

マルチモーダル情報と機械翻訳

東京大学 大学院情報理工学系研究科
創造情報学専攻 准教授
中山 英樹



MACHINE PERCEPTION GROUP

自己紹介

- ▶ 中山英樹
 - 東京大学 創造情報学専攻 准教授
 - 産総研人工知能センター 招聘研究員



- ▶ 研究分野
 - コンピュータビジョン
 - 自然言語処理
 - 深層学習



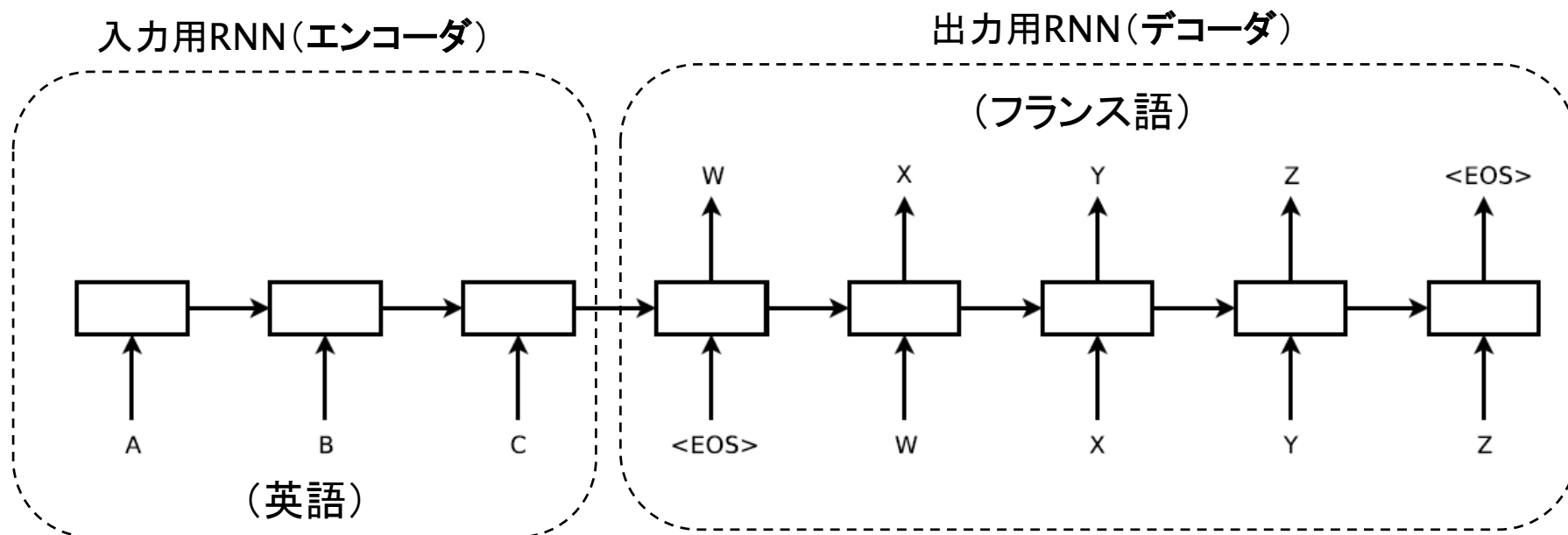
Machine Perception Group

目次

- ▶ 1. エンコーダ・デコーダモデルの発展
- ▶ 2. クロスモーダル・マルチモーダルな機械翻訳

リカレントニューラルネットワーク(RNN)を用いた機械翻訳

- ▶ Sequence to sequence [Sutskever+, NIPS'14]
 - 二つのRNN (LSTM) を接続し、ソース言語・ターゲット言語の文（単語列）の入出力関係を学習
 - **エンコーダ・デコーダモデル**

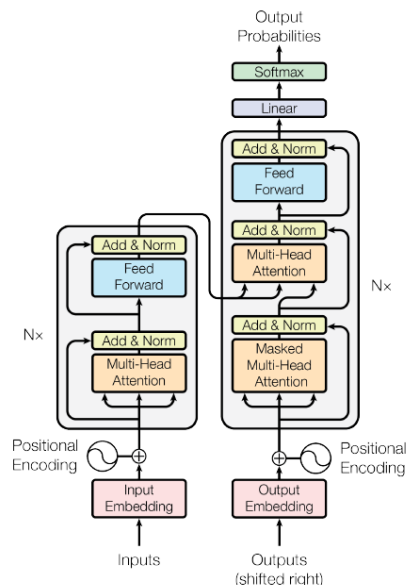


Sutskever et al., "Sequence to Sequence Learning with Neural Networks", In Proc. of NIPS, 2014.

各分野におけるネットワークの発展

▶ 自然言語処理

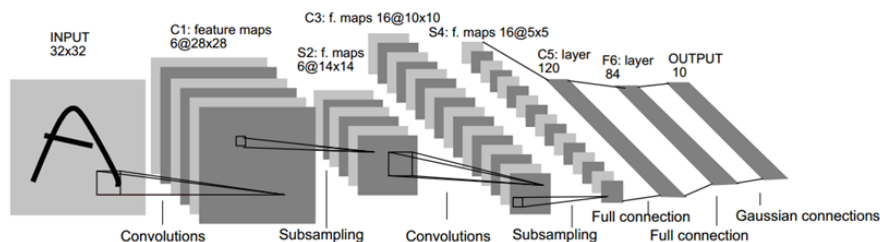
- **Transformer [Vaswani+, 2017]**
- 時系列方向の集積を行わない
- フィードフォワードと注意機構のみで大域的情報を利用



Vaswani et al., “Attention Is All You Need”, In Proc. of NIPS, 2017.

