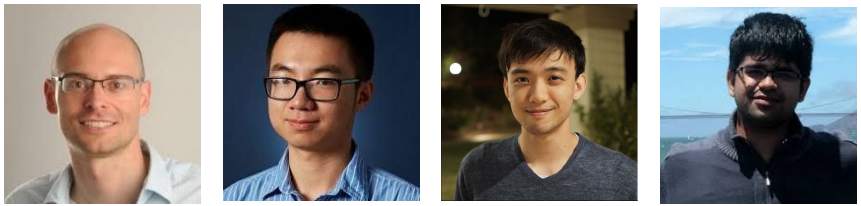
CS294-158 Deep Unsupervised Learning

Lecture 9: Unsupervised Distribution Alignment



Pieter Abbeel, Peter Chen, Jonathan Ho, Aravind Srinivas

UC Berkeley

Distribution Alignment Problem

- Image to Image

(Isola et al, 2017)

Distribution Alignment Problem

- Image to Image

(Isola et al, 2017)

Distribution Alignment Problem

- Image to Image

(Isola et al, 2017)

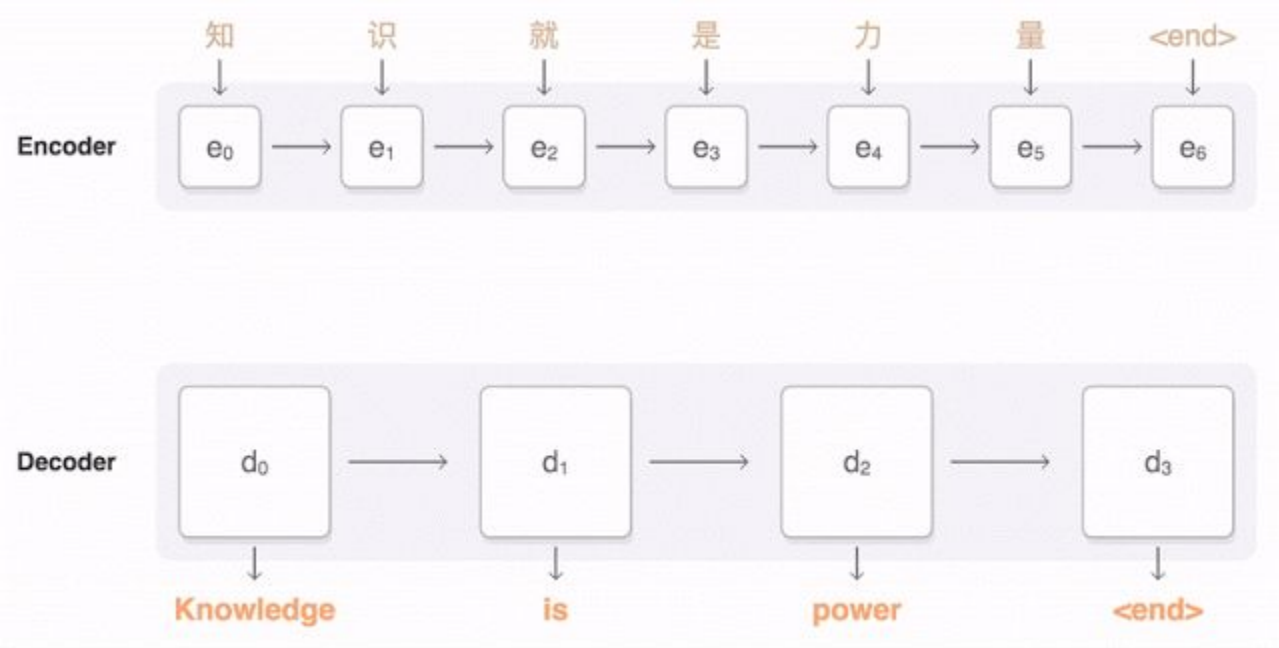
Distribution Alignment Problem

- Image to Image

(Isola et al, 2017)

Distribution Alignment Problem

- Text to text



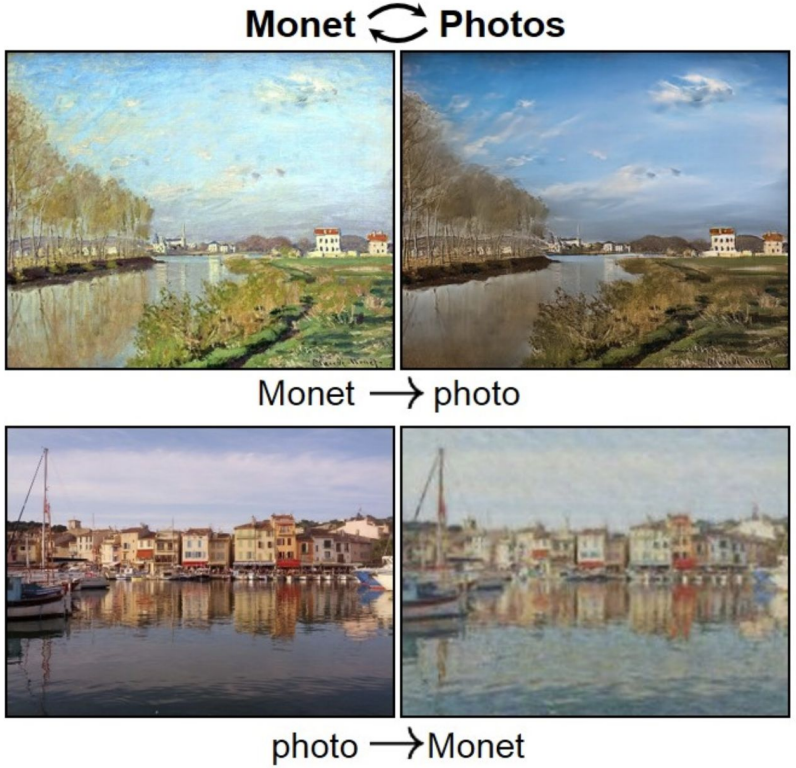
https://ai.googleblog.com/2016/09/a-neural-network-for-machine.html

