マルチモーダル深層学習の発展

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

自己紹介

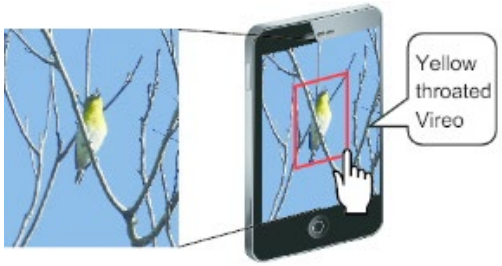
- 中山英樹
  - 東京大学 創造情報学専攻 准教授
  - 産総研人工知能センター 招聘研究員

- 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

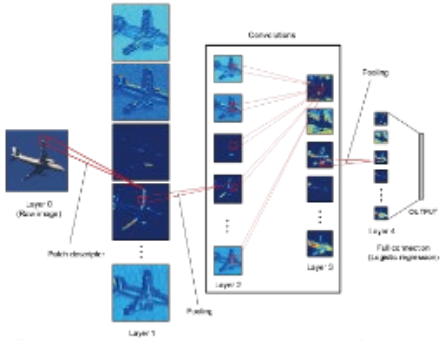
研究例：画像認識

Large-scale image tagging
ICPR'16, CVPR'10, ECCV'10

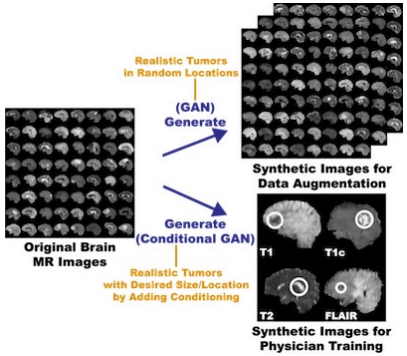


Fine-grained recognition
ICME'13, CLEF'13

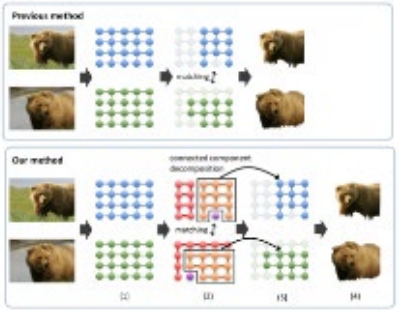
Scene text detection
ICDAR'17



Visual representation learning
BMVC'13

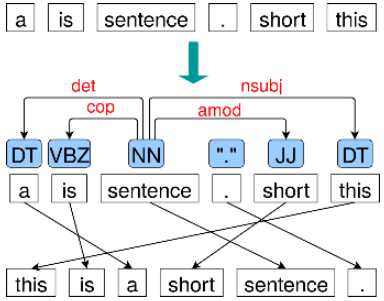


Medical image analysis
ISBI'18

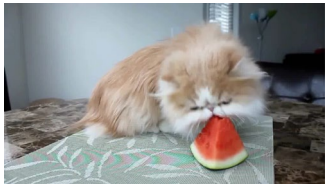


Object discovery
ACMMM'15,17

研究例：自然言語処理・マルチモーダル

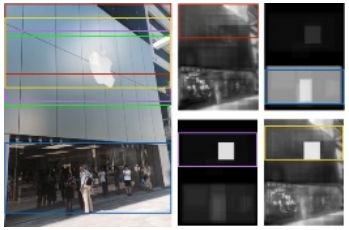


Word representation learning
ICLR'18, IJCNLP'17

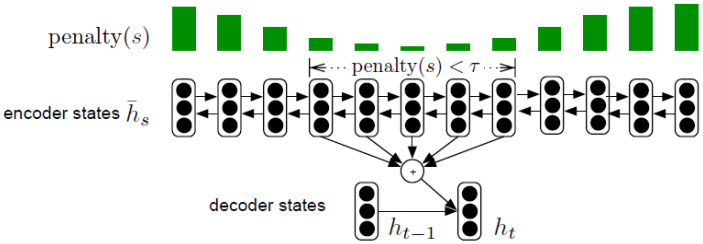


a cat is trying to eat the food

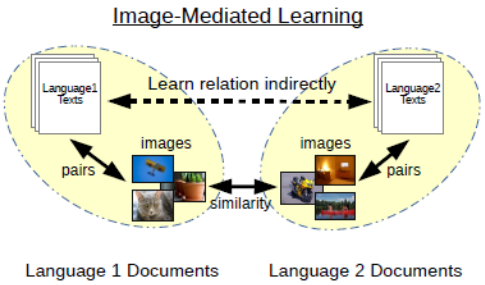
Image/video caption generation
LREC'18, COLING'16



Multimodal thumbnailing
WWW'16



Machine translation
MT'17, NMT@ACL'17, ACL'18



Cross-lingual retrieval
EMNLP'15

本講演のメッセージ

- 深層学習は精度向上だけが全てではない！
  - それぞれの分野で定番のニューラルネットワークが確立
    → 異なるモダリティのシームレスな接続が可能に

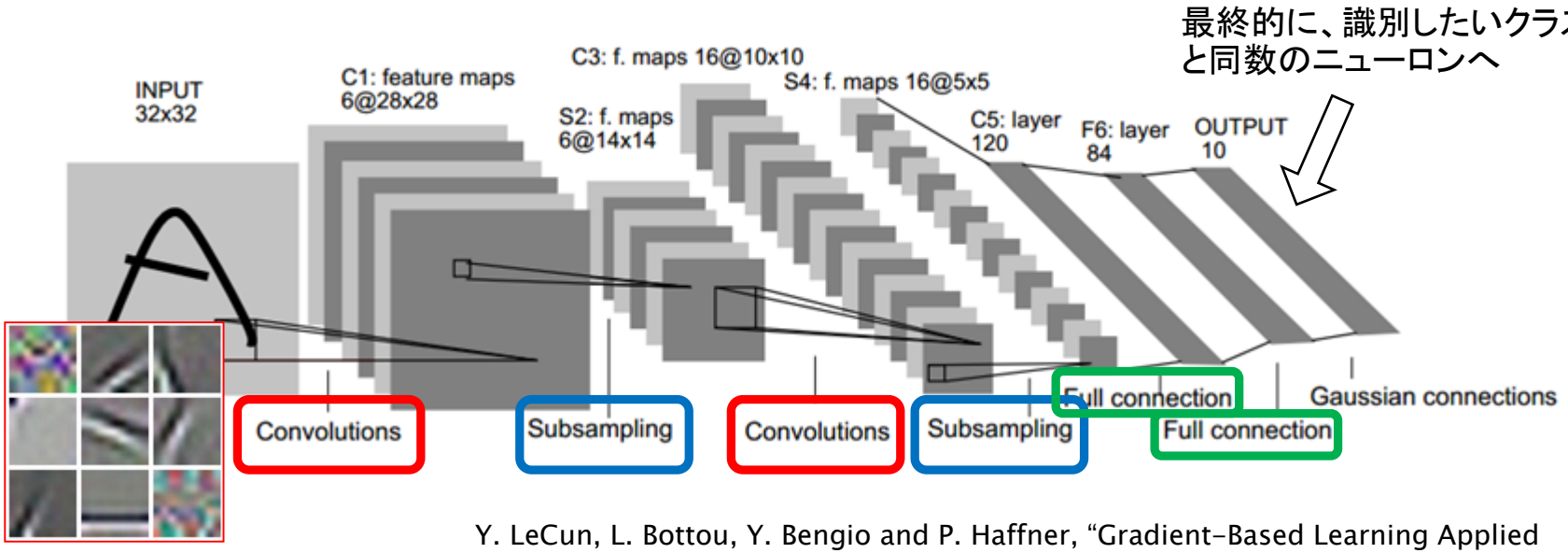
- 学際的な領域で面白い成果が次々に生まれている
  - 分野の垣根（参入障壁）が急速になくなりつつある

目次

- １．各分野における定番ネットワークの進化
- ２．マルチモーダル（クロスモーダル）深層学習
  - エンコーダ・デコーダモデルとマルチモーダル表現
  - One-to-one タスク
  - Many-to-one タスク
  - Many-to-many タスク
- ３．研究紹介
  - 画像を媒介としたゼロショット機械翻訳