▶ 画像認識

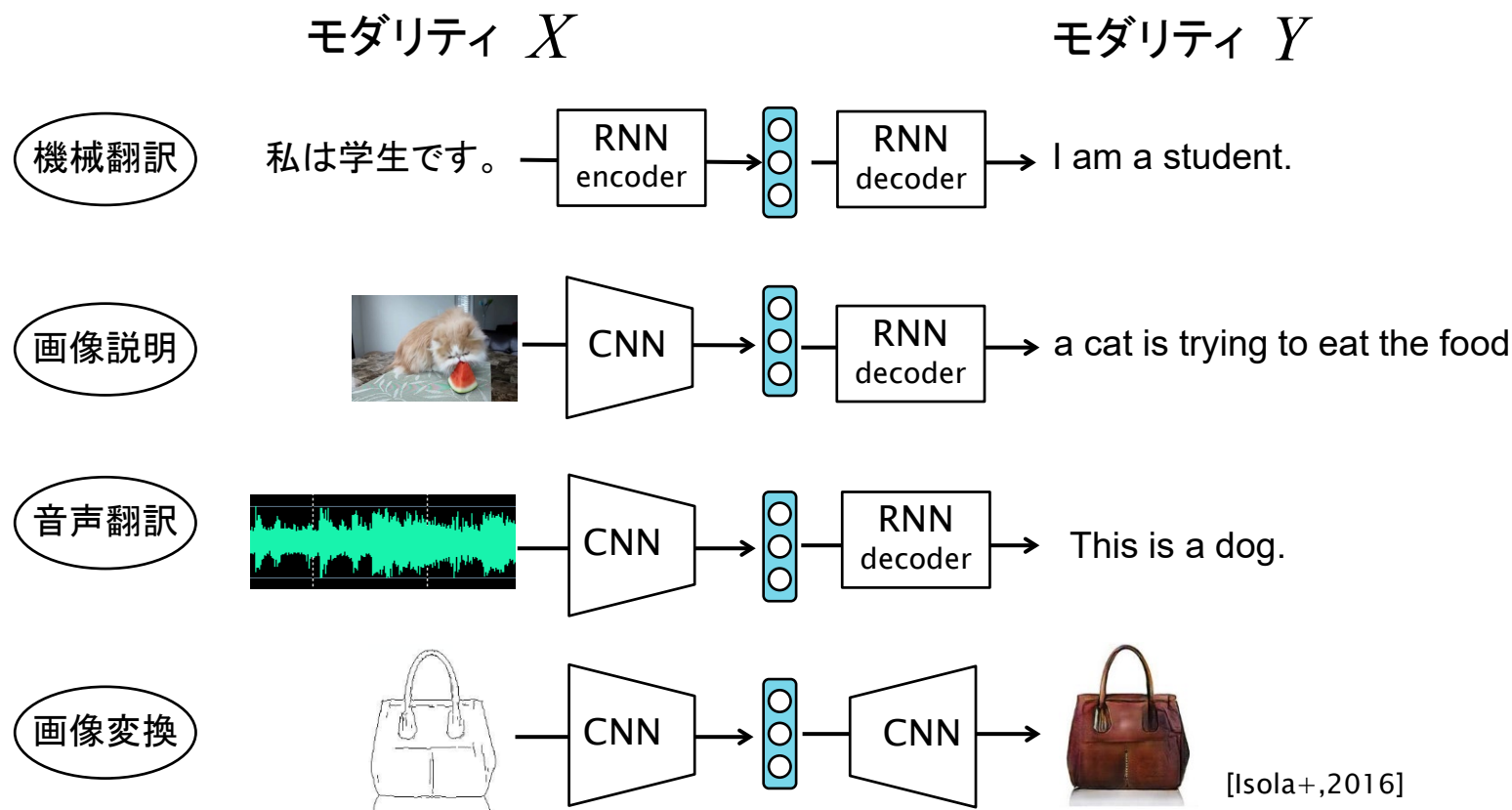
- **畳み込みニューラルネットワーク (CNN) [LeCun+, 1998]**
- 脳のV1視覚野に関する知見をもとに設計
- CNNは系列データ全般でかなり有効



Y. LeCun et al., “Gradient-Based Learning Applied to Document Recognition”, Proceedings of the IEEE, 86(11):2278–2324, 1998.

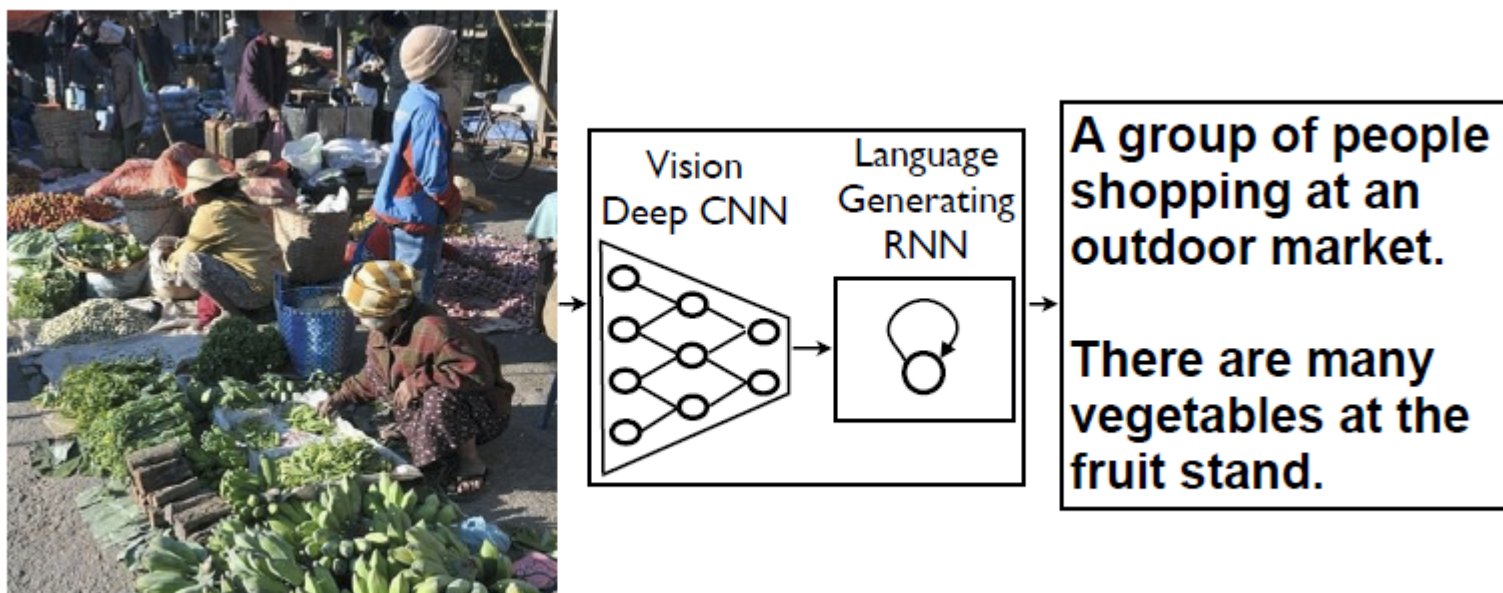
クロスモーダル技術の発展

- ▶ それぞれの分野で定番の**エンコーダ・デコーダ**が確立
- ▶ 柔軟にアプリケーションの設計ができるように



画像説明文生成

- ▶ CNN (画像エンコーダ) をRNN (テキストデコーダ) へ接続
 - RNN側の誤差をCNN側までフィードバック (end-to-end)
 - **画像から言語への“翻訳”**