Supervised Distribution Alignment

- Image to Image: pix2pix (Isola et al, 2017)
- Text to text: machine translation
- Image to text: captioning
- Text to Image, voice to text, text to voice….
- It's simply fitting conditional distribution p(a|b)
  - We have access to (a, b) pairs
  - What if (a, b) pairs are expensive to obtain or just don't exist?

Unsupervised Distribution Alignment

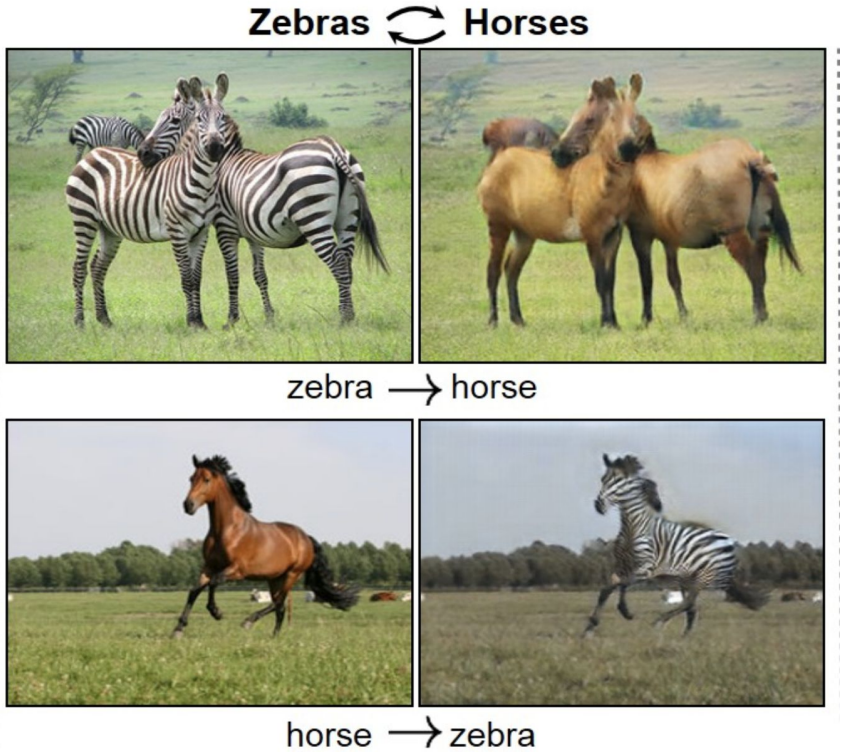
- Image to Image



(Zhu et al, 2017)

Unsupervised Distribution Alignment

- Image to Image



(Zhu et al, 2017)

Unsupervised Distribution Alignment

- A lot of applications:
  - translation to/from small languages that are not economical to label
  - augment labeled examples
  - style transfer etc
- Is this even a feasible problem?
  - We have access to samples from p(a) and p(b)
  - Without any access to samples from p(a, b) we need to estimate p(a|b) and p(b|a)

Marginal Matching

- We will try to learn the relationship between A and B:
  - Approximate p(a|b) by q_θ(a|b) and p(a|b) by q_θ(a|b)
- The marginals induced by approximate mapping q should match original margins:

$$q(b) = \mathbb{E}_{a \sim p(a)} \left[q(b|a) \right] \approx \mathbb{E}_{a \sim p(a)} \left[p(b|a) \right] = p(b)$$

$$q(a) = \mathbb{E}_{b \sim p(b)} \left[q(a|b) \right] \approx \mathbb{E}_{b \sim p(b)} \left[p(a|b) \right] = p(a)$$

- In literature, q(b|a) is oftentimes just a deterministic mapping, which we call G_{AB} : A -> B

Marginal Matching

- [hand-draw 1d example]

Marginal Matching

- [1d example, ambiguity]

Cycle Consistency

- Many names in the literature: Cycle Consistency, Dual Learning, Back translation, …
- Core idea:

$$\mathbb{E}_{b\sim q(b|a)}\left[q(a'|b)\right] \approx \mathbb{E}_{b\sim p(b|a)}\left[p(a'|b)\right] \forall a$$

- In the case of deterministic mapping:

$$G_{BA}(G_{AB}(a)) = a$$
$$G_{AB}(G_{BA}(b)) = b$$

Marginal Matching

- [1d example, with reduced ambiguity]

(partial) Unsupervised Alignment Principles

- We have come up with two invariances that are true for all alignment problems and we can use them as learning signals
  - Marginal Matching
  - Cycle Consistency
- These are obviously not enough in general.
  - In practice, researchers inject additional inductive biases into learning systems by selecting architectures, loss functions, and problems.