画像認識に用いるネットワーク

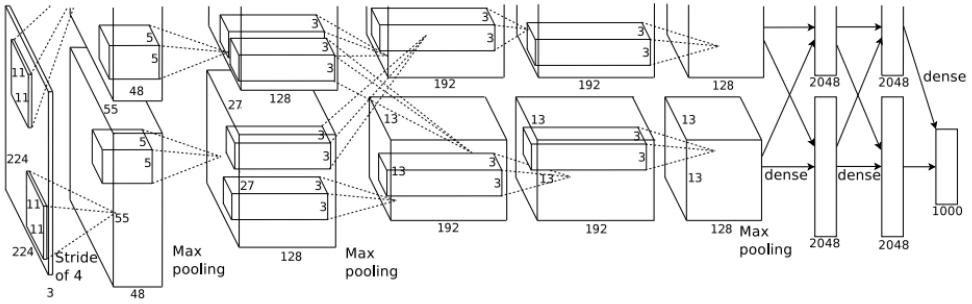
- 畳み込みニューラルネットワーク（CNN)
  - 局所領域（受容野）の畳み込みとプーリングを繰り返す多層パーセプトロン
  - V1視覚野に関する知見をもとに設計
  - 原形は日本初（福島邦彦先生、1980年代）



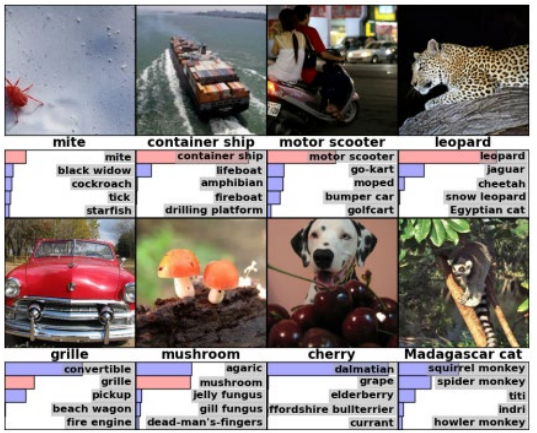
Y. LeCun, L. Bottou, Y. Bengio and P. Haffner, “Gradient-Based Learning Applied to Document Recognition”, Proceedings of the IEEE, 86(11):2278-2324, 1998.

ILSVRC 2012 での衝撃

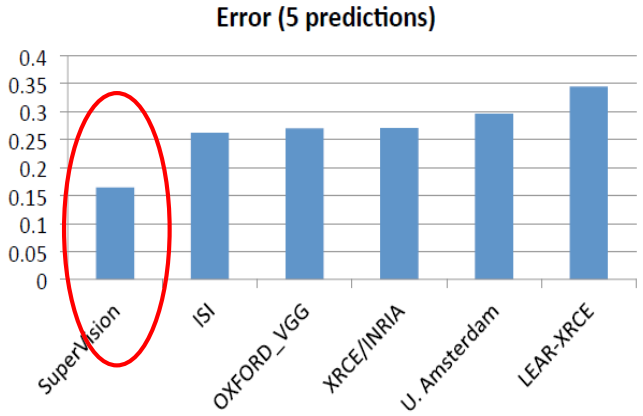
- 画像認識のコミュニティにおける中心的なコンペティション
- 1000クラス識別タスクで、CNNを用いたシステムが圧勝
  - トロント大学Hinton先生のグループ (AlexNet)



[A. Krizhevsky *et al.*, NIPS'12]



エラー率が一気に10%以上減少！
(※過去数年間での向上は1~2%)

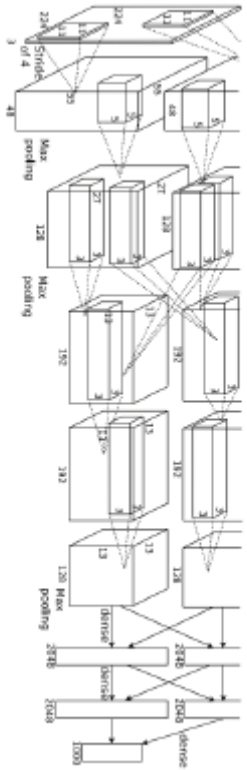


更に深く、広く…

2015 ResNet
(152層)

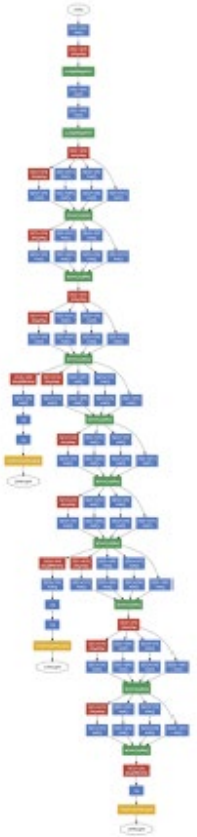
- 2012年以降劇的な向上が続いてきた

2012 AlexNet
(8層)



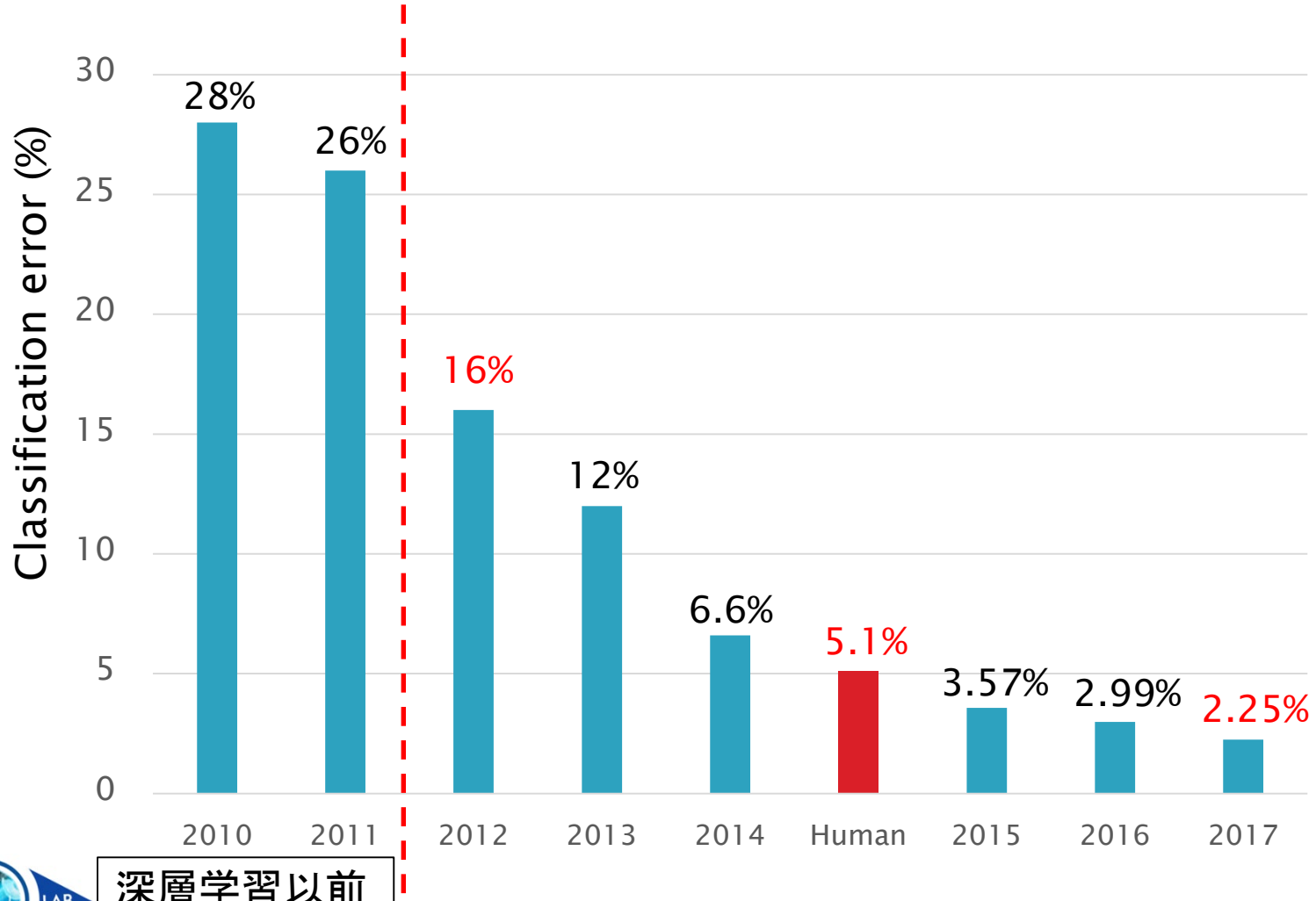
2014 VGG
(19層)

