NEURAL NETWORK TECHNIQUES FOR PRODUCING INGREDIENT REPRESENTATIONS IN VECTOR SPACE

YOUYANG GU

12/10/15

CONTRIBUTIONS

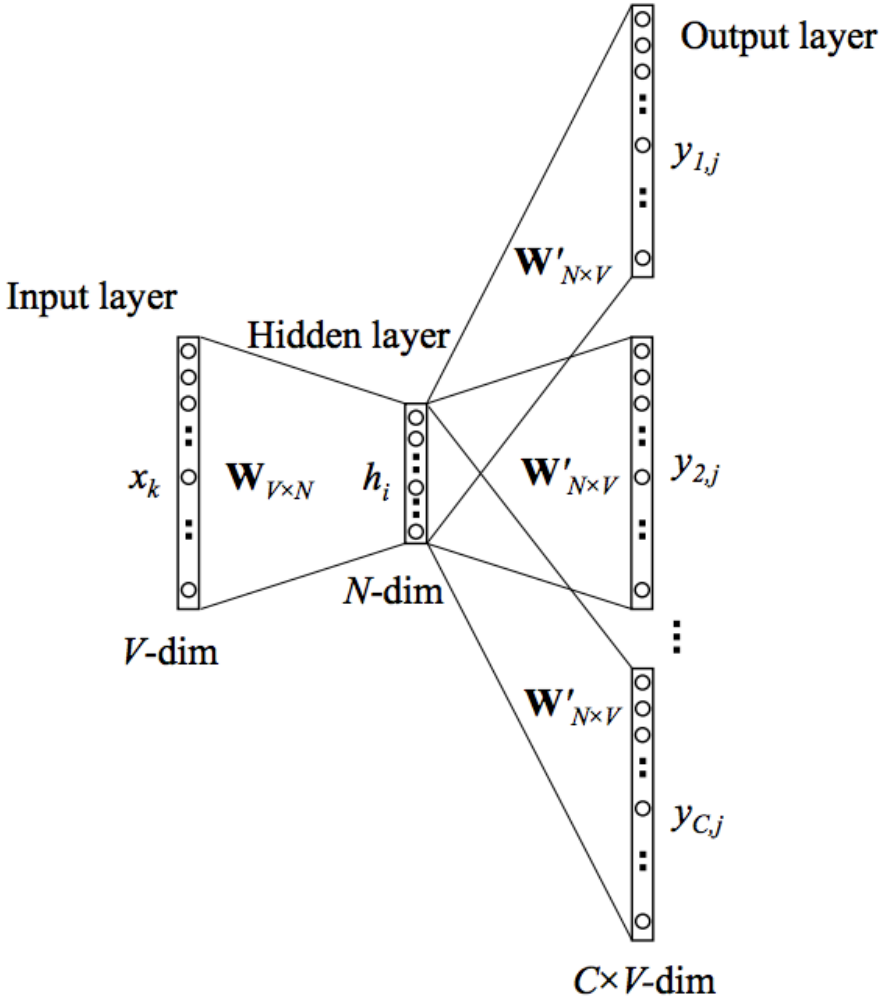
- Generate embeddings for ingredients using the skip-gram model from NLP. Embeddings show high proximity between ingredients that have similar semantic meanings.
- Given list of ingredients, correctly predict whether it is a valid combination 90%+ of the time.
- Given list of ingredients, correctly predict category 80%+ of the time.
- For newly unseen ingredients, use property-level information of ingredient to generate a map to an embedding → perform predictions.
- Pave the way for analysis of adulterants/illegal ingredients that normally would not show up in the ingredient list.