CycleGAN

- Marginal matching w/ GAN

$$\mathcal{L}_{\text{GAN}}(G, D_Y, X, Y) = \mathbb{E}_{y \sim p_{\text{data}}(y)}[\log D_Y(y)] + \mathbb{E}_{x \sim p_{\text{data}}(x)}[\log(1 - D_Y(G(x))], \quad (1)$$

- Cycle consistency w/ deterministic mapping & L1 loss

$$\mathcal{L}_{\text{cyc}}(G, F) = \mathbb{E}_{x \sim p_{\text{data}}(x)}[\|F(G(x)) - x\|_1] + \mathbb{E}_{y \sim p_{\text{data}}(y)}[\|G(F(y)) - y\|_1].$$

(Zhu et al, 2017)

CycleGAN

Loss	Per-pixel acc.	Per-class acc.	Class IOU
CoGAN [32]	0.45	0.11	0.08
BiGAN/ALI [9, 7]	0.41	0.13	0.07
SimGAN [46]	0.47	0.11	0.07
Feature loss + GAN	0.50	0.10	0.06
CycleGAN (ours)	**0.58**	**0.22**	**0.16**
pix2pix [22]	0.85	0.40	0.32

Table 3: Classification performance of photo→labels for different methods on cityscapes.

(Zhu et al, 2017)

CycleGAN

Loss	Per-pixel acc.	Per-class acc.	Class IOU
Cycle alone	0.10	0.05	0.02
GAN alone	0.53	0.11	0.07
GAN + forward cycle	0.49	0.11	0.07
GAN + backward cycle	0.01	0.06	0.01
CycleGAN (ours)	**0.58**	**0.22**	**0.16**

Table 5: Ablation study: classification performance of photo→labels for different losses, evaluated on Cityscapes.

(Zhu et al, 2017)

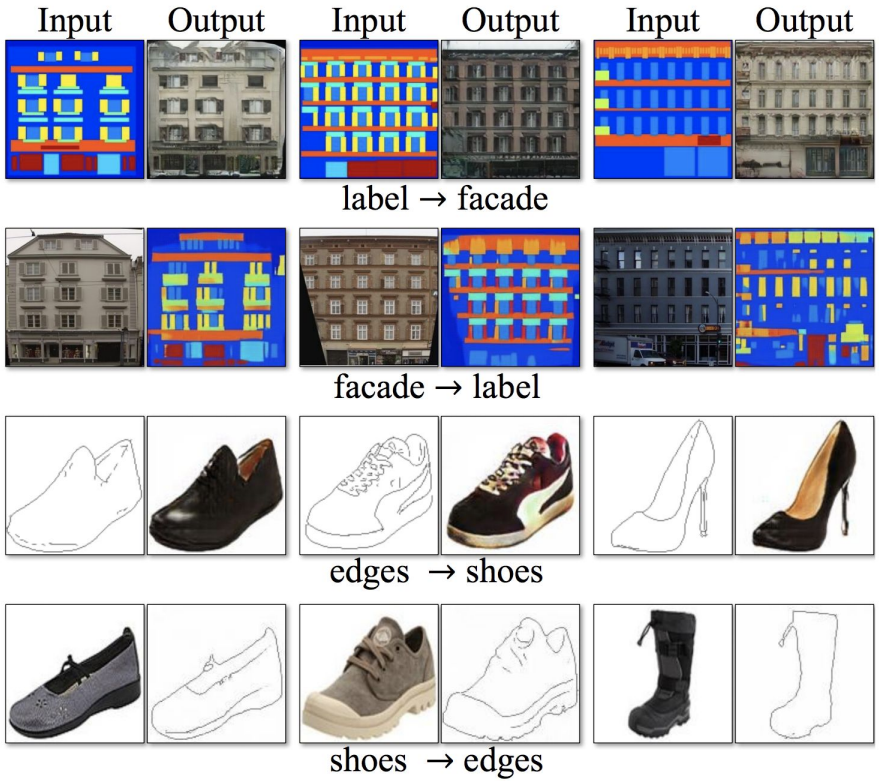
CycleGAN

Loss	Per-pixel acc.	Per-class acc.	Class IOU
Cycle alone	0.22	0.07	0.02
GAN alone	0.51	0.11	0.08
GAN + forward cycle	**0.55**	**0.18**	**0.12**
GAN + backward cycle	0.39	0.14	0.06
CycleGAN (ours)	0.52	0.17	0.11

Table 4: Ablation study: FCN-scores for different variants of our method, evaluated on Cityscapes labels→photo.

(Zhu et al, 2017)

CycleGAN



(Zhu et al, 2017)

CycleGAN

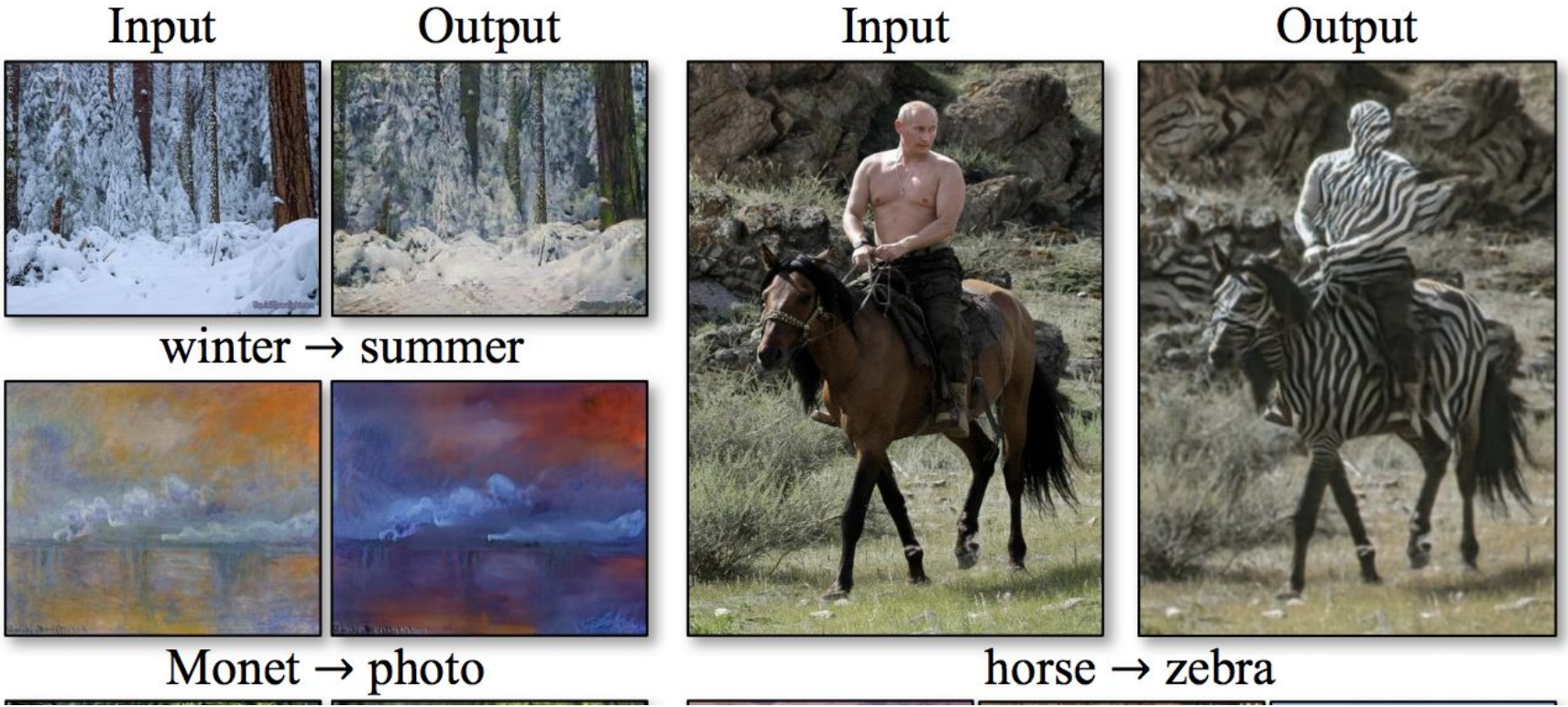
summer Yosemite → winter Yosemite

apple → orange

orange → apple

(Zhu et al, 2017)

CycleGAN failure cases



(Zhu et al, 2017)

Stochastic Mapping

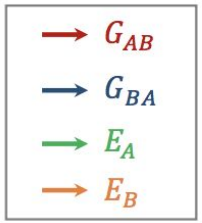
- Sometimes deterministic (one-to-one) mappings are too restricted, e.g. semantic mask <> image
- One straightforward way to extend CycleGAN is to make the mapping take in an additional noise source (DualGAN)

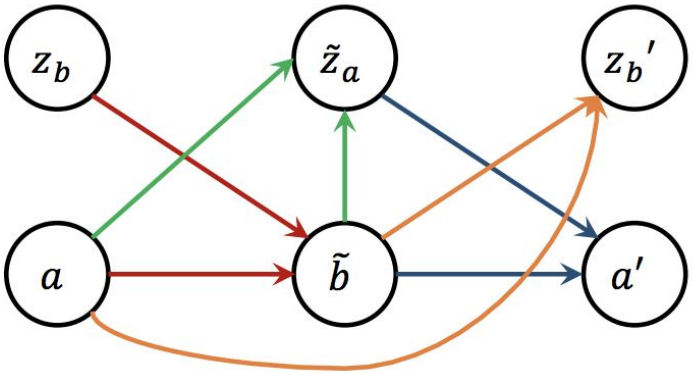
$$G_{AB}(a, z) : A \times Z \rightarrow B \quad \text{instead of } G_{AB}(a) : A \rightarrow B$$

- However, we also need to change cycle-consistency l1 loss since otherwise z will simply be ignored
  - $G_{BA}(G_{AB}(a, z), z') = a$, which means the choice of z, z' is irrelevant

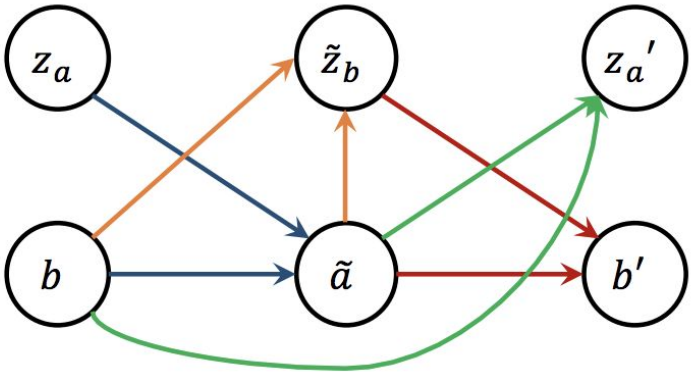
(Almahairi et al, 2017)

Augmented CycleGAN





Cycle starting from $A \times Z_b$



Cycle starting from $B \times Z_a$

(Almahairi et al, 2017)

Augmented CycleGAN

$$\mathcal{L}^A_{\text{CYC}}(G_{AB}, G_{BA}, E_A) = \mathop{\mathbb{E}}_{\substack{a \sim p_d(a) \\ z_b \sim p(z_b)}} \left\| a' - a \right\|_1,$$

$$\tilde{b} = G_{AB}(a, z_b), \; \tilde{z}_a = E_A(a, \tilde{b}), \; a' = G_{BA}(\tilde{b}, \tilde{z}_a). \quad (9)$$

(Almahairi et al, 2017)

CycleGAN can "cheat"

- It "can" generate diverse mappings

(a) AugCGAN (b) StochCGAN

Figure 5: Given an edge from the data distribution (leftmost column), we generate shoes by sampling five $z_b \sim p(z_b)$. Models generate diverse shoes when edges are from the data distribution.

(Almahairi et al, 2017)

CycleGAN can "cheat"

- And be cycle-consistent at the same time

(c) AugCGAN (d) StochCGAN

Figure 6: Cycles from both models starting from a real edge and a real shoe (left and right respectively in each subfigure). The ability for StochCGAN to reconstruct shoes is surprising and is due to the "steganography" effect (see text).

(Almahairi et al, 2017)

Augmented CycleGAN

(a) AugCGAN

(b) StochCGAN

(Almahairi et al, 2017)

Augmented CycleGAN

(a) AugCGAN

(b) StochCGAN

(Almahairi et al, 2017)

Can we do better?

- Can we do better than only relying on 1) Marginal matching and 2) Cycle consistency?
  - Not clear what other invariances we can rely on (good open problem)
- Core problem is that aligning p(a) and p(b) without knowing what's inside a & b is too difficult.
  - One direction: a & b are usually high-dimensional; there is additional structure in them
  - (current image-image alignment works that use ConvNet or patch-based discriminator are already implicitly using this principle)

An NLP Example

- Let's say A = all english sentences, B = all french sentences
  - Two semantically unrelated sentences a, b might have the same frequency p(a) = p(b) and cycle-consistency won't rule that out either
  - Nevertheless, we know each sentence is made up of words and it's unlikely the words in those two unrelated sentences have the same statistics.
- Here we start to make additional assumption, sub-components (e.g. words) of a large random variable (e.g. sentences) can have their own alignment.
  - And we can make use of co-occurrence statistics of the sub-components. word2vec!

recap: word2vec - Skip Gram

Skip-gram model

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log p(w_{t+j}|w_t)$$

$$p(w_O|w_I)=\frac{\exp\left({v'_{w_O}}^{\top}v_{w_I}\right)}{\sum_{w=1}^{W}\exp\left({v'_{w}}^{\top}v_{w_I}\right)}$$

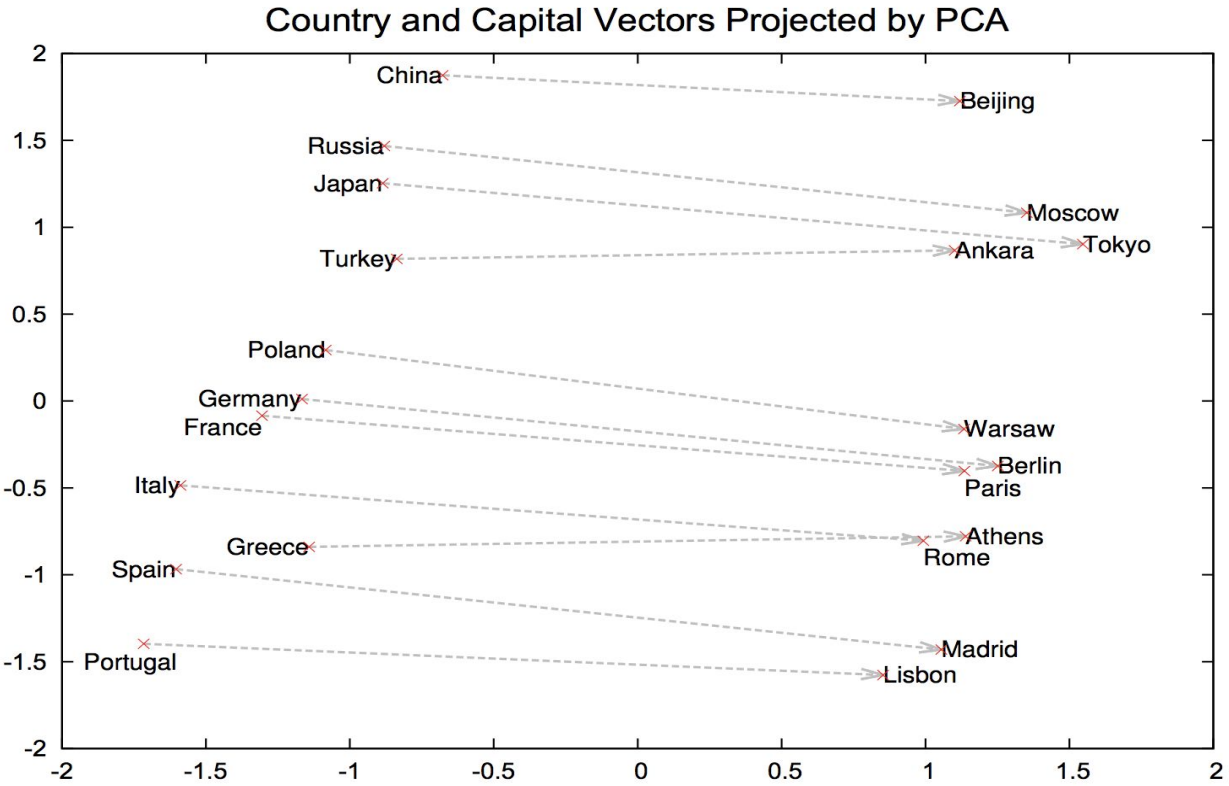
Don't have to have the denominator over all words in the vocabulary

- Can use negative sampling

$$\log\sigma({v'_{w_O}}^{\top}v_{w_I})+\sum_{i=1}^{k}\mathbb{E}_{w_i\sim P_n(w)}\left[\log\sigma(-{v'_{w_i}}^{\top}v_{w_I})\right]$$

$$P(w_i)=1-\sqrt{\frac{t}{f(w_i)}}$$

recap: word2vec



recap: word2vec

	NEG-15 with 10^{-5} subsampling	HS with 10^{-5} subsampling
Vasco de Gama	Lingsugur	Italian explorer
Lake Baikal	Great Rift Valley	Aral Sea
Alan Bean	Rebbeca Naomi	moonwalker
Ionian Sea	Ruegen	Ionian Islands
chess master	chess grandmaster	Garry Kasparov

Table 4: Examples of the closest entities to the given short phrases, using two different models.

Czech + currency	Vietnam + capital	German + airlines	Russian + river	French + actress
koruna	Hanoi	airline Lufthansa	Moscow	Juliette Binoche
Check crown	Ho Chi Minh City	carrier Lufthansa	Volga River	Vanessa Paradis
Polish zolty	Viet Nam	flag carrier Lufthansa	upriver	Charlotte Gainsbourg
CTK	Vietnamese	Lufthansa	Russia	Cecile De

Table 5: Vector compositionality using element-wise addition. Four closest tokens to the sum of two vectors are shown, using the best Skip-gram model.

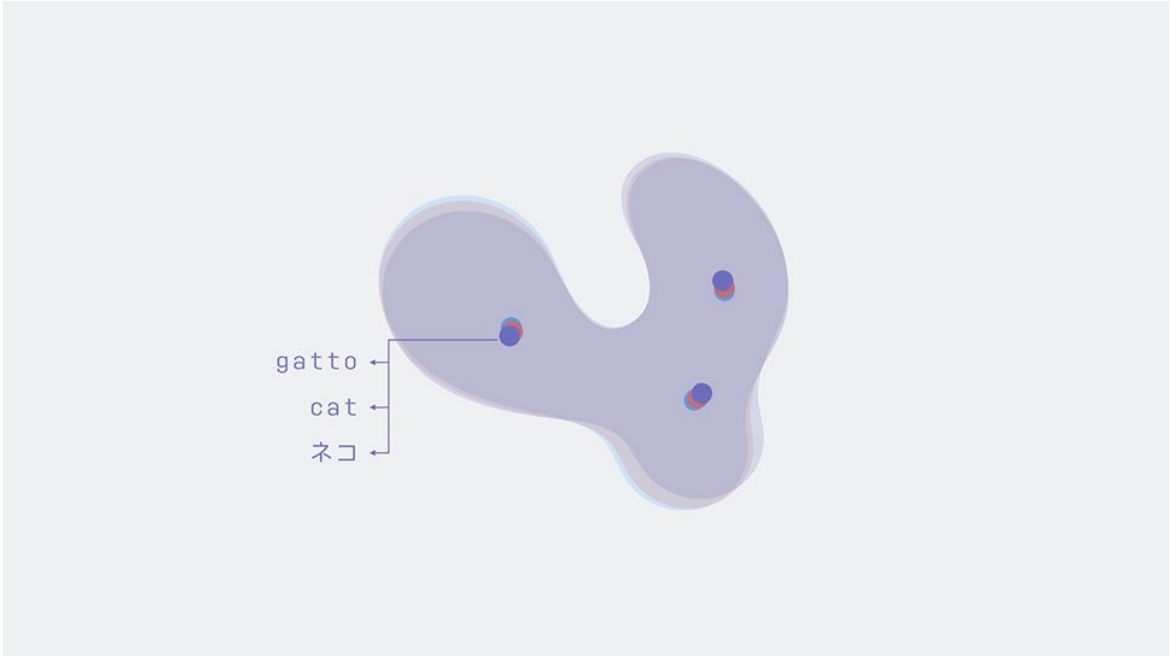
word2vec alignment

- If similar vector calculus holds in all languages, can we align word embeddings in different languages just by uncovering some affine transformation?

(Facebook blog)

word2vec alignment

- Magically it's only a rotation away (Mikolov et al. 2013, Xing et al 2015)



(Facebook blog)

Unsupervised Word Alignment

- (Conneau et al. 2018) proposed the following alignment algorithm
  a. Approximate marginal matching by adversarial training
  b. Refined rotation by solving for exact alignment from a) top pairs

Unsupervised Word Alignment

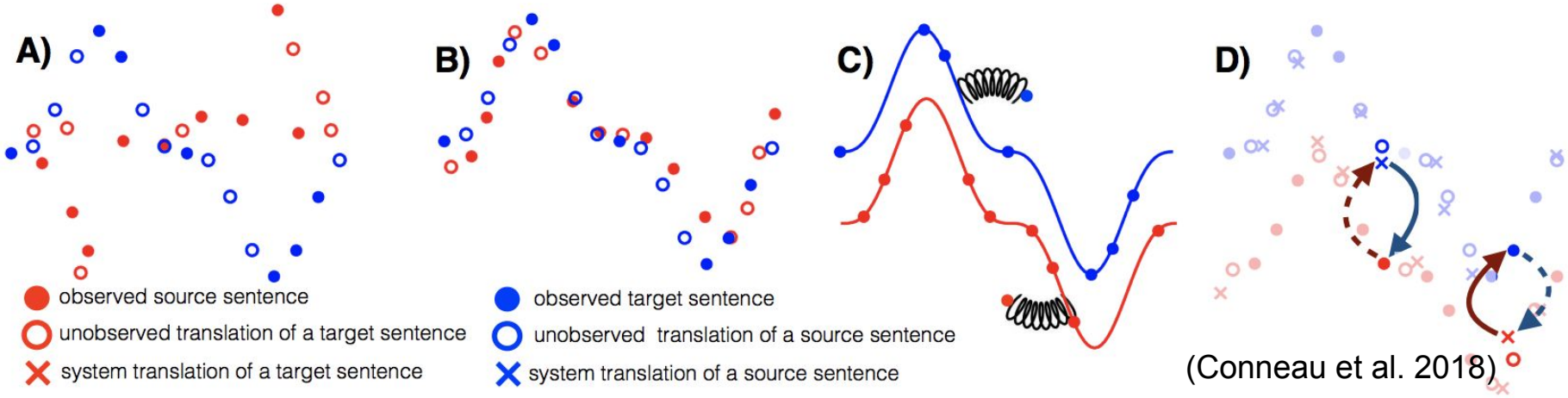
	en-es	es-en	en-fr	fr-en	en-de	de-en	en-ru	ru-en	en-zh	zh-en	en-eo	eo-en
Methods with cross-lingual supervision and fastText embeddings												
Procrustes - NN	77.4	77.3	74.9	76.1	68.4	67.7	47.0	58.2	40.6	30.2	22.1	20.4
Procrustes - ISF	81.1	82.6	81.1	81.3	71.1	71.5	49.5	63.8	35.7	**37.5**	29.0	27.9
Procrustes - CSLS	81.4	82.9	81.1	**82.4**	73.5	**72.4**	**51.7**	**63.7**	**42.7**	36.7	**29.3**	25.3
Methods without cross-lingual supervision and fastText embeddings												
Adv - NN	69.8	71.3	70.4	61.9	63.1	59.6	29.1	41.5	18.5	22.3	13.5	12.1
Adv - CSLS	75.7	79.7	77.8	71.2	70.1	66.4	37.2	48.1	23.4	28.3	18.6	16.6
Adv - Refine - NN	79.1	78.1	78.1	78.2	71.3	69.6	37.3	54.3	30.9	21.9	20.7	20.6
Adv - Refine - CSLS	**81.7**	**83.3**	**82.3**	82.1	**74.0**	72.2	44.0	59.1	32.5	31.4	28.2	**25.6**

Table 1: Word translation retrieval P@1 for our released vocabularies in various language pairs. We consider 1,500 source test queries, and 200k target words for each language pair. We use fastText embeddings trained on Wikipedia. NN: nearest neighbors. ISF: inverted softmax. ('en' is English, 'fr' is French, 'de' is German, 'ru' is Russian, 'zh' is classical Chinese and 'eo' is Esperanto)

(Conneau et al. 2018)

Unsupervised Machine Translation

- (Lample et al. 2018) leverages all 3 core principles:
  - word-level alignment (= sub-component level statistics)
  - monolingual language models (= marginal matching)
  - back translation (= cycle consistency)



(Conneau et al. 2018)

Unsupervised Machine Translation

	en → fr	fr → en	en → de	de → en	en → ro	ro → en	en → ru	ru → en
Unsupervised PBSMT								
Unsupervised phrase table	-	17.50	-	15.63	-	14.10	-	8.08
Back-translation - Iter. 1	24.79	26.16	15.92	22.43	18.21	21.49	11.04	15.16
Back-translation - Iter. 2	27.32	26.80	17.65	22.85	20.61	22.52	12.87	16.42
Back-translation - Iter. 3	27.77	26.93	17.94	22.87	21.18	22.99	13.13	16.52
Back-translation - Iter. 4	27.84	27.20	17.77	22.68	21.33	23.01	13.37	**16.62**
Back-translation - Iter. 5	**28.11**	27.16	-	-	-	-	-	-
Unsupervised NMT								
LSTM	24.48	23.74	14.71	19.60	-	-	-	-
Transformer	25.14	24.18	17.16	21.00	21.18	19.44	7.98	9.09
Phrase-based + Neural network								
NMT + PBSMT	27.12	26.29	17.52	22.06	21.95	23.73	10.14	12.62
PBSMT + NMT	27.60	**27.68**	**20.23**	**25.19**	**25.13**	**23.90**	**13.76**	**16.62**

(Conneau et al. 2018)

Unsupervised Machine Translation

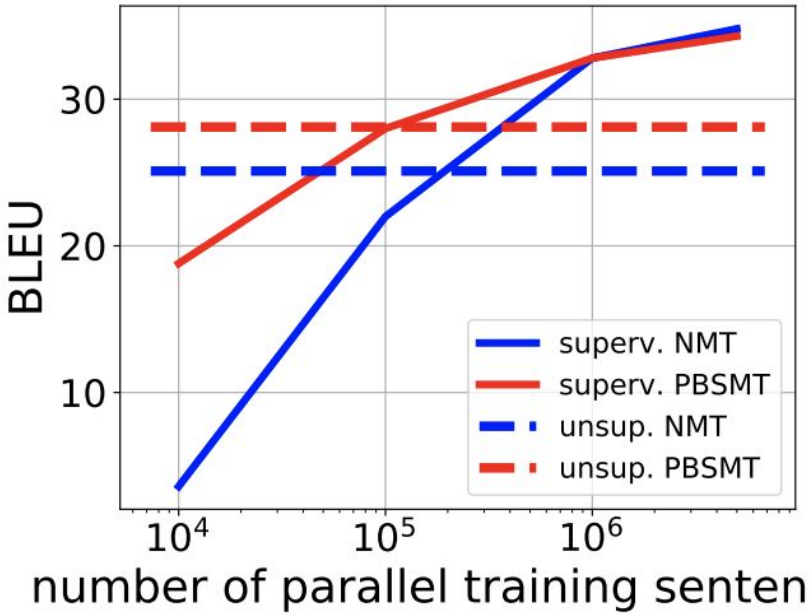
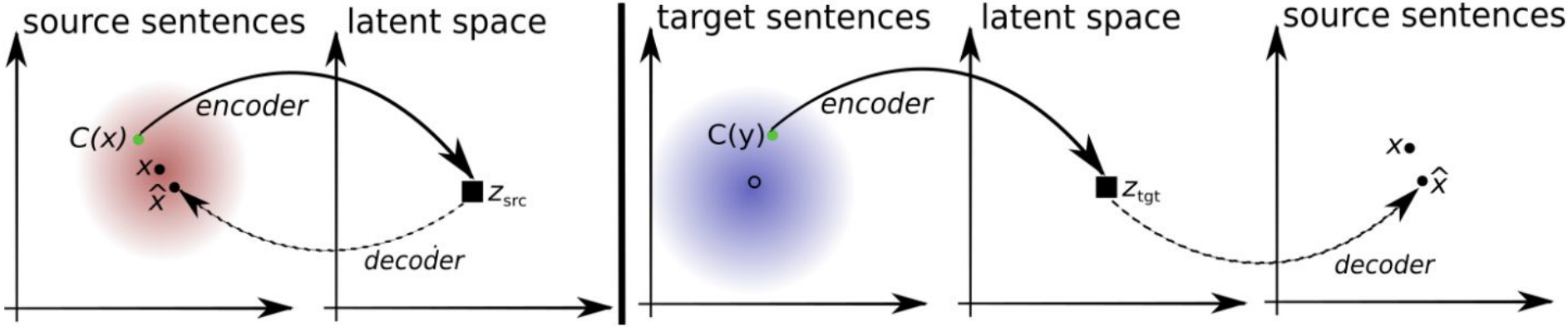


Figure 2: Comparison between supervised and unsupervised approaches on WMT'14 En-Fr, as we vary the number of parallel sentences for the supervised methods.

(Conneau et al. 2018)

Unsupervised Machine Translation



- problem setup: two domains, they are aligned in some way.
  - examples of them: images, video, text…
- Can we learn to translate one to another? We have access to p(x) and p(y)
- first glance the problem is hopeless, where do you get training signal?
  - there are actually some learning signal:
    - marginal matching
      - show how it works in 1d
      - and show how it breaks on frequency parity cases
    - cycle consistency
      - help reduce part of search space
- Based on these priciples, CycleGAN & DualGAN *DiscoGAN
  - arch, loss
  - results
- Fundamentally unimodal, how to handle multi modality?
  - hint: adding z doesn't help
  - augmented cyclegan
- Discuss failure cases, fundamentally marginal matching + cycle consistency isn't enough information to align
  - one key insight is to look into p(x), what's inside each x
  - x is oftentimes high-dim, look at the mutual information between each dimension..
- one example is word2vec, co-occurence shapes embedding space, then apply marginal matching
- build up from word, you can use marginal matching + cycle consistency to achieve alignment at sentence level

Unsupervised Machine Translation