METRIC LEARNING FOR CROSSMODAL ALIGNMENT

Wednesday 20th December, 2017

Micael Carvalho

PhD candidate in Computer Science / Deep Learning

Université Pierre et Marie Curie

Recipe1M dataset

Proposed by Salvador et al. at CVPR 2017

Dataset composed of pairs image-recipe

Ingredients

pasta
ground beef
taco seasoning
water
cream cheese
cheese

Instructions

1. Preheat oven to 350F.
2. Boil pasta until just cooked.
3. Brown ground beef and then drain.
4. Add taco seasoning and water to meat and simmer for 5 minutes.

...

5. Put half of the shredded cheese over pasta, then cover with hamburger meat and mix gentle.
6. Sprinkle remaining cheese over the top.
7. Cook in the oven uncovered for 15-20 minutes.

Ingredients

butter
olive oil
sweet onions
portabella mushrooms
celery
carrot
garlic cloves
...

Instructions

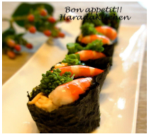
1. Melt 1 tablespoon butter with 1/2 tablespoon olive oil in saucepan over medium heat.
2. Add onions and saute, stirring every few minutes, until they are caramelized, about 15-20 minutes.

...

3. (If soup is too thick, thin with a little more hot broth).
4. Season to suit your taste with salt and freshly-cracked black pepper.
5. Serve in deep bowls, garnished with a sprinkle of minced, fresh parsley.

Task 1: Image to Recipe retrieval

Query Image



Retrieved Recipe

Ingredients

sushi rice
salmon
avocado
cream cheese
nori

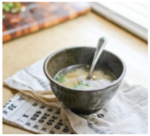
Instructions

1. Make 2 bowls of sushi rice.
2. Slice the salmon into 24 ultra-thin slices, and cut the avocado and cream cheese into long, thin strips.
3. Place a small bowl-worth of sushi rice on plastic wrap and spread it out to the size of a nori sheet.

...

4. Cut the rolls while wiping the knife with a wet cloth between each cut.
5. Shown in the photo on the left is avocado, and to the right is mini cucumber.

Query Image



Retrieved Recipe

Ingredients

butter
olive oil
sweet onions
portabella
mushrooms
celery
carrot
garlic cloves
...

Instructions

1. Melt 1 tablespoon butter with 1/2 tablespoon olive oil in saucepan over medium heat.
2. Add onions and saute, stirring every few minutes, until they are caramelized, about 15-20 minutes.

...

3. (If soup is too thick, thin with a little more hot broth).
4. Season to suit your taste with salt and freshly-cracked black pepper.
5. Serve in deep bowls, garnished with a sprinkle of minced, fresh parsley.

Task 2: Recipe to Image retrieval

Query Recipe

Ingredients

butter
garlic cloves
all - purpose flour
kosher salt
milk
chicken broth
mozzarella cheese
parmesan cheese
onion
...

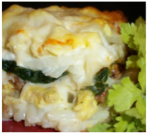
Instructions

1. Heat butter in 2 qt saucepan over low heat until melted
2. Add garlic.
3. Stir in flour and salt.
4. Cook, stirring constantly until bubbly.
5. Remove from heat and stir in milk and broth.

...

6. Cook uncovered at 350F 20-30 minutes until nice and bubbly.
7. Let stand 10 minutes before cutting.

→ **Retrieved Image**



Query Recipe

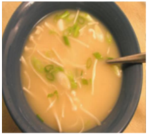
Ingredients

dashi stock
hot water
miso
firm tofu
green onion

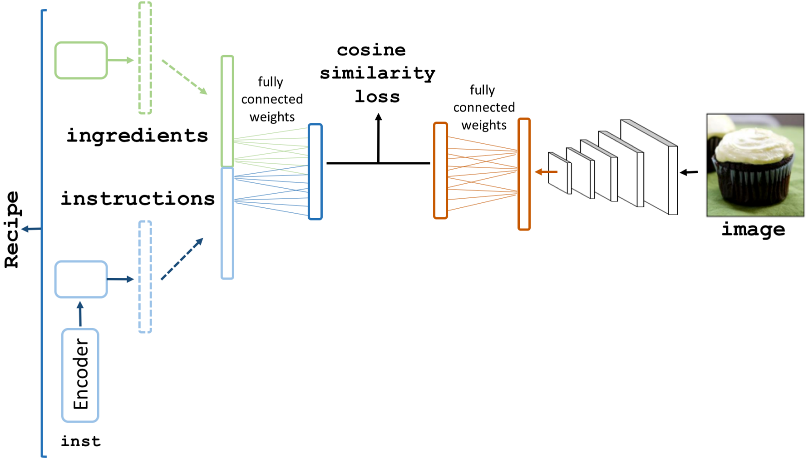
Instructions

1. Transfer dashi to a small soup pot over medium-low heat.
2. Meanwhile, stir together hot water and miso until miso is dissolved.
3. Pour watery miso mixture into the pot.
4. Add cubed tofu.
5. Bring the pot to a simmer.
6. To serve, sprinkle sliced green onions and a pinch of katsuobushi on top.

→ **Retrieved Image**



Architecture — learning to align with a similarity loss



Metric learning bases

Challenge – part 1

Challenge: What is the distance between... ?

1275

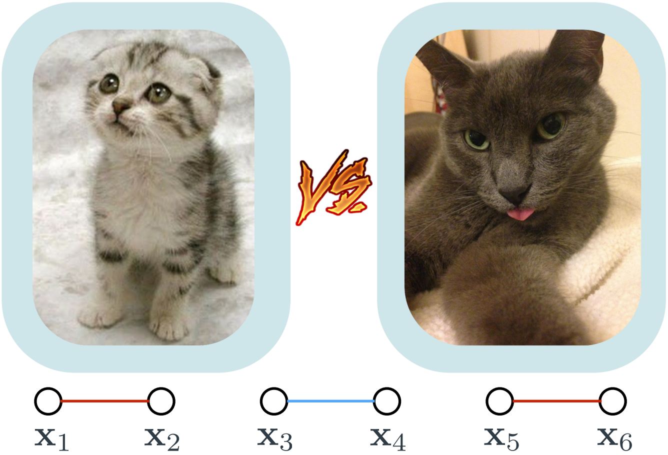
VS

1317

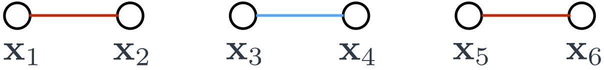
It's *easier* to find distances between numbers than between images

Challenge – part 1

Challenge: What is the distance between... ?



Pairwise/contrastive



Trained on the paired data $\{(\mathbf{x}_i, \mathbf{x}_j, y_{i,j})\}$, with the cost function

$$y_{i,j} D_{i,j}^2 + (1 - y_{i,j})[\alpha - D_{i,j}]_+^2$$

$$y_{i,j} \in \{0, 1\}, \qquad D_{i,j} = ||f(\mathbf{x}_i) - f(\mathbf{x}_j)||_2, \qquad [\cdot]_+ = max(0, \cdot)$$

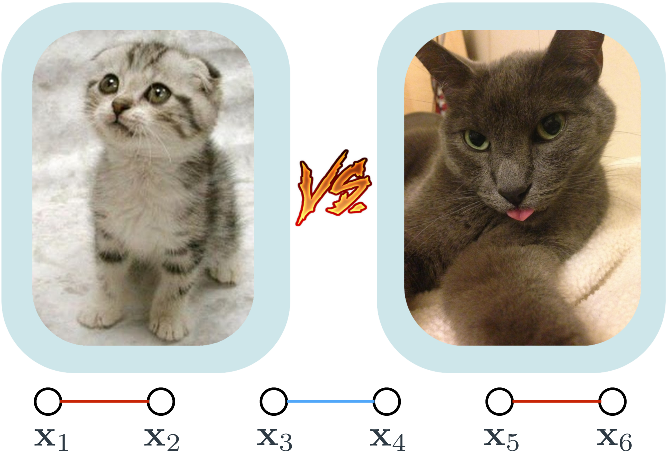
[☺] Approaches positive pairs and distances negative pairs by α;

[☹] Forces positive examples to have distance 0;

[☹] (...) Other problems, lets just agree it's *not optimal*.

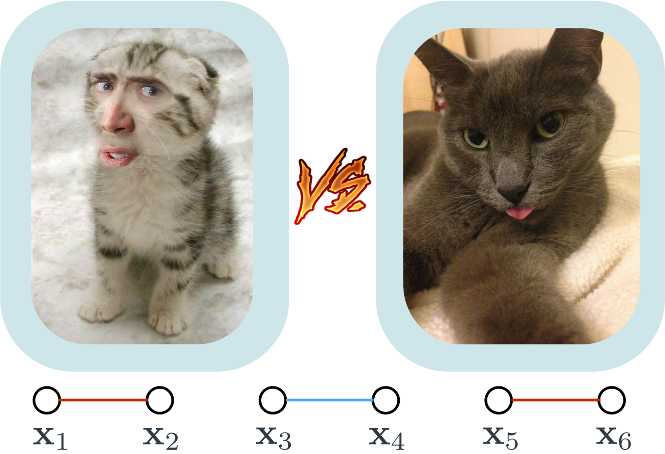
Challenge – part 2

Challenge: What is the distance between... ?



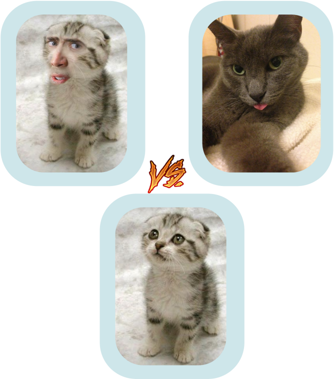
Challenge – part 2

Challenge: What is the distance between... ?



Challenge – part 2

Challenge: What is the distance between... ?



$\mathbf{x}_1$
$\mathbf{x}_2$
$\mathbf{x}_3$
$\mathbf{x}_4$
$\mathbf{x}_5$
$\mathbf{x}_6$

Triplet

$\mathbf{x}_1$
$\mathbf{x}_2$
$\mathbf{x}_3$
$\mathbf{x}_4$
$\mathbf{x}_5$
$\mathbf{x}_6$

Trained on $\{(\mathsf{x}_a, \mathsf{x}_p, \mathsf{x}_n)\}$, with the cost function

$$[D^2_{ia,ip} - D^2_{ia,in} + \alpha]_+$$

[☺] Approaches positive examples and distances negative examples;

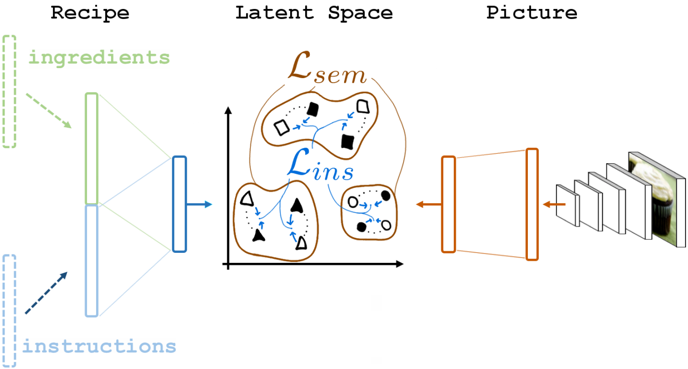
[☺] Pushes away *the negative example* and closer *the positive example* if the negative one is inside $D^2_{ia,ip} + \alpha$;

Problem-driven proposal

Micael Carvalho⋆, Rémi Cadène⋆,

David Picard, Nicolas Thome, and Matthieu Cord

Semantic loss



Semantic-based loss $\mathcal{L}_{sem}$ added to organize the feature space. Its triplets are constructed with respect to the class of each sample, instead of their instance information.

Total loss

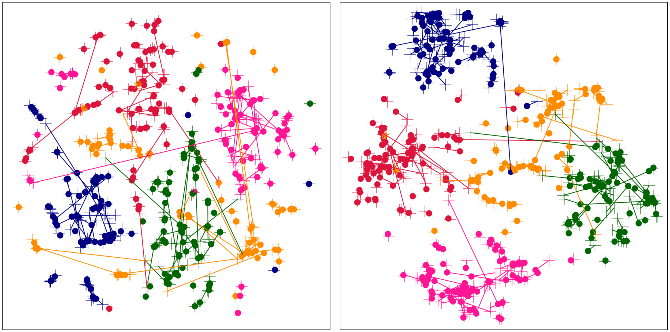
$$\mathcal{L}_{total} = \mathcal{L}_{ins} + \lambda \mathcal{L}_{sem}$$

$\mathcal{L}_{ins}$ and $\mathcal{L}_{sem}$ are triplet-based losses:

$$\ell_{tri}(\theta, x_q, x_p, x_n) = [d(x_q, x_p) + \alpha - d(x_q, x_n)]_+$$

The problem is symmetrical in modalities:
$(\mathbb{Q}, \mathbb{P}_q, \mathbb{N}_q) \in (\mathcal{V}, \mathcal{T}, \mathcal{T})$ or $(\mathbb{Q}, \mathbb{P}_q, \mathbb{N}_q) \in (\mathcal{T}, \mathcal{V}, \mathcal{V})$.

Qualitative studies - t-SNE

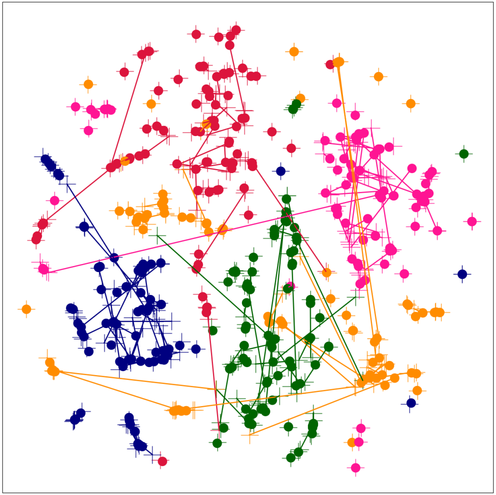


(a) Triplet (b) Ours

Figure 1: **t-SNE visualization.** Image (resp. Recipe) points are denoted with the + (resp. •) symbol. Matching pairs are connected with a trace. Blue points are associated to the cupcake class, orange to hamburger, pink to green beans, green to pork chops, and red to pizza.

Sampling strategies

Organizing the space



How to choose which triplets should be used?

Average triplet mining

To adjust the parameters θ of a network with SGD $\theta(t+1) = \theta(t) - \eta\delta$, the update term δ can be calculated as follows:

$$\delta_{avg} = \sum_{x_q \in \mathbb{Q}} \left(\sum_{x_p \in \mathbb{P}^{\mathbb{B}}_{q,v}} \sum_{x_n \in \mathbb{N}^{\mathbb{B}}_{q,v}} \frac{\nabla \ell_{tri}(\theta, x_q, x_p, x_n)}{|\mathbb{Q}| \cdot |\mathbb{N}^{\mathbb{B}}_{q,v}| \cdot |\mathbb{P}^{\mathbb{B}}_{q,v}|} + \sum_{x_p \in \mathbb{P}^{\mathbb{B}}_{q,s}} \sum_{x_n \in \mathbb{N}^{\mathbb{B}}_{q,s}} \lambda \frac{\nabla \ell_{sem}(\theta, x_q, x_p, x_n)}{|\mathbb{Q}| \cdot |\mathbb{N}^{\mathbb{B}}_{q,s}| \cdot |\mathbb{P}^{\mathbb{B}}_{q,s}|} \right)$$

where $\mathbb{Q}$ is the ensemble of query items, and $\mathbb{P}^{\mathbb{B}}_q$ and $\mathbb{N}^{\mathbb{B}}_q$ are their crossmodal ensemble of positive and negative matches, respectively

Hard(est) triplet mining

$$\delta_{max} = \sum_{x_q \in \mathbb{Q}} \left(\sum_{x_p \in \mathbb{P}^{\mathbb{B}}_{q,v}} \max_{x_n \in \mathbb{N}^{\mathbb{B}}_{q,v}} \frac{\nabla \ell_{tri}(\theta, x_q, x_p, x_n)}{|\mathbb{Q}| \cdot |\mathbb{P}^{\mathbb{B}}_{q,v}|} + \sum_{x_p \in \mathbb{P}^{\mathbb{B}}_{q,s}} \max_{x_n \in \mathbb{N}^{\mathbb{B}}_{q,s}} \lambda \frac{\nabla \ell_{sem}(\theta, x_q, x_p, x_n)}{|\mathbb{Q}| \cdot |\mathbb{P}^{\mathbb{B}}_{q,s}|} \right)$$

Adaptive triplet mining (ours)

$$\delta_{adm} = \sum_{x_q \in \mathbb{Q}} \left(\sum_{x_p \in \mathbb{P}^{\mathbb{B}}_{q,v}} \sum_{x_n \in \mathbb{N}^{\mathbb{B}}_{q,v}} \frac{\nabla \ell_{tri}(\theta, x_q, x_p, x_n)}{\beta'_r} + \sum_{x_p \in \mathbb{P}^{\mathbb{B}}_{q,s}} \sum_{x_n \in \mathbb{N}^{\mathbb{B}}_{q,s}} \lambda \frac{\nabla \ell_{sem}(\theta, x_q, x_p, x_n)}{\beta'_s} \right)$$

with β'_r and β'_s compensating for uninformative triplets:

$$\beta'_r = \sum_{x_q \in \mathbb{Q}} \sum_{x_p \in \mathbb{P}^{\mathbb{B}}_{q,v}} \sum_{x_n \in \mathbb{N}^{\mathbb{B}}_{q,v}} \mathbb{1}_{\ell_{tri} \neq 0}$$

$$\beta'_s = \sum_{x_q \in \mathbb{Q}} \sum_{x_p \in \mathbb{P}^{\mathbb{B}}_{q,s}} \sum_{x_n \in \mathbb{N}^{\mathbb{B}}_{q,s}} \mathbb{1}_{\ell_{sem} \neq 0}$$

Experiments

State-of-the-art comparison

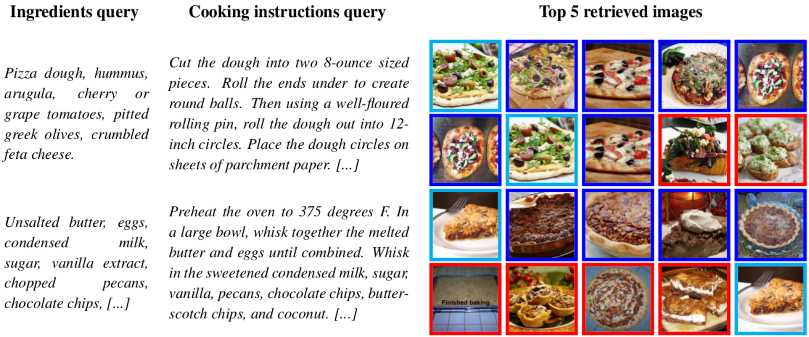
	im2recipe @ 1k			recipe2im @ 1k		
	MedR	R@1	R@10	MedR	R@1	R@10
CCA [1]	15.7	14.0	43.0	24.8	9.0	35.0
PWC [1]	5.2	24.0	65.0	5.1	25.0	65.0
PWC++ (pairwise, ours)	3.3 ± 0.4	25.8 ± 1.6	67.1 ± 1.4	3.5 ± 0.5	24.8 ± 1.1	67.1 ± 1.2
Ours	**1.0** ± 0.1	**39.8** ± 1.8	**77.4** ± 1.1	**1.0** ± 0.1	**40.2** ± 1.6	**78.7** ± 1.3

	im2recipe @ 10k			recipe2im @ 10k		
	MedR	R@1	R@10	MedR	R@1	R@10
PWC [1]	41.9	-	-	39.2	-	-
PWC++ (pairwise, ours)	34.6 ± 1.0	7.6 ± 0.2	30.3 ± 0.4	35.0 ± 0.9	6.8 ± 0.2	28.8 ± 0.3
Ours	**13.2** ± 0.4	**14.9** ± 0.3	**45.2** ± 0.2	**12.2** ± 0.4	**14.8** ± 0.3	**46.1** ± 0.3

Table 1: **State-of-the-art comparison.** MedR means Median Rank (lower is better). R@K means Recall at K (between 0% and 100%, higher is better). The mean and std values over 10 (resp. 5) bags of 1k (resp. 10k) pairs each are reported for the top (resp. bottom) table.

[1] Salvador et al., "Learning Cross-modal Embeddings for Cooking Recipes and Food Images," CVPR'17.

Qualitative studies - Visualization



Ingredients query	Cooking instructions query	Top 5 retrieved images
Pizza dough, hummus, arugula, cherry or grape tomatoes, pitted greek olives, crumbled feta cheese.	*Cut the dough into two 8-ounce sized pieces. Roll the ends under to create round balls. Then using a well-floured rolling pin, roll the dough out into 12-inch circles. Place the dough circles on sheets of parchment paper. [...]*	
Unsalted butter, eggs, condensed milk, sugar, vanilla extract, chopped pecans, chocolate chips, [...]	*Preheat the oven to 375 degrees F. In a large bowl, whisk together the melted butter and eggs until combined. Whisk in the sweetened condensed milk, sugar, vanilla, pecans, chocolate chips, butter-scotch chips, and coconut. [...]*	

Figure 2: **Recipe-to-images visualization.** For each recipe, we have the top row, indicating the top 5 images retrieved by our model for a given recipe query, and the bottom row, indicating the top 5 images by the triplet loss for the same recipe. In cyan, the matching image. In blue, images belonging to the same class than the recipe. In red, images belonging to a different class.

Qualitative studies - Visualization

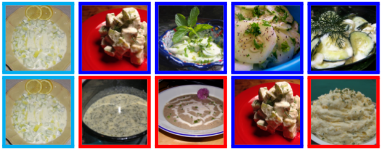
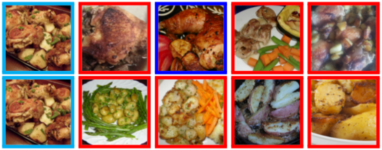
Ingredients query	Cooking instructions query	Top 5 retrieved images
Yogurt, cucumber, salt, garlic clove, fresh mint.	*Stir yogurt until smooth. Add cucumber, salt, and garlic. Garnish with mint. Normally eaten with pita bread. Enjoy!*	
Olive oil, balsamic vinegar, thyme, lemons, chicken drumsticks with bones and skin, garlic, potatoes, parsley.	*Whisk together oil, mustard, vinegar, and herbs. Season to taste with a bit of salt and pepper and a large pinch or two of brown sugar. Place chicken in a non-metal dish and pour marinade on top to coat. [...]*	

Figure 3: **Recipe-to-images visualization.** For each recipe, we have the top row, indicating the top 5 images retrieved by our model for a given recipe query, and the bottom row, indicating the top 5 images by the triplet loss for the same recipe. In cyan, the matching image. In blue, images belonging to the same class than the recipe. In red, images belonging to a different class.

Thank you

Micael Carvalho

micael.carvalho@lip6.fr

micaelcarvalho.com

Laboratoire d'Informatique de Paris 6; Sorbonne Universités, UPMC, Paris, France