O. Vinyals et al., “Show and Tell: A Neural Image Caption Generator”, In Proc. CVPR, 2015.

動画画像キャプションニング [Laokulrat+, COLING'16]

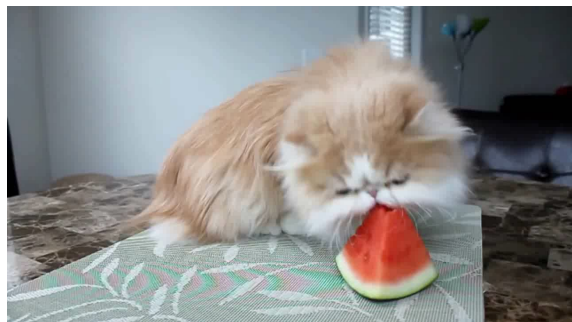
産総研AIRCでの成果

この成果は、国立研究開発法人新エネルギー・産業技術総合開発機構 (NEDO) の委託業務の結果得られたものです

認識結果



a woman is slicing some vegetables



a cat is trying to eat the food

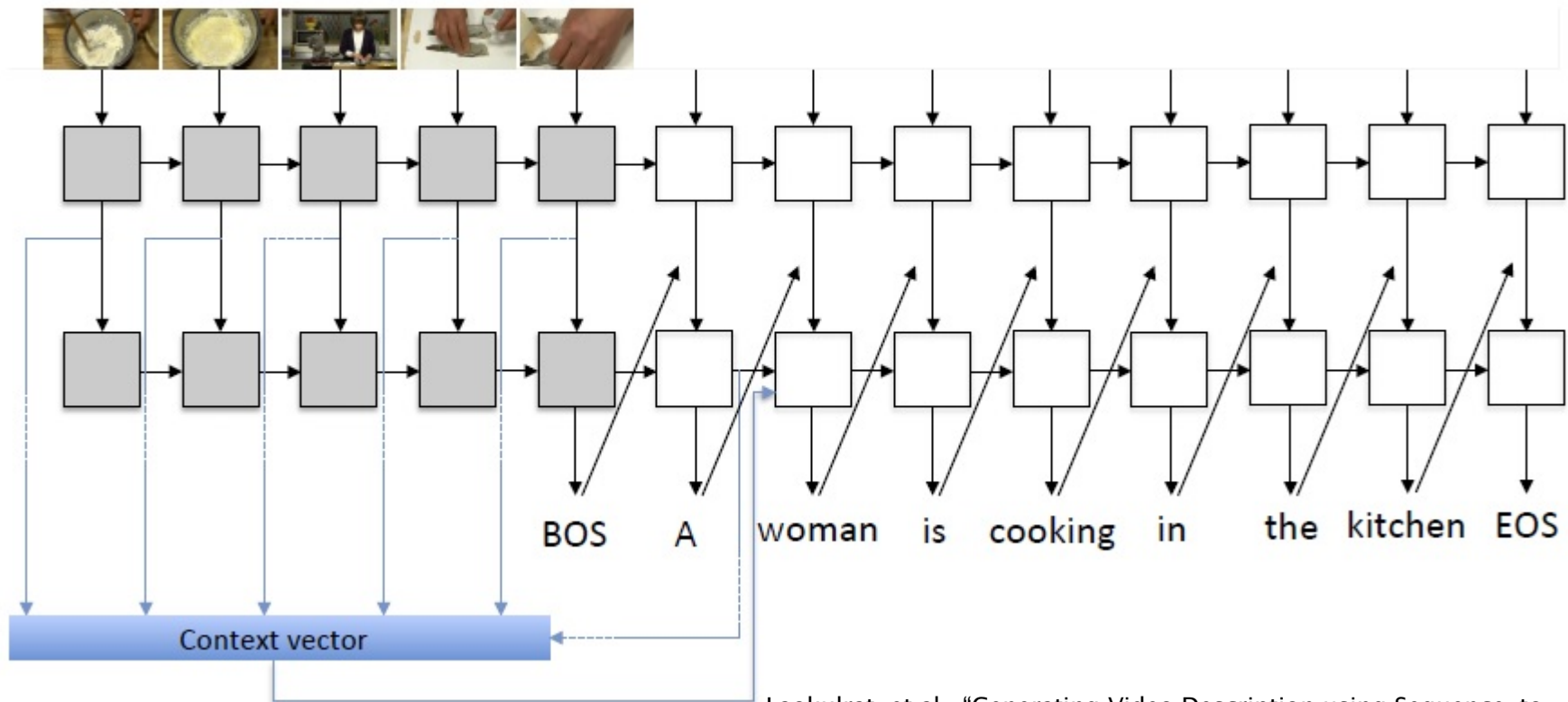


a dog is swimming in the pool

Laokulrat et al., "Generating Video Description using Sequence-to-sequence Model with Temporal Attention", In Proc. of COLING, 2016.

動画画像キャプションング [Laokulrat+, COLING'16]

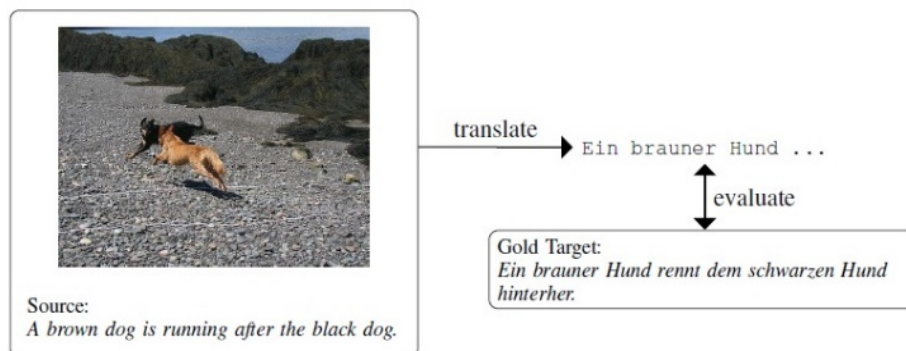
- ▶ CNNにより動画のフレームごとに特徴抽出を行い、時系列データとしてRNNへ入力
- ▶ アテンション機構により、重要なフレームへ重みづけ



その他の言語+画像タスクの例

▶ マルチモーダル機械翻訳

- 機械翻訳の曖昧性解消に画像を活用



[Specia+, 2016]

▶ マルチモーダル対話応答

- 画像内容を前提とした対話
- 中身を理解しないと会話が成立しない



User1: My son is ahead and surprised!
User2: Did he end up winning the race?
User1: Yes he won, he can't believe it!