2014 GoogLeNet
(22層)



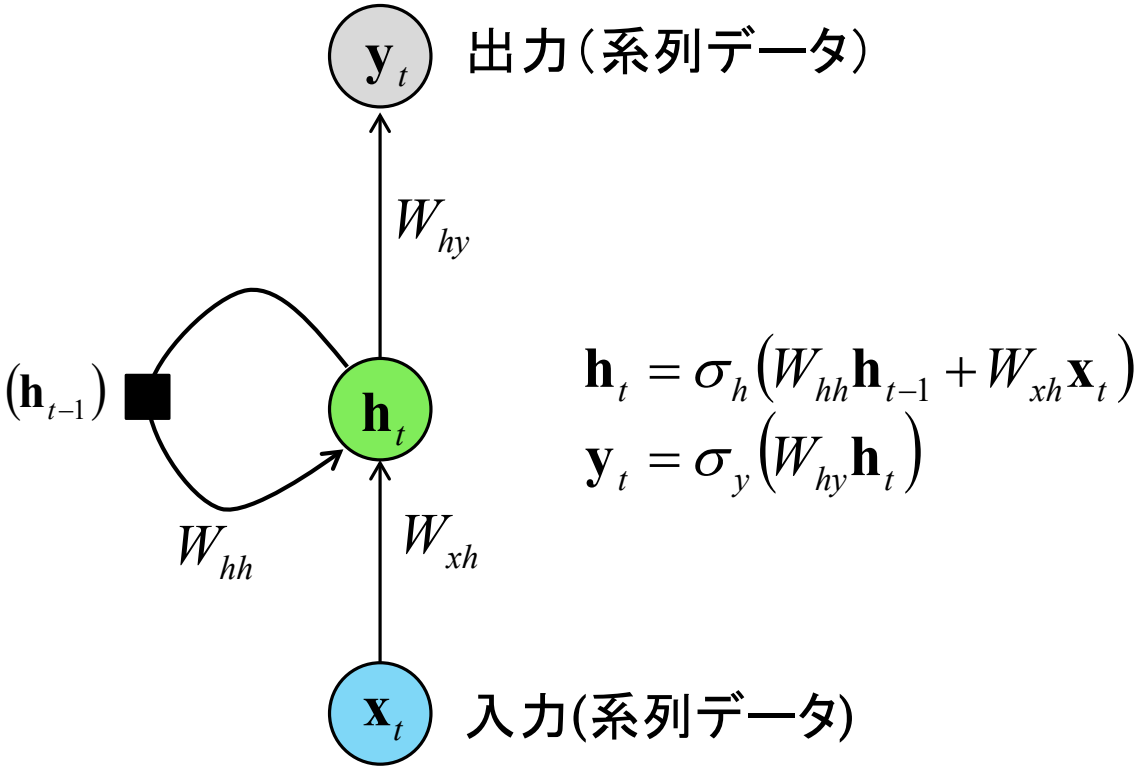
圧倒的な性能向上

- エラー率が 16% (2012) → 2.3% (2017)



NLPに用いるネットワークの例

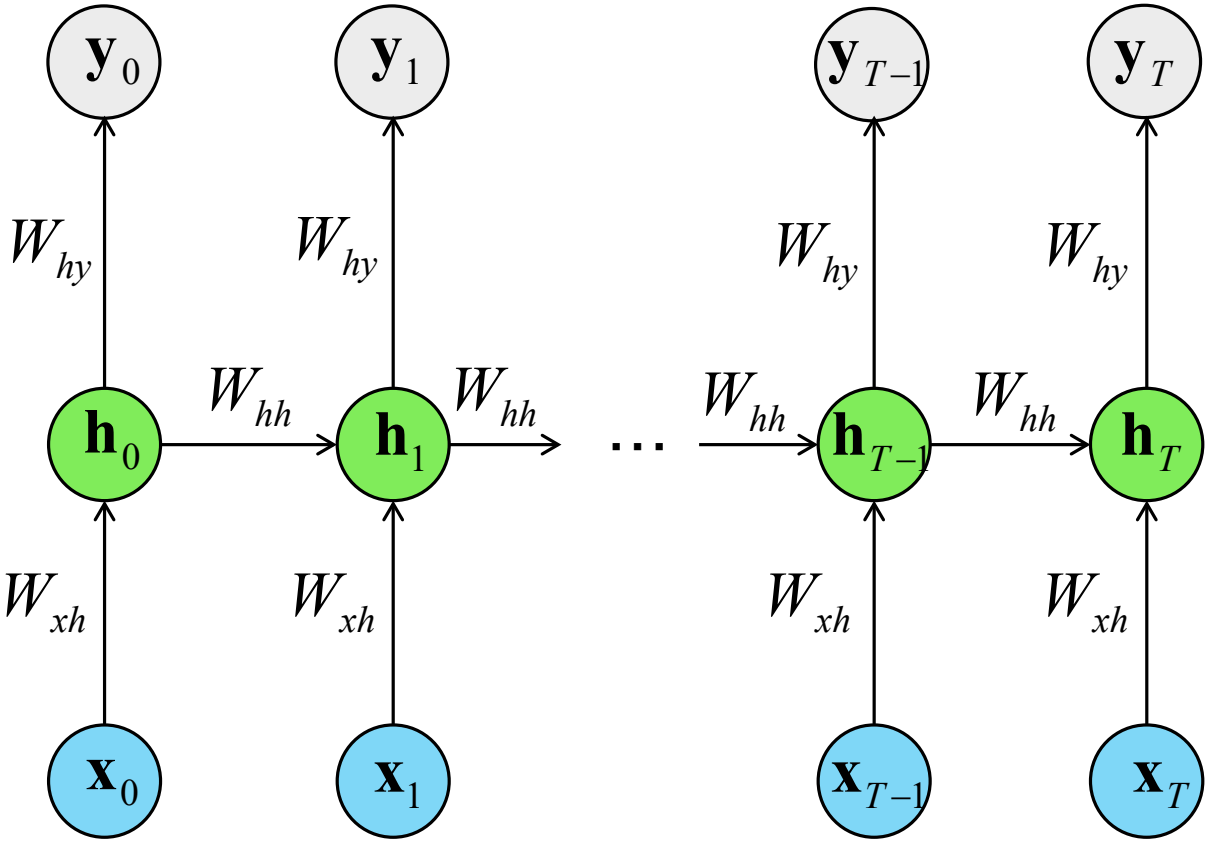
- Recurrent Neural Network (RNN)
  - 自分の一個前の隠れ状態を再入力するネットワーク
  - 隠れ状態は、入力系列の情報を記憶した分散表現（ベクトル表現）となる
  - 理論的には、任意のタイムスケールでの入出力依存関係を表現可能



$$\mathbf{h}_t = \sigma_h\left(W_{hh}\mathbf{h}_{t-1} + W_{xh}\mathbf{x}_t\right)$$
$$\mathbf{y}_t = \sigma_y\left(W_{hy}\mathbf{h}_t\right)$$