DATA

- 150,000 products, 100,000 unique ingredients
- 3000 ingredients occur in 50+ products
- Top 10 ingredients:
  - *Salt, water, sugar, citric acid, riboflavin, folic acid, thiamine mononitrate, niacin, natural flavor, soy lecithin*

INGREDIENTS: ENRICHED FLOUR (WHEAT FLOUR, NIACIN, REDUCED IRON, THIAMIN MONONITRATE [VITAMIN B_1], RIBOFLAVIN [VITAMIN B_2], FOLIC ACID), CORN SYRUP, SUGAR, SOYBEAN AND PALM OIL (WITH TBHQ FOR FRESHNESS), CORN SYRUP SOLIDS, DEXTROSE, HIGH FRUCTOSE CORN SYRUP, FRUCTOSE, GLYCERIN, CONTAINS 2% OR LESS OF COCOA (PROCESSED WITH ALKALI), POLYDEXTROSE, MODIFIED CORN STARCH, SALT, DRIED CREAM, CALCIUM CARBONATE, CORNSTARCH, LEAVENING (BAKING SODA, SODIUM ACID PYROPHOSPHATE, MONOCALCIUM PHOSPHATE, CALCIUM SULFATE), DISTILLED MONOGLYCERIDES, HYDROGENATED PALM KERNEL OIL, SODIUM STEAROYL LACTYLATE, GELATIN, COLOR ADDED, SOY LECITHIN, DATEM, NATURAL AND ARTIFICIAL FLAVOR, VANILLA EXTRACT, CARNAUBA WAX, XANTHAN GUM, VITAMIN A PALMITATE, YELLOW #5 LAKE, RED #40 LAKE, CARAMEL COLOR, NIACINAMIDE, BLUE #2 LAKE, REDUCED IRON, YELLOW #6 LAKE, PYRIDOXINE HYDROCHLORIDE (VITAMIN B_6), RIBOFLAVIN (VITAMIN B_2), THIAMIN HYDROCHLORIDE (VITAMIN B_1), CITRIC ACID, FOLIC ACID, RED #40, YELLOW #5, YELLOW #6, BLUE #2, BLUE #1.

SKIP-INGREDIENT MODEL



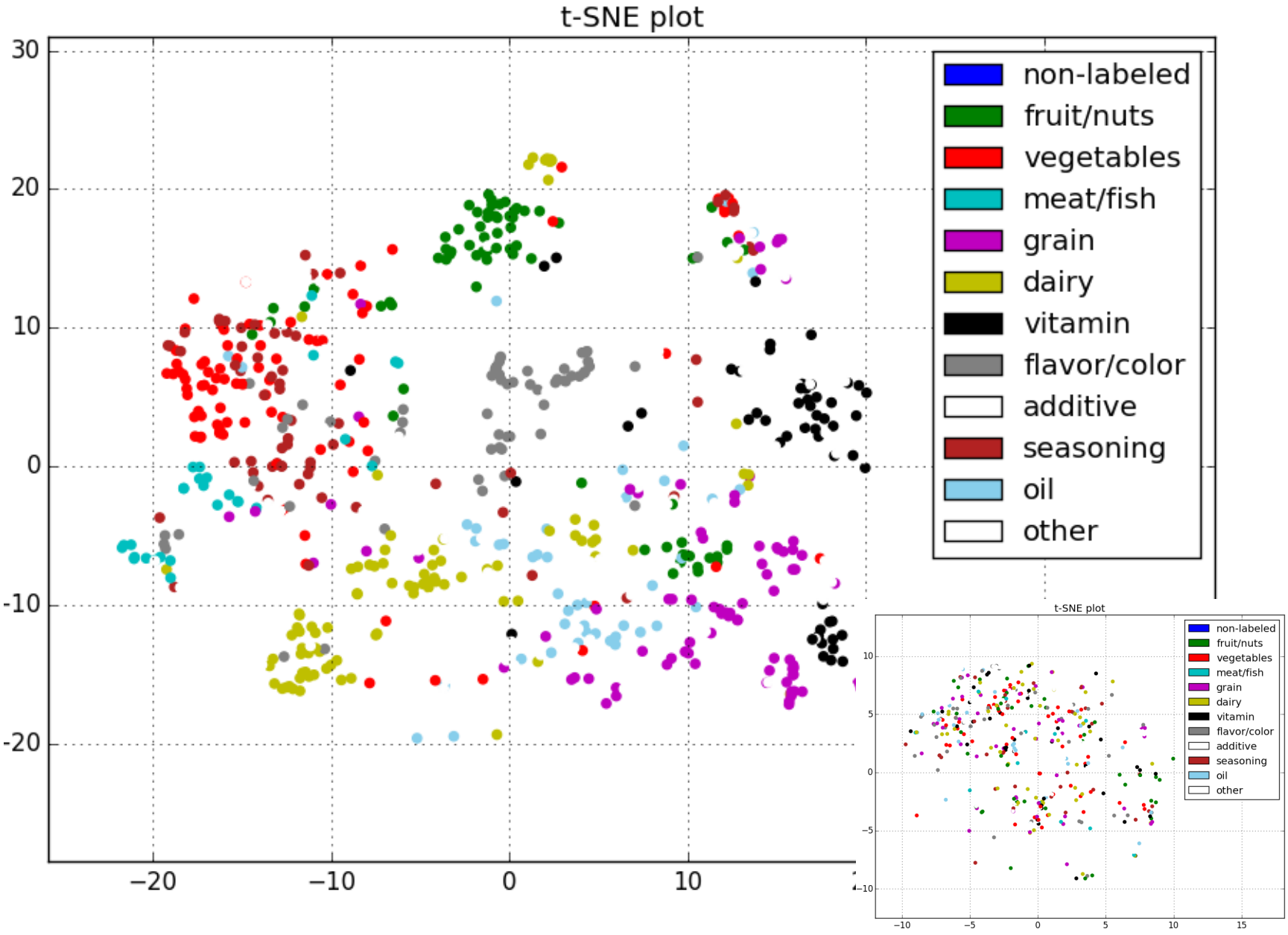
Goal: Predict context given ingredient.