[Mostafazadeh+, 2017]

Specia et al., "A Shared Task on Multimodal Machine Translation and Crosslingual Image Description", In Proc. of WMT, 2016.

Mostafazadeh et al., "Image-Grounded Conversations: Multimodal Context for Natural Question and Response Generation", In Proc. of IJCNLP, 2017.

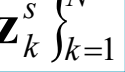
画像を媒介としたゼロショット機械翻訳

[Nakayama and Nishida, 2017]

- ▶ 一般的な方法
(教師付き学習)
 - 大規模な平行コーパスが必要

X		Y	
Japanese		English	

- ▶ **提案法 (画像ピボット)**
 - 画像付きの単一言語ドキュメントのみ
 - Webから容易に収集可能

X		Z	Y	
Japanese		Image	English	
				
				
				
				
				
				

$\mathcal{T}^t = \{\mathbf{z}_k^t, \mathbf{y}_k\}_{k=1}^{N^t}$

$\mathcal{T}^s = \{\mathbf{x}_k, \mathbf{z}_k^s\}_{k=1}^{N^s}$

Nakayama and Nishida, "Zero-resource machine translation by multimodal encoder-decoder network with multimedia pivot", Machine Translation Journal, 2017.

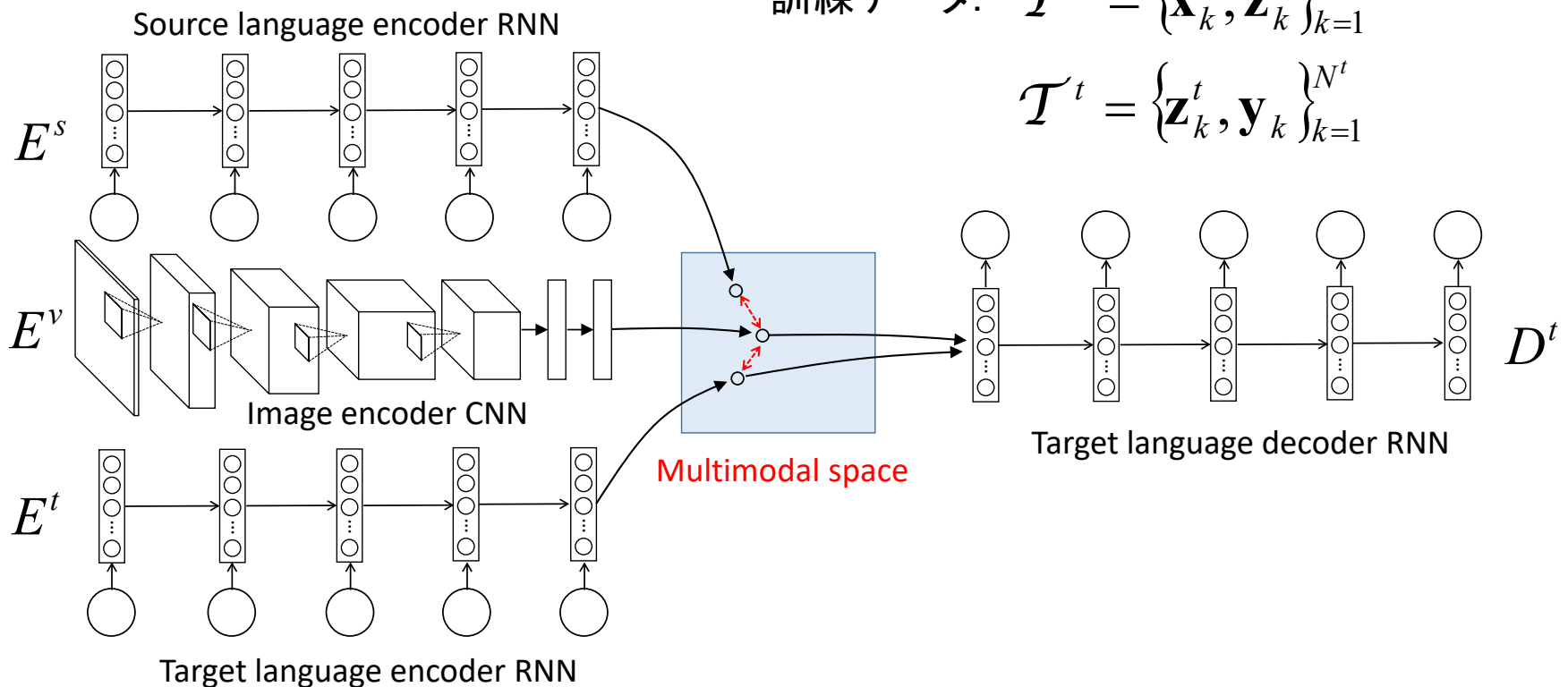
Funaki and Nakayama, "Image-mediated learning for zero-shot cross-lingual document retrieval", In Proc. of EMNLP, 2015.