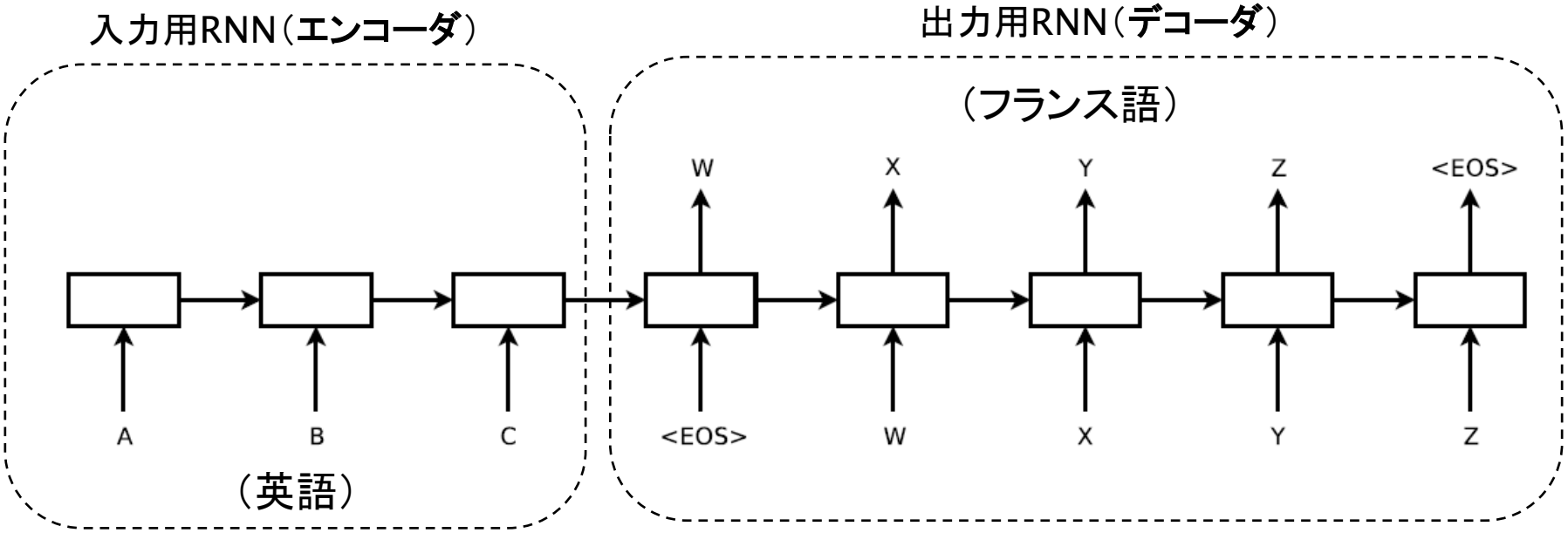
時間方向に展開してみると…

- 静的な(深い)ネットワークとして書ける
  - 普通のパーセプトロンと同様、誤差逆伝播による学習が可能
- 他の深層モデル同様、誤差消失により実際には遠い依存関係の学習が困難であったが、LSTM [Hochreiter+, 1997] により大幅な進展



RNNを用いた機械翻訳

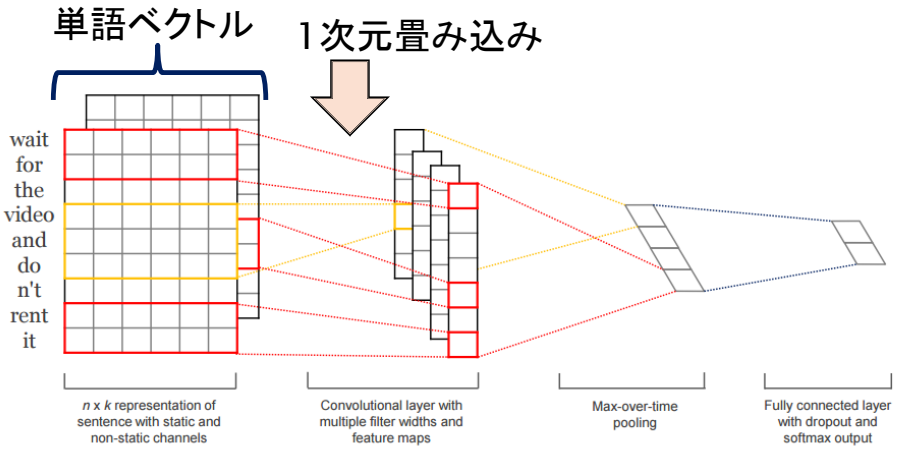
- Sequence to sequence [Sutskever+, NIPS'14]
  - 二つのRNN (LSTM) を接続し、英語・フランス語単語列の入出力関係を学習
  - 自然言語処理における深層学習の最初のブレークスルーの一つ



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

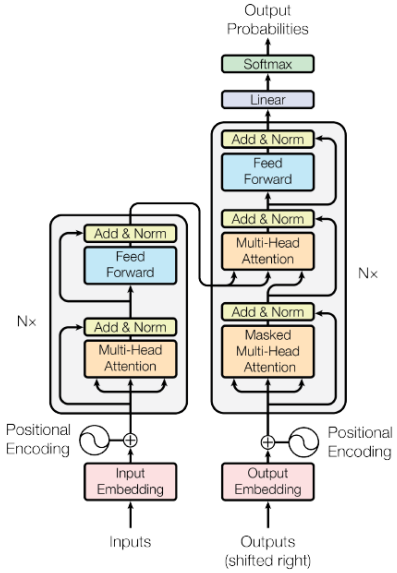
ネットワーク技術は日進月歩

- CNNもNLPでブームに
  - CNNは系列データ全般でかなり有効



Yoon Kim, “Convolutional Neural Networks for Sentence Classification”, In Proc. of EMNLP, 2014.

- 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
  - 学習済みモデル(BERT)が話題 [Devlin+, 2018]



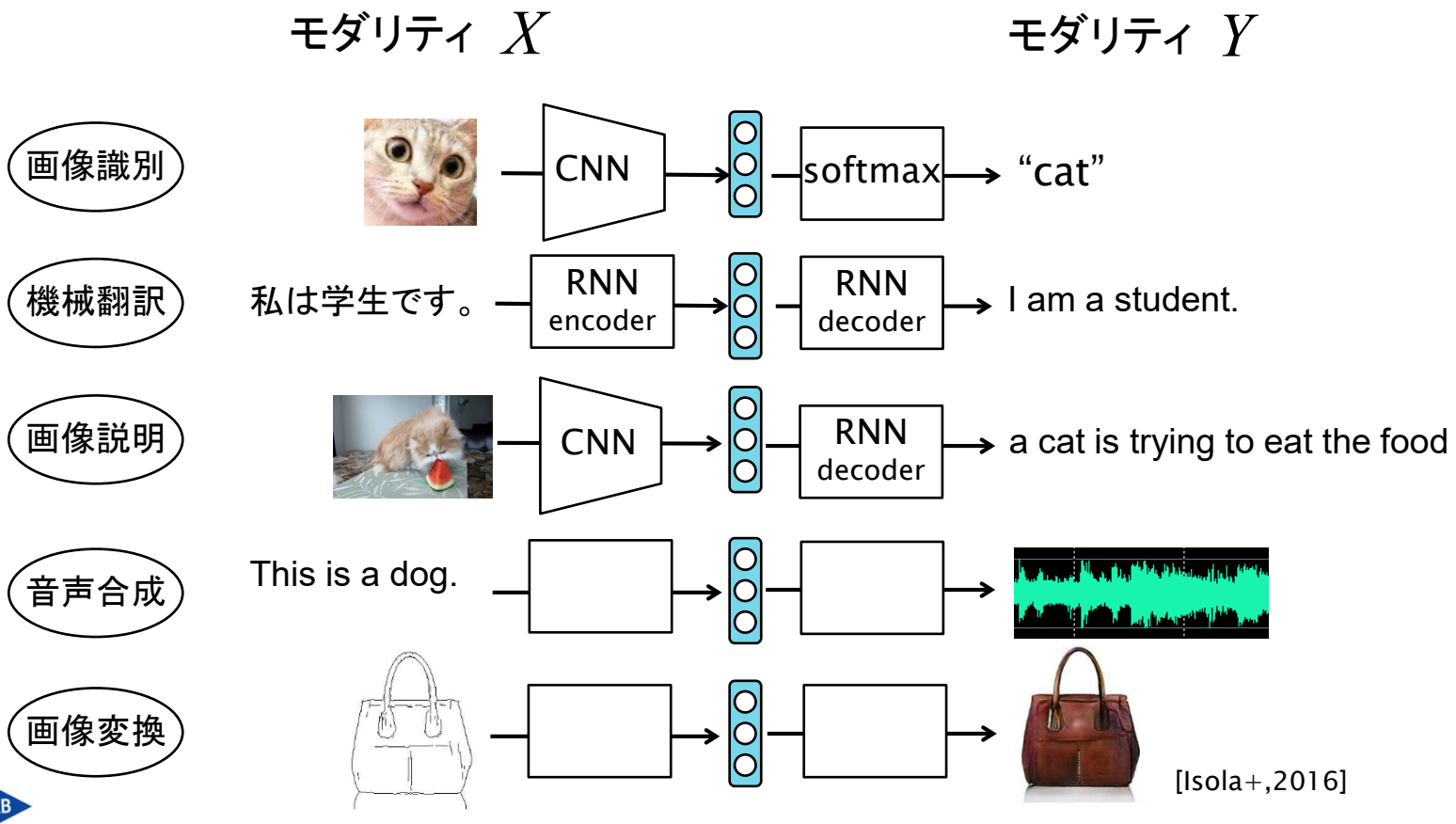
Vaswani et al., “Attention Is All You Need”, In Proc. of NIPS, 2017.

