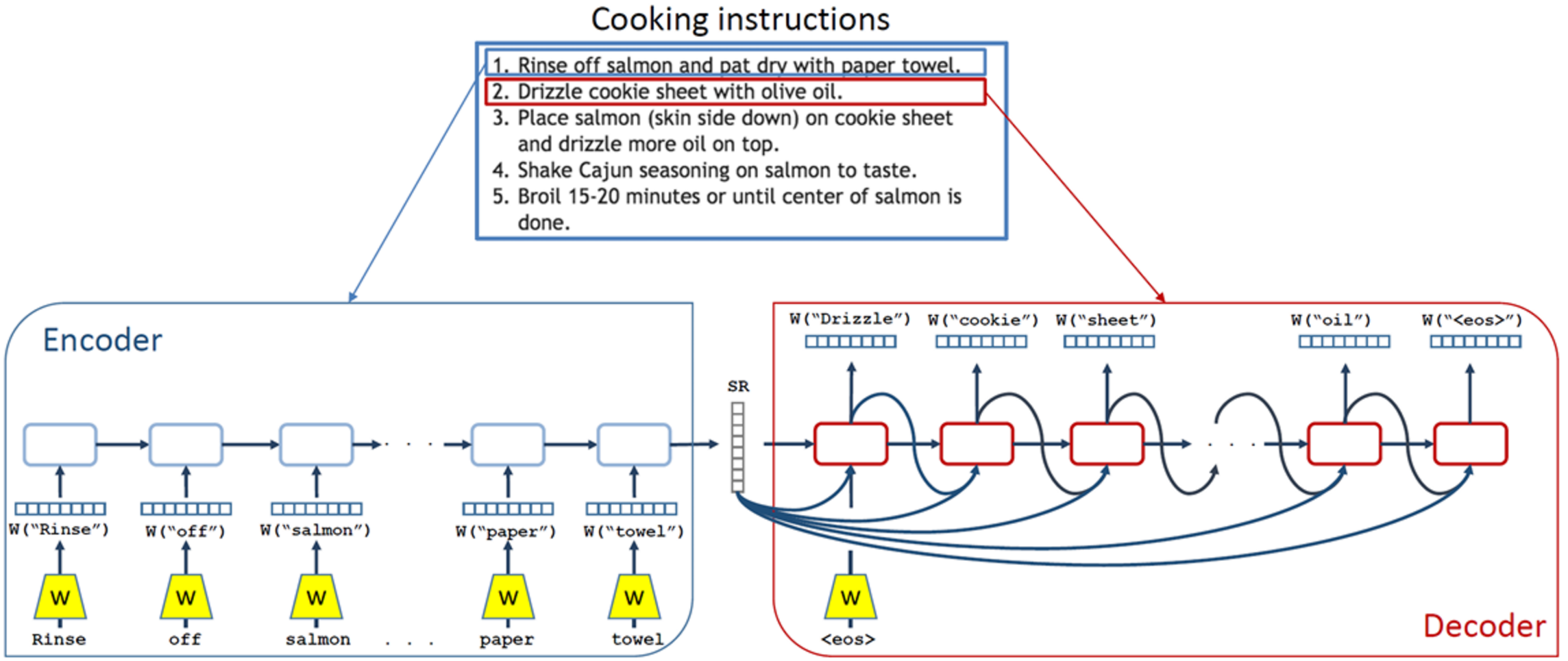


Fig. 7. **Skip-instructions model.** During training the encoder learns to predict the next instruction.

Classification & Semantic Regularization

Class Probabilities:

- **Calculate from recipe (r) and image (v) embeddings:**
- $p_r = W^c\phi^r$ and $p_v = W^c\phi^v$ followed by a softmax activation.

Semantic Regularization:

- **Shared weights (Wc) ensure semantic alignment between recipe and image embeddings.**

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

Optimization

Two-Stage Optimization:

- **Train recipe network first, with frozen image network (pre-trained on ImageNet).**
- **Then, train image network with frozen recipe network.**
- **Reason: Prevents instability and divergence during joint training.**

Final Optimization:

- **Fine-tune both networks together, with minimal changes.**
- **Validation: Use Median Rank (MedR) for stable performance.**

6.1Implementation Details

1. Task Definition

Goal: Retrieve the correct recipe given a food image (im2recipe) or vice versa (recipe2im).

Evaluation:

Metrics: Median Rank (MedR, lower is better) & Recall@K (R@K, higher is better).

Test Protocol: 1,000 random recipe-image pairs, repeated 10x for robustness.

2. Baseline Comparison (CCA Models)

Weak Baselines:

CCA + GoogleNews word2vec + skip-thoughts: MedR = 25.2.

CCA + ingredient word2vec + skip-instructions: Better but still limited (MedR = 15.7).

Limitations:

Uses averaged features (loses recipe structure).

Image features (ResNet-50) are generic (not food-optimized).

3. Our Model’s Performance

- Outperforms CCA:
  - MedR = 5.2 (vs. 15.7 for best baseline).
  - R@1 = 24%, R@5 = 51%, R@10 = 65%.
- Semantic Regularization Boost:
  - Reduces MedR from 7.2 → 5.2 (Table 3).
  - Aligns embeddings using 1,047 food categories.

4. Qualitative Results

- Successes:
  - Correctly retrieves recipes for visually distinct dishes (e.g., salads, baked goods).
- Failures:
  - Misses invisible ingredients (e.g., beef in lasagna).
  - Confuses visually similar ingredients (e.g., shrimp vs. salmon).

5. Comparison to Subsequent Work

- Attention Models: Minor improvements (MedR = 4.6).
- Double-Triplet Learning: Significant gains (MedR = 1.0, R@1 = 40%).

Component Analysis (Ablation Study)

- Visual Backbones Matter:
  - ResNet-50 (MedR=7.9) outperforms VGG-16 (MedR=15.3) due to better feature extraction.
  - Fine-tuning ResNet further reduces MedR to 7.2.
- Semantic Regularization Boost:
  - Adds ~2-point MedR improvement (e.g., 7.2 → 5.2).

Recipe1M vs. Recipe1M+

- Recipe1M+ Wins:
  - MedR drops 5 points (13.6 → 8.6 on im2recipe).
  - R@10 jumps 10% (46% → 54%).
- Generalization Test (Food-101):
  - Recipe1M+ model achieves MedR=2.6 (vs. 4.75 for Recipe1M) in recipe2im.
  - Confirmed no data leakage between datasets.

- Why ResNet-50?
- Residual connections enable deeper networks without gradient issues, capturing finer visual details (e.g., grill marks vs. raw ingredients).
- Semantic Regularization
- Forces embeddings to align with 1,047 food categories, improving cross-modal consistency.
- Failure Cases:
  - Occluded ingredients (e.g., beef in lasagna).
  - Subtle differences (shrimp vs. salmon, sushi fillings).

Metric	Recipe1M	Recipe1M+	Improvement
MedR (im2recipe)	13.6	8.6	↓ 37%
R@10 (recipe2im)	67.5%	76.3%	↑ 13%
Food-101 MedR	4.75	2.60	↓ 45%

Key Takeaway: Scaling data with web-sourced images (Recipe1M+) consistently boosts performance **without overfitting**.

Human vs. AI Performance

- Human Baseline (AMT Workers):
  - Easy (all recipes): 81.6% accuracy.
  - Hard (same dish name): ~53.2% accuracy.
- Our Model:
  - Outperforms humans in easy/medium tasks (84.8% vs. 81.6%).
  - Struggles with fine-grained cases (e.g., sushi types, smoothies).

TABLE 3
Im2recipe retrieval comparisons on Recipe1M. Median ranks and recall rate at top K are reported for baselines and our method. Note that the joint neural embedding models consistently outperform all the baseline methods.

	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
attention + SR. [18]	4.6	0.26	0.54	0.67	4.6	0.26	0.54	0.67
AdaMine [19]	1.0	0.40	0.69	0.77	1.0	0.40	0.68	0.79

	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

TABLE 5

Comparison with human performance on im2recipe task on Recipe1M. 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	**81.6 ± 8.9**	52.0	70.0	34.0	58.0	56.0	**54.0 ± 13.0**	54.0	48.0	58.0	52.0	48.0	46.0	54.0	58.0	**52.2 ± 04.6**
joint-emb. only	**83.6 ± 3.0**	76.0	68.0	38.0	24.0	62.0	**53.6 ± 21.8**	58.0	58.0	58.0	64.0	38.0	58.0	62.0	42.0	**54.8 ± 09.4**
joint-emb.+semantic	**84.8 ± 2.7**	74.0	82.0	56.0	30.0	62.0	**60.8 ± 20.0**	52.0	60.0	62.0	68.0	42.0	68.0	62.0	44.0	**57.2 ± 10.1**

TABLE 6

Comparison between models trained on Recipe1M vs. Recipe1M+. 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. In this table, Recipe1M refers to the *intersection* dataset.

	Recipe1M test set				Recipe1M+ test set			
	im2recipe							
	medR	R@1	R@5	R@10	medR	R@1	R@5	R@10

7 CONCLUSION

This work presents Recipe1M+, the largest structured recipe dataset (1M recipes + 13M images), enabling breakthroughs in cross-modal food AI. Our neural embedding model with semantic regularization achieves human-competitive recipe retrieval accuracy while demonstrating remarkable capabilities like "recipe arithmetic" for culinary creativity. The expanded dataset maintains quality despite web-scale collection, showing strong generalization on benchmarks like Food-101. Beyond cooking, our framework opens new possibilities for understanding procedural knowledge in domains from furniture assembly to industrial processes, particularly for predicting visual outcomes of action sequences. This research lays the foundation for AI systems that blend comprehension and creativity in multimodal instruction following.