画像ピボットを用いる機械翻訳モデル

- ▶ ソース言語・ターゲット言語・画像に共通の分散表現を学習
- ▶ ターゲット言語のデコーダをマルチモーダル表現に接続

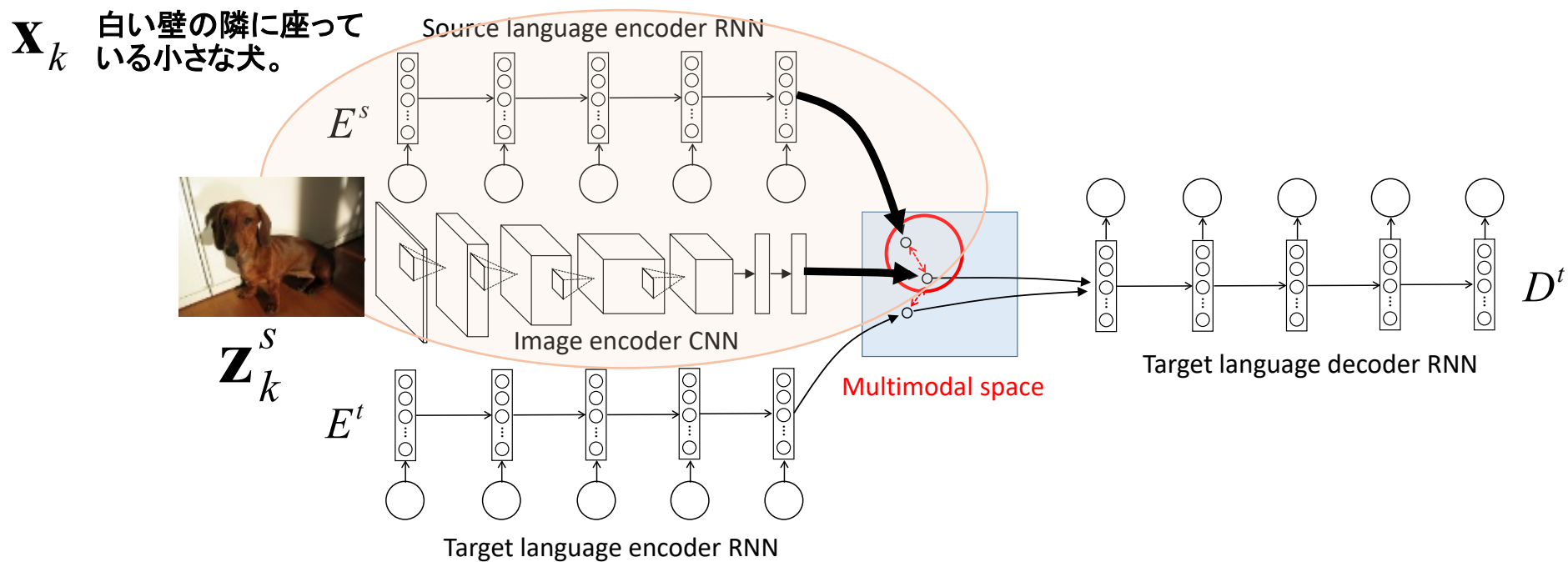
訓練データ: $\mathcal{T}^s = \{\mathbf{x}_k, \mathbf{z}_k^s\}_{k=1}^{N^s}$

$$\mathcal{T}^t = \{\mathbf{z}_k^t, \mathbf{y}_k\}_{k=1}^{N^t}$$



目的関数 1

- ▶ ソース言語と画像をマルチモーダル空間上でアラインメント

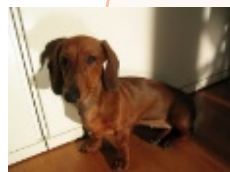


目的関数 1

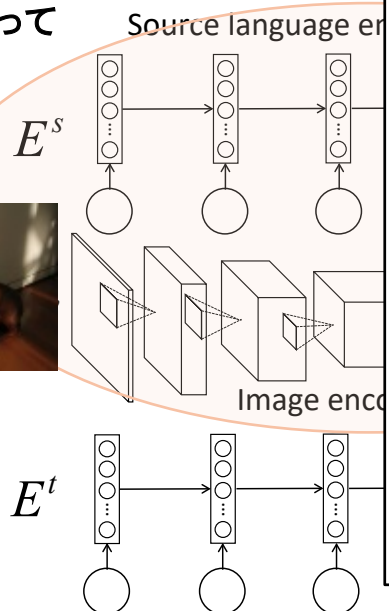
- ソース言語と画像をマルチモーダル空間上でアラインメント

$\mathbf{x}_k$

白い壁の隣に座っている小さな犬。



$\mathbf{z}_k^s$



Target language encoder RNN

$$\mathcal{T}^s = \{\mathbf{x}_k, \mathbf{z}_k^s\}_{k=1}^{N^s}$$

Pair-wise Rank Loss

[Frome+, NIPS'13]

$$\mathcal{L}^s = \sum_k \sum_{i \neq k} \max\{0, \alpha - s(E^v(\mathbf{z}_k^s), E^s(\mathbf{x}_k)) + s(E^v(\mathbf{z}_k^s), E^s(\mathbf{x}_i))\}$$



$s(\)$: Similarity score function

Margin
(Hyper
parameter)