目次

- １．各分野における定番ネットワークの進化
- **２．マルチモーダル（クロスモーダル）深層学習**
  - エンコーダ・デコーダモデルとマルチモーダル表現
  - One-to-one タスク
  - Many-to-one タスク
  - Many-to-many タスク
- ３．研究紹介
  - 画像を媒介としたゼロショット機械翻訳

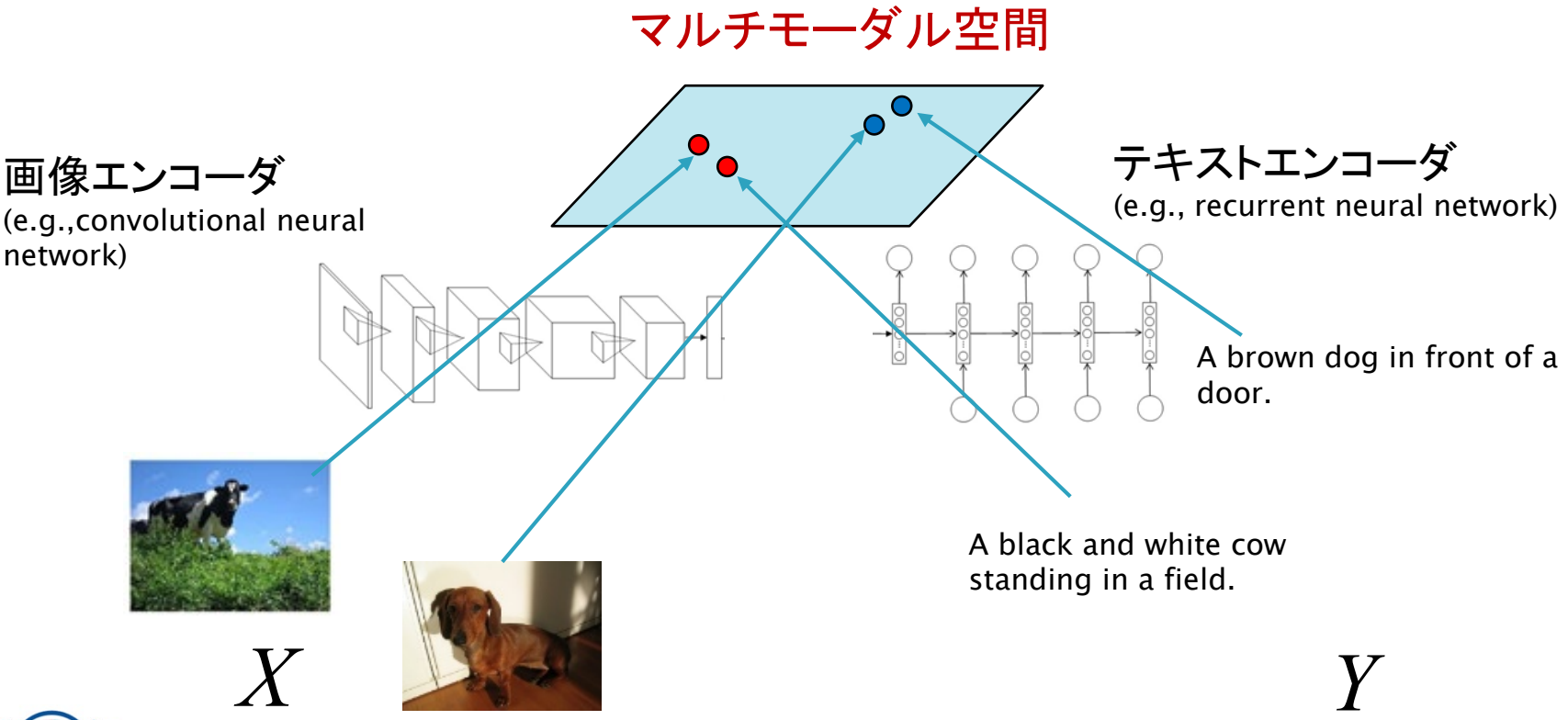
クロスモーダル技術の発展

- 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



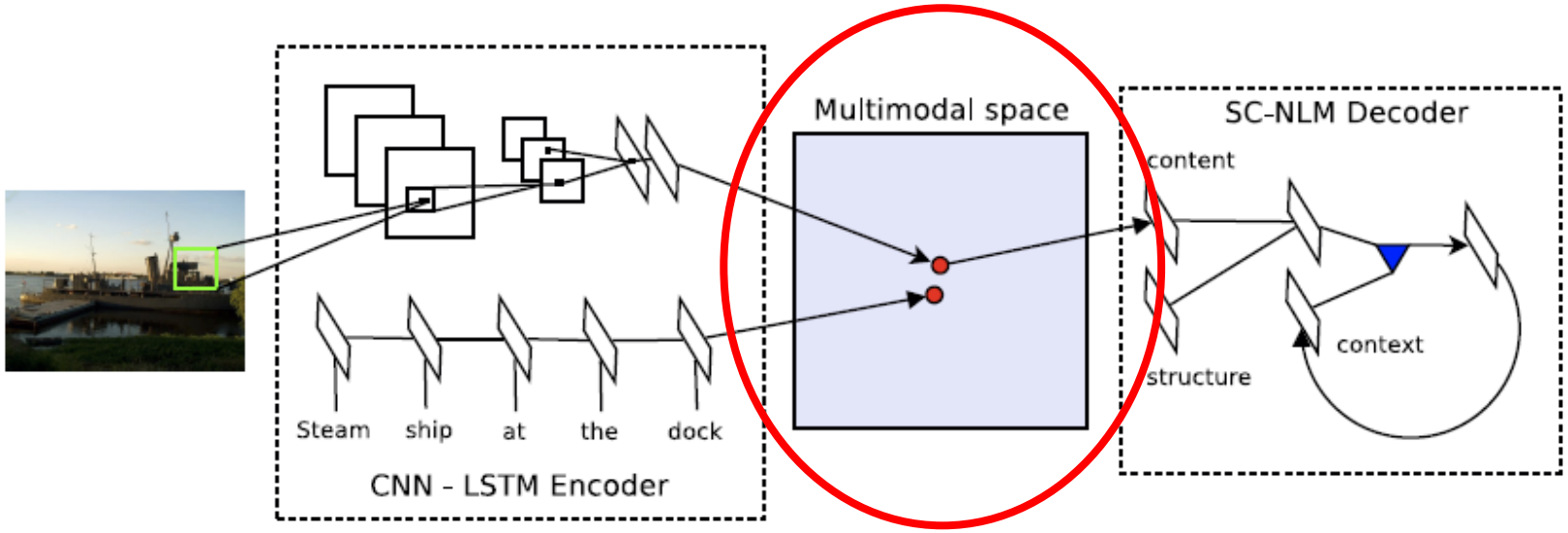
クロスモーダル表現学習