Parameters:

N: 120-20000
learning rate: 0.005-0.1
lambda: 0.0005
m = 10-30
d = 10-30
n_epochs = 8-25
batch_size: 100-200
max_output_len: 4-12
dropout: yes

NEAREST NEIGHBORS

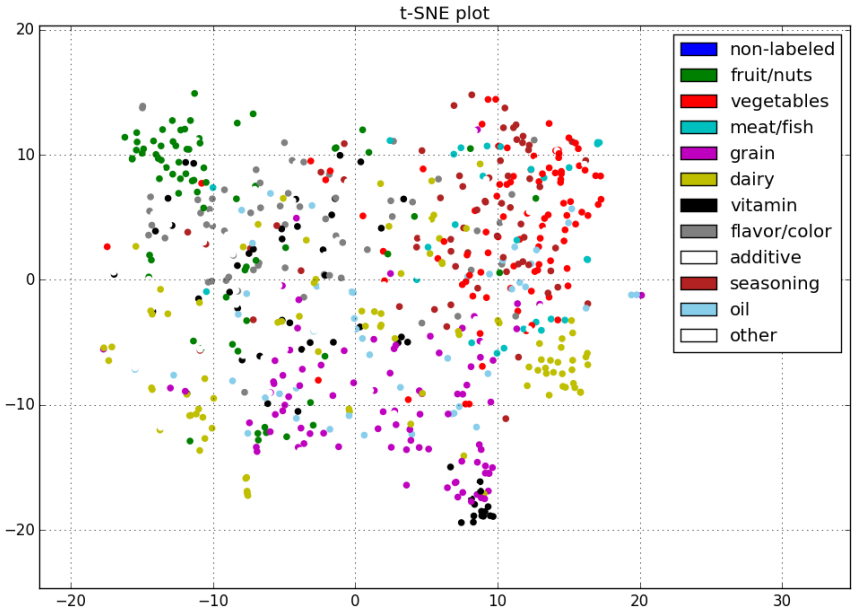
parmesan	--> ['romano cheese' "pasteurized cow's milk" 'parmesan cheese']
whole grain wheat flour	--> ['whole wheat flour' 'granola' 'graham flour']
oat bran	--> ['buckwheat' 'flaxseed' 'millet']
anticaking agent	--> ['anti-caking agent' 'parsley flake' 'paprika oleoresin']
unsweetened chocolate	--> ['chocolate chip' 'chocolate liquor' 'dark chocolate']
malted barley	--> ['malted barley flour' 'flour' 'barley malt']
herb	--> ['roasted garlic' 'garlic' 'green pepper']
apricot	--> ['grape' 'pineapple' 'raspberry']
garlic puree	--> ['red pepper' 'green onion' 'apple cider vinegar']
lime juice	--> ['shallot' 'cilantro' 'white vinegar']
chicken flavor	--> ['chicken fat' 'beef extract' 'yeast extract']
cabbage	--> ['dehydrated vegetable' 'celery' 'water chestnut']
oil	--> ['vegetable oil' 'canola' 'cottonseed']
white pepper	--> ['cumin' 'other spices' 'coriander']
lime juice concentrate	--> ['lemon juice concentrate' 'dried onion' 'mustard flour']
crisp rice	--> ['whole grain oat' 'granola' 'whole grain rolled oat']
red wine vinegar	--> ['white vinegar' 'balsamic vinegar' 'crushed tomato']
banana	--> ['raspberry' 'cherry' 'blueberry']


t-SNE plot
non-labeled
fruit/nuts
vegetables
meat/fish
grain
dairy
vitamin
flavor/color
additive
seasoning
oil
other
30
20
10
0
-10
-20
−20
−10
0
10

PREDICTING VALID COMBINATIONS

- **{'salt' 'sugar' 'peanut' 'peanut oil' 'roasted peanut'}**
  - → Valid combination
- **{'monoglyceride' 'natural flavorings' 'fd&c red 40' 'pasteurized cultured milk' 'vitamin k1'}**
  - → Invalid combination

- **Substitutions**
- **Additions**
- **Removals**



PREDICTING CATEGORIES

- **3 types of categories:**
  - food_category - 1124 choices – “Soda – Diet Cola”
  - shelf - 128 choices – “Soda”
  - aisle - 16 choices – “Drinks”
- **Example input:** {enzymes, pasteurized part skim milk, bifidus, chive}
- **Expected output:** cheese
- **3 models:**
  - Mixture model + alpha-smoothing
  - Max entropy
  - Neural network

Model	N	V	I	I weighted
Logistic regression	120	0.575	0.989	0.508
Logistic regression	1000	0.612	0.976	0.441
Logistic regression	5000	0.003	0.984	0.993
Neural network	120	0.942	0.914	0.844
Neural network	1000	0.861	0.968	0.950
Neural network	5000	0.877	0.956	0.936

model	aisle	shelf	food_category
mixture model	0.672	0.584	0.430
logistic regression	0.764	0.677	0.455
neural network	0.778	0.699	0.503
neural network (5k ings)	0.810	0.735	0.536

UNSEEN INGREDIENTS

Ingredient	*Neighbor 1*	*Neighbor 2*	*Neighbor 3*
vanilla flavor	organic vanilla extract	vanilla	organic vanilla
raisin paste	organic tomato paste	tomato paste	potato flake
dill	herb	organic basil	mustard
cheese sauce	sauce	worcestershire sauce	tomato sauce
green	red	artificial color	color
bleached flour	enriched bleached flour	corn flour	partially defatted peanut flour
sausage	pepperoni	ham	bacon
cane syrup	organic dried cane syrup	dried cane syrup	glucose-fructose syrup
organic almond	almond	tree nut	hazelnut
light tuna	fish	anchovy	sardines

Table 11. Given previously unseen ingredients, we can use the UMLS database to find the nearest neighbors in the seen ingredients. We can then use this new representation to make various predictions (such as the ones presented in this paper).

Example hierarchy for monosodium glutamate (MSG):

monosodium glutamate → glutamic acid → amino acid → carboxylic acid → organic compound.