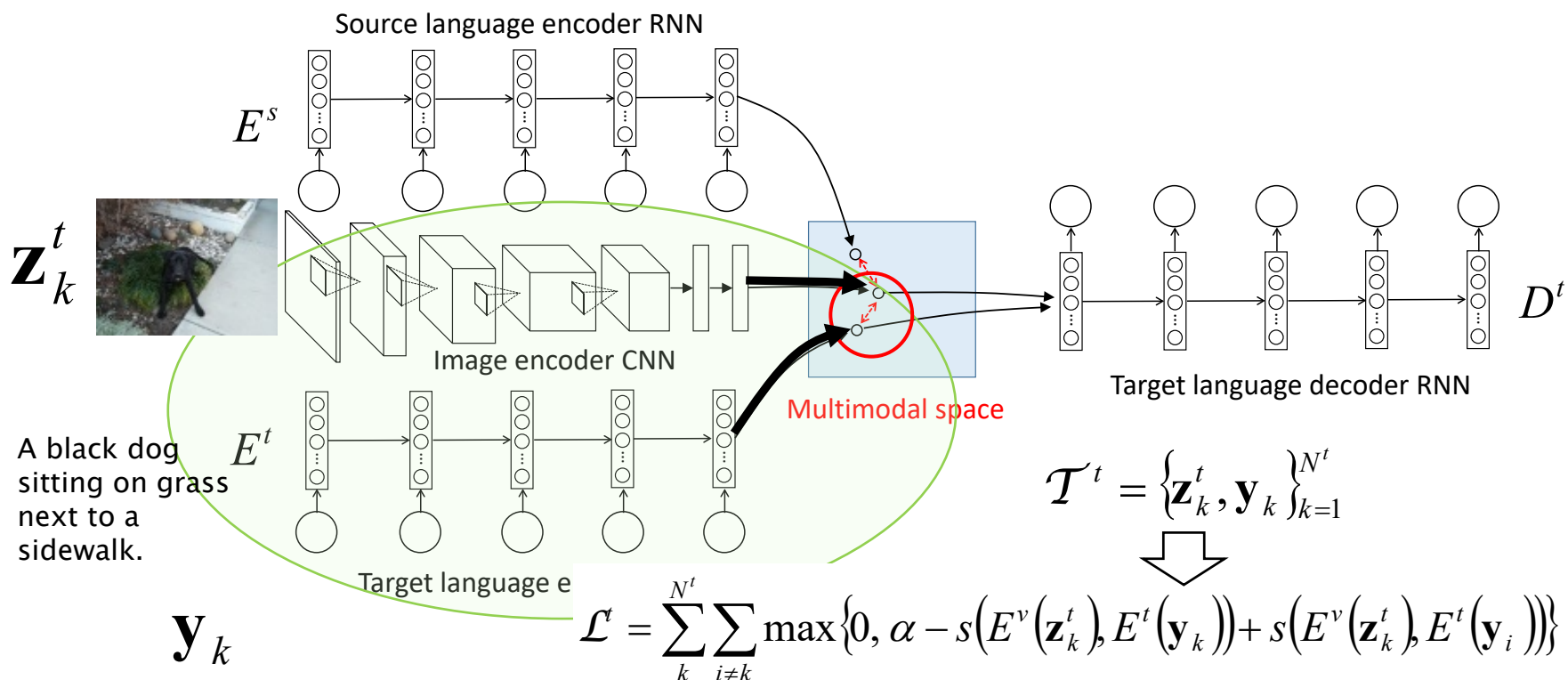
An image

Paired
text

Negative
(not paired)
text

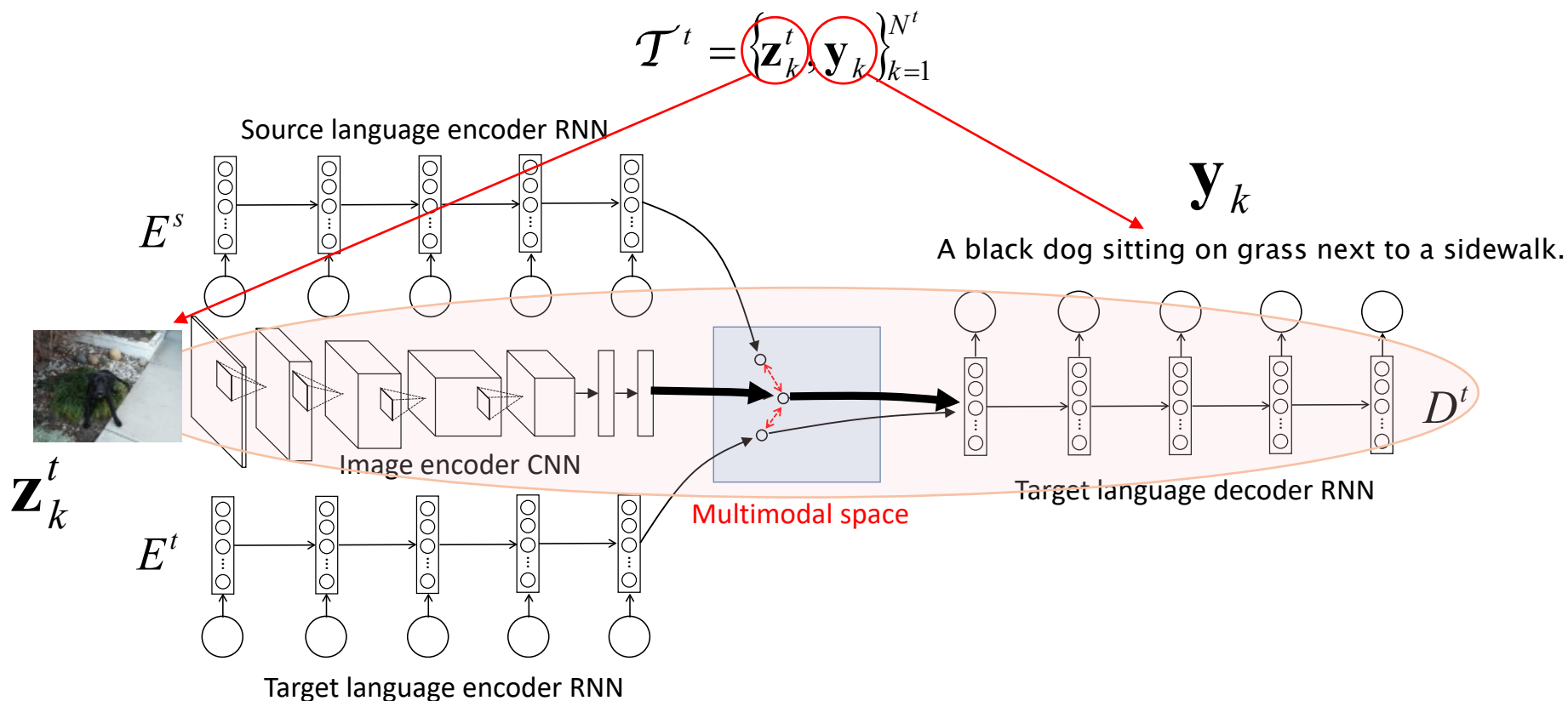
目的関数 2

- ▶ ターゲット言語と画像をマルチモーダル空間上でアラインメント



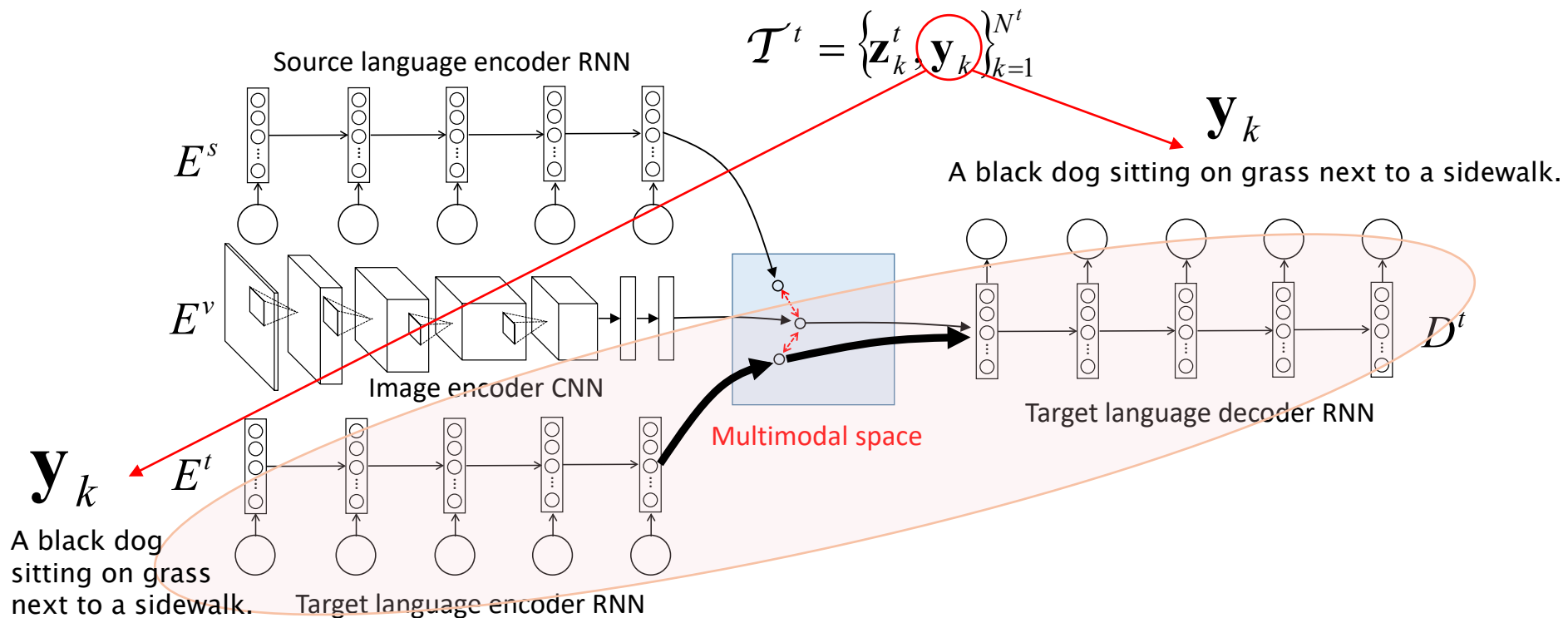
目的関数 3

- ▶ 画像を入力、ターゲット言語テキストをデコード
- ▶ クロスエントロピー損失



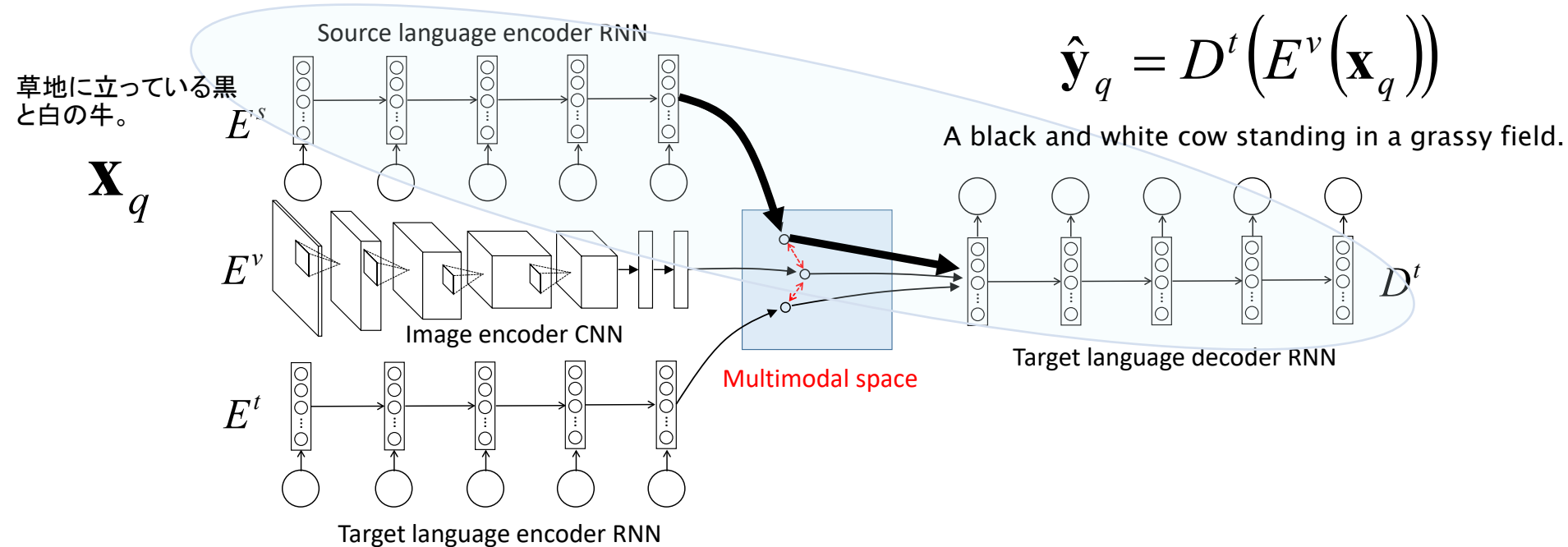
目的関数 4

- ▶ ターゲット言語テキストを入力、再構築



テスト時

- ▶ エンコーダ・デコーダをフィードフォワードするだけ
- ▶ テスト時には画像は必要ない



データセット

- ▶ IAPR-TC12 [Grubinger+, 2006]
 - 二万枚の英独キャプション付き画像



a photo of a brown sandy beach; the dark blue sea with small breaking waves behind it; a dark green palm tree in the foreground on the left; a blue sky with clouds on the horizon in the background;



ein Photo eines braunen Sandstrands; das dunkelblaue Meer mit kleinen brechenden Wellen dahinter; eine dunkelgrüne Palme im Vordergrund links; ein blauer Himmel mit Wolken am Horizont im Hintergrund;



- ▶ Multi30K [Elliott+, 2016]
 - 約三万枚の英独キャプション付き画像
- ▶ ランダムにデータを分け、ゼロショットの独英翻訳を評価

評価結果

- ▶ 評価指標: BLEU値 (大きいほど良い)

提案法
(ゼロショット)

Topology	Training Strategy	Decoder training	IAPR-TC12	Multi30K
De → En				
3-way	end-to-end	image	24.3 (17.2)	18.1 (3.4)
3-way	end-to-end	description	26.2 (19.8)	18.9 (4.2)
3-way	end-to-end	image + description	26.7 (20.2)	18.7 (3.9)

教師付き学習
(理想値)

Data Size	IAPR-TC12	Multi30K
De → En		
9000/14000	47.2 (42.5)	24.5 (9.8)
3000	32.9 (26.5)	<u>18.9 (3.8)</u>
2000	<u>29.2 (22.6)</u>	17.7 (2.8)
1000	25.6 (18.4)	16.3 (2.2)