- 異なるモダリティに属するデータを共通の空間へ写像
  - 意味的にアラインメントされた表現が得られる

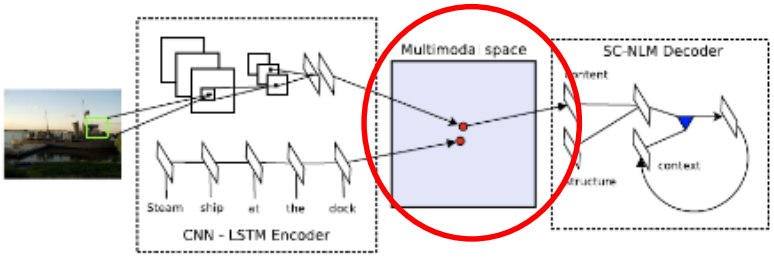


画像とテキストのマルチモーダル分散表現

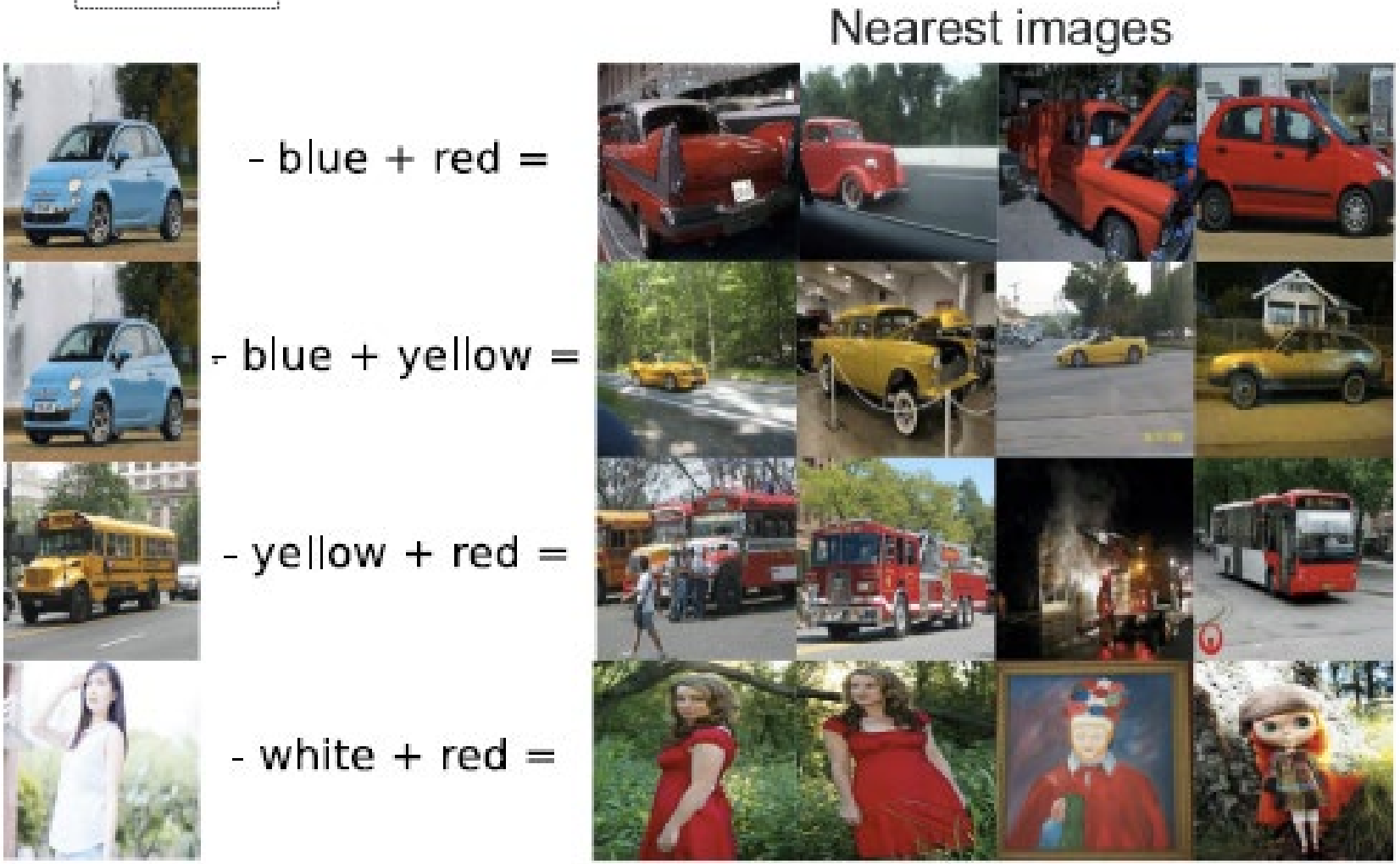
- 共通の潜在空間へマッピング [Kiros et al., 2014]
  - 異なるモダリティ間での“演算”が可能



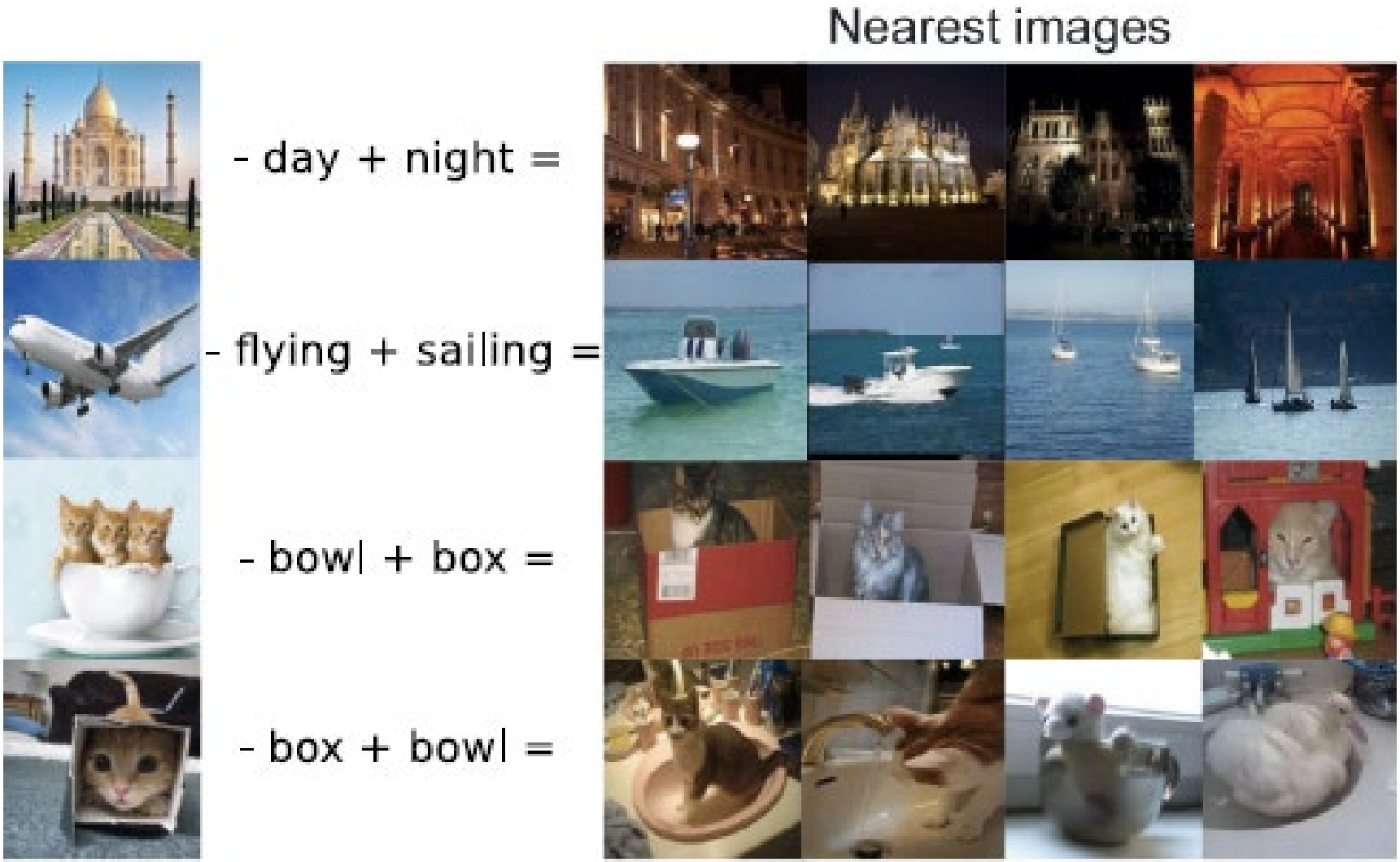
R. Kiros et al., “Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models”, TACL, 2015.



R. Kiros et al., "Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models", TACL, 2015.



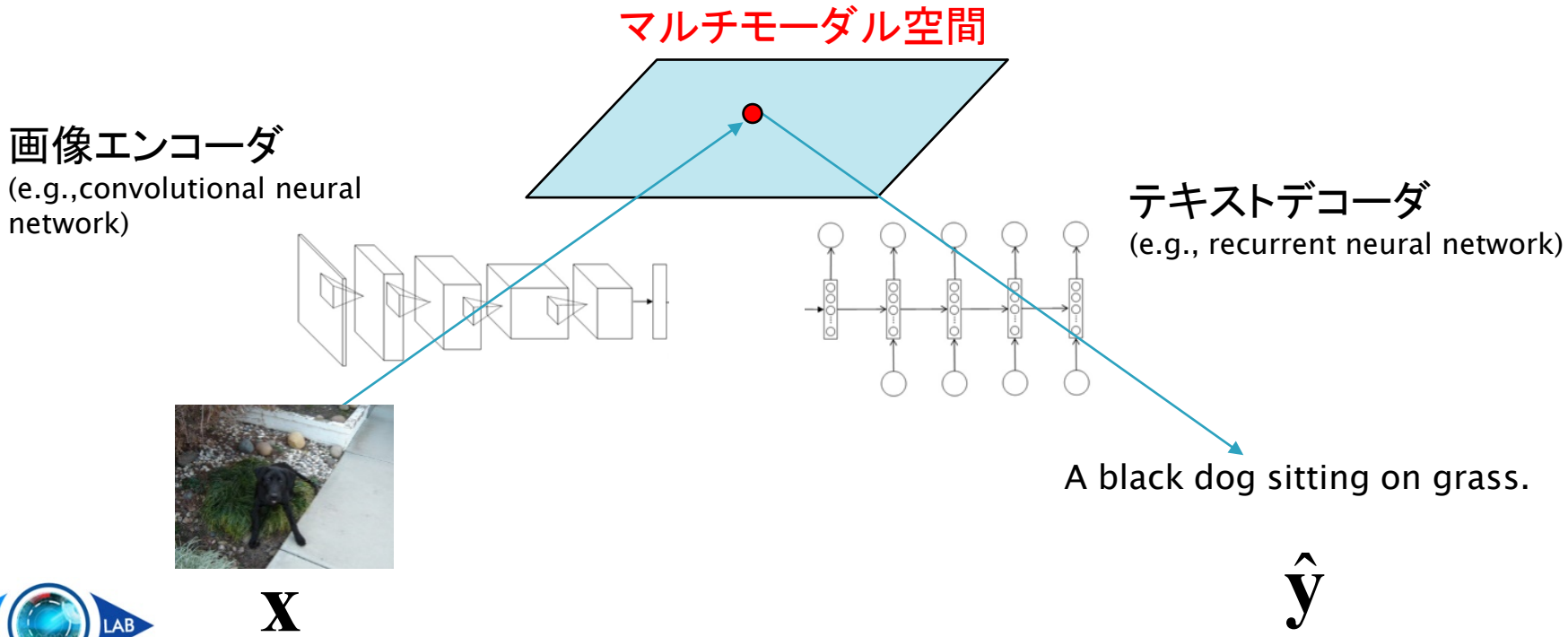
[Kiros et al., 2014]



[Kiros et al., 2014]

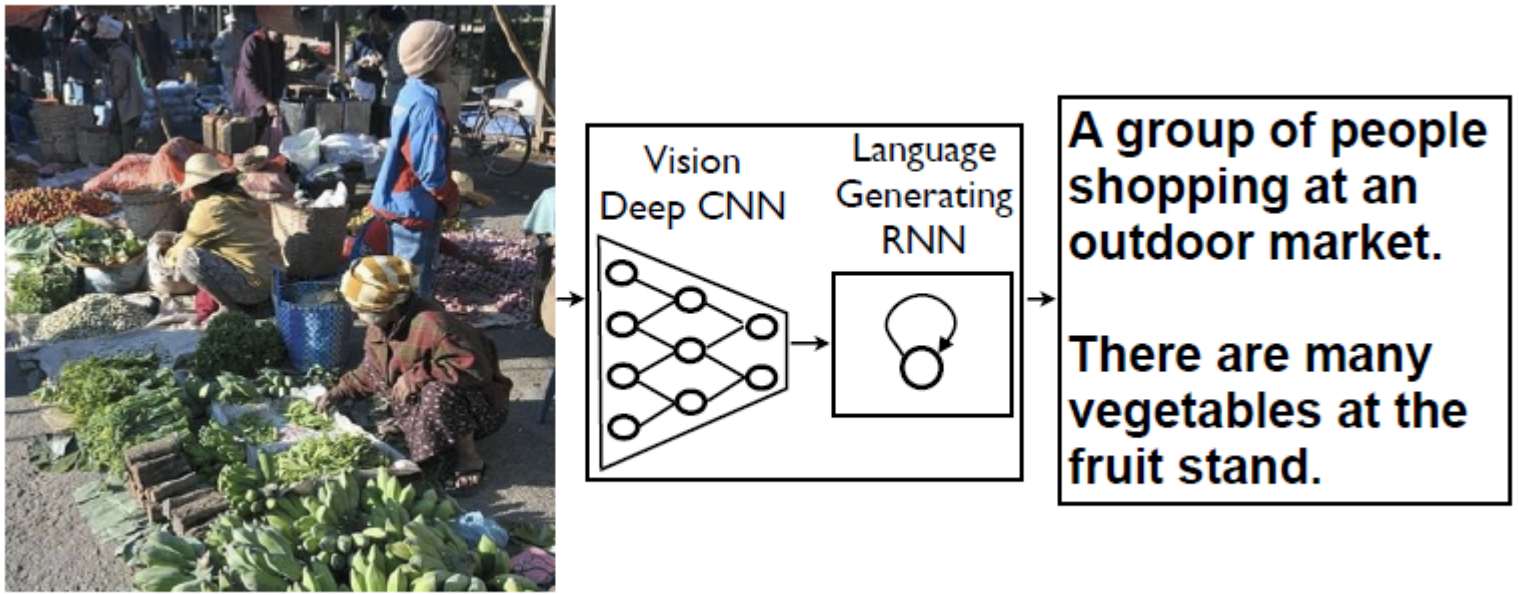
エンコーダ・デコーダモデル

- 入力をマルチモーダル表現へマッピングし、所望の出力形式へデコードする
- 誤差逆伝播法により、入力から出力へ至る全てのネットワークパラメータの最適化(一貫学習)を行うものが多い



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)



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

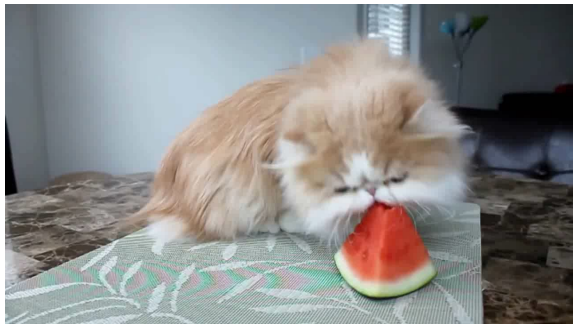
動画像キャプショニング [Laokulrat+, COLING'16]

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

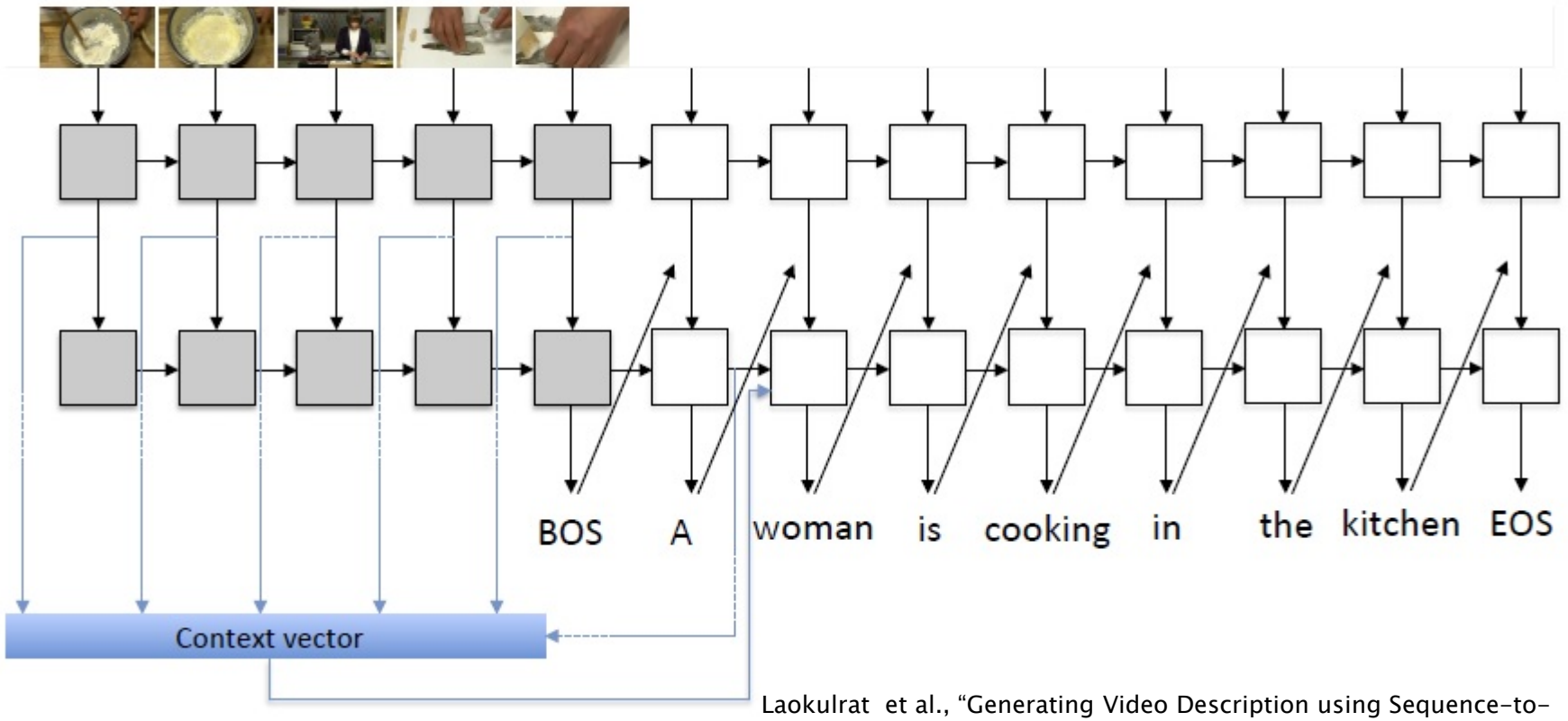
a dog is swimming in the pool

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

動画像キャプショニング [Laokulrat+, COLING'16]

産総研AIRCでの成果

- 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



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

文章からの画像生成

- 敵対的生成ネットワーク（GAN）の登場により、画像のデコーディングも飛躍的に進化

this small bird has a pink breast and crown, and black primaries and secondaries.

this magnificent fellow is almost all black with a red crest, and white cheek patch.

the flower has petals that are bright pinkish purple with white stigma

this white and yellow flower have thin white petals and a round yellow stamen

Reed et al., “Generative Adversarial Text to Image Synthesis”, In Proc. of ICML, 2016.

画像スタイル変換

- Image-to-image translation
  [Isola+, CVPR'16]

Isola et al., "Image-to-Image Translation with Conditional Adversarial Networks", In Proc. IEEE CVPR, 2017.

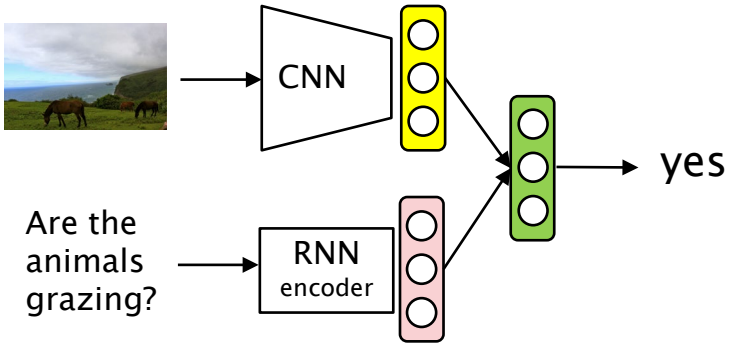
- Cycle GAN
  [Zhu+, ICCV'17]

Zhu et al., "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks", In Proc. IEEE ICCV, 2017.

時代はよりマルチモーダルへ

Many-to-one

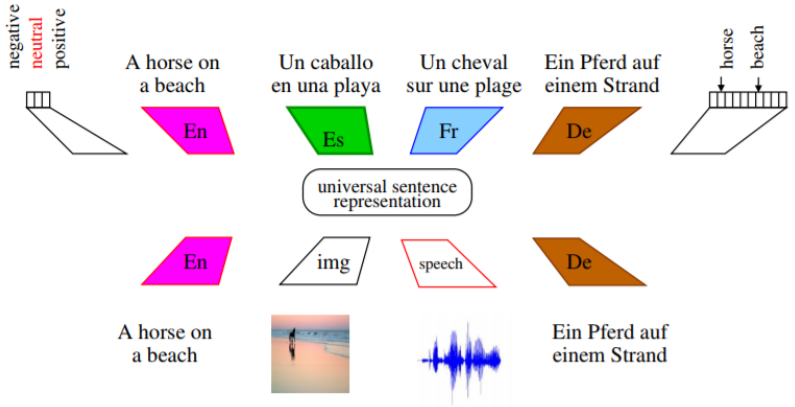
(e.g., 画像質問応答,
マルチセンサ識別)



- 認識精度・頑健性の向上
- 複数のモダリティを駆使した新規AIタスク

Many-to-many

(e.g., マルチモーダル機械翻訳)



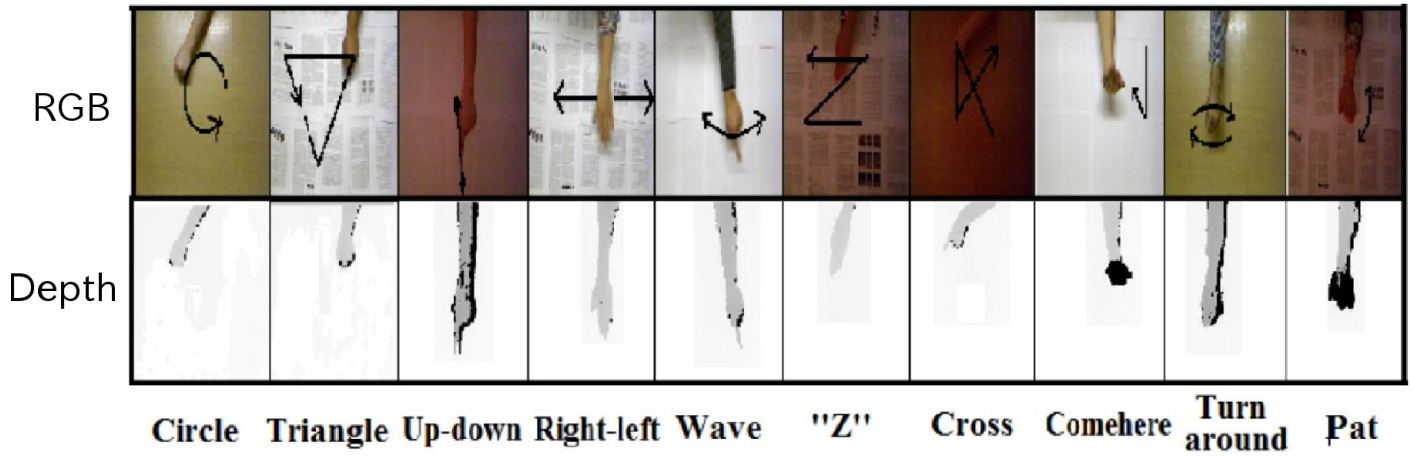
Schwenk and Douze, "Learning Joint Multilingual Sentence Representations with Neural Machine Translation", 2017.

マルチモーダルジェスチャー認識

[Nishida and Nakayama, 2015]

Nishida and Nakayama, “Multimodal gesture recognition using multi-stream recurrent neural network”, In Proc. of PSIVT, 2015.