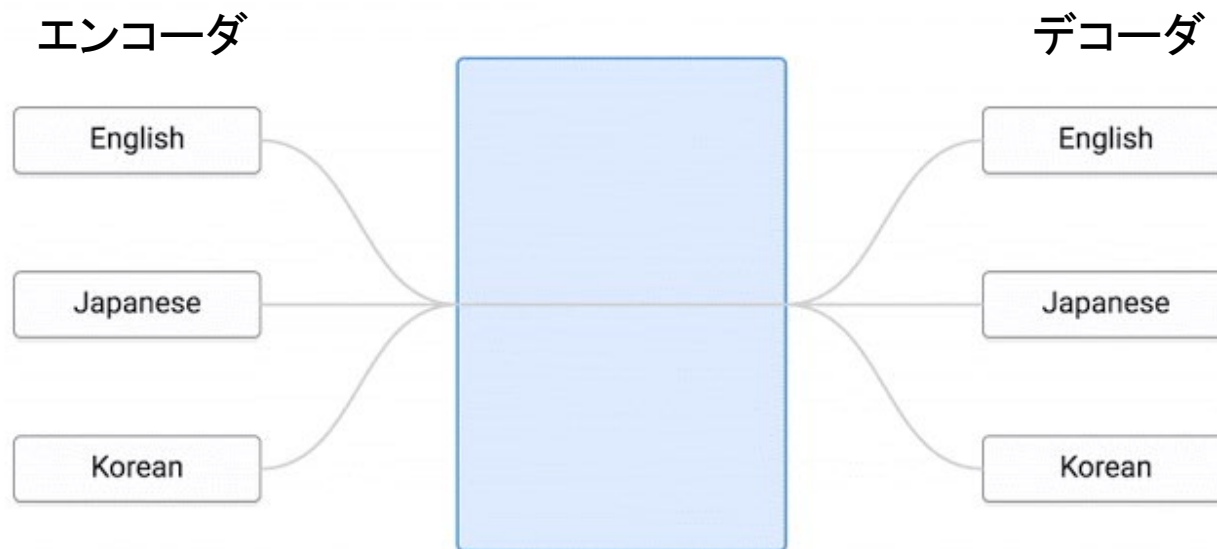
- ▶ 教師付きの場合に用いるパラレルコーパスの5倍程度の画像付き単一ドキュメントを用いると同等の性能

エラー分析

Attribute, counting errors	
ein dunkelhäutiges mädchen mit langen schwarzen haaren und einem blauen pullover steht an einem braunen ufer im vordergrund; ein dunkelblauer see dahinter; weiße wolken an einem blauen himmel im hintergrund;	a dark-skinned boy with long black hair and a white sweater is standing in a brown shore in the foreground; a dark blue lake behind it; white clouds in a blue sky in the background; (a dark-skinned girl with long black hair and a blue pullover is standing on a brown shore in the foreground; a dark blue lake behind it; white clouds in a blue sky in the background;)
eine frau in einem rosa kleid hält ein baby.	a young in a blue shirt is holding a baby. (a woman in a pink skirt is holding a baby.)
drei männer stehen auf einem siegerpodium mit einer gelbblauweißen wand dahinter;	a men are standing on a podium with a yellow, blue and white wall behind it; (three men are standing on a podium with a yellow, blue and white wall behind it;)
ein blondes kind schaukelt auf einer schaukel.	a little boy is on a swing. (a blond child swinging on a swing.)
Gramatical errors	
eine braune berglandschaft mit einigen schneebedeckten bergen;	a brown mountain landscape with a snow snow covered mountains; (a brown mountain landscape with a few snow covered peaks;)
blick auf die häuser einer stadt am meer mit grauen wolken an einem blauen himmel im hintergrund;	view of a houses of a city at a sea; a clouds in the city sky in the background; (view of the houses of a city at the sea with grey clouds in a blue sky in the background;)

マルチモーダルによる知識転移 [Johnson+, TACL'17]

- ▶ グーグルの機械翻訳 (many-to-manyモデル)
 - 共通の中間表現を介することで、**直接教示していない言語対についても翻訳が（ある程度）可能に**
 - 例) 日⇔英、韓⇔英のみ学習すると、日⇔韓の翻訳ができる
 - あるモダリティ（この場合英語）が仲立ちした知識転移

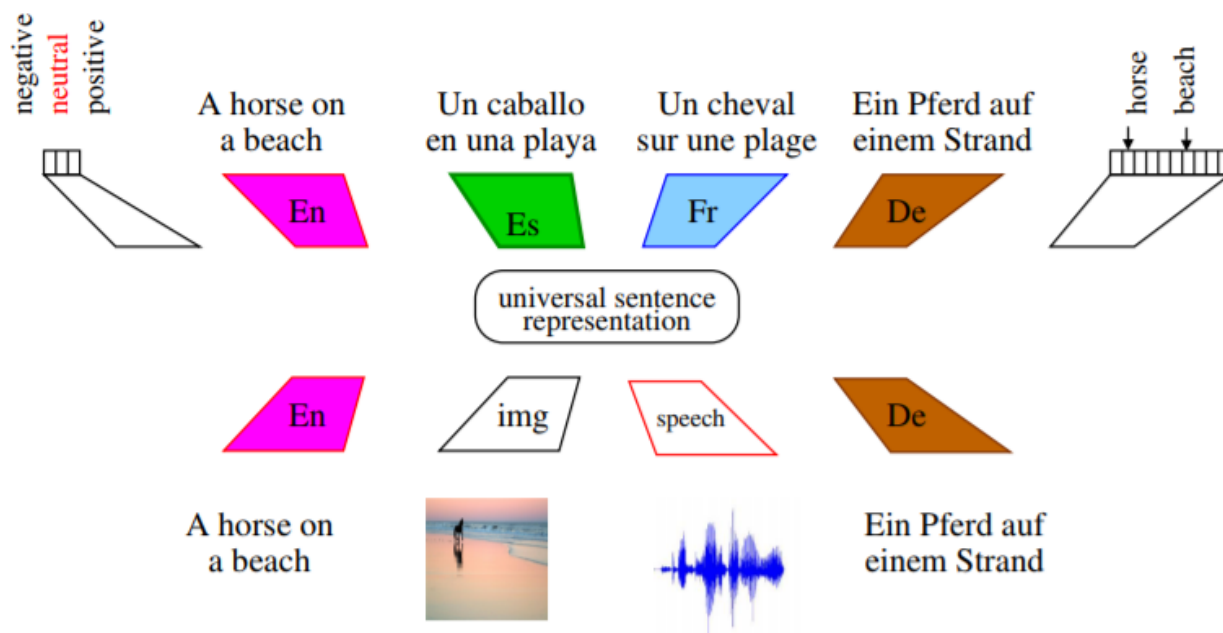


<https://ai.googleblog.com/2016/11/zero-shot-translation-with-googles.html>

Johnson et al., "Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation", Transactions of the ACL, 2017.

Many-to-manyモデルの可能性

- ▶ マルチ入力・マルチタスク
- ▶ ゆくゆくは、さまざまなモダリティ・タスクを横断する汎用的表現を獲得？
- ▶ **知識転移・メタ学習**はホットなトピック



まとめ

- ▶ 深層学習が各分野で浸透
 - 共通の道具（ニューラルネット）で異なるドメインをシームレスに接続することが可能に
 - 機械翻訳においてもさまざまな言語外情報を活用できる
- ▶ マルチモーダルのご利益
 - 精度・頑健性の向上
 - 言語外の文脈情報の取り込み
 - 知識転移・メタ学習などへの応用
- ▶ アイデア次第でいろいろな面白いことが出来る時代
 - 分野間コラボレーションがますます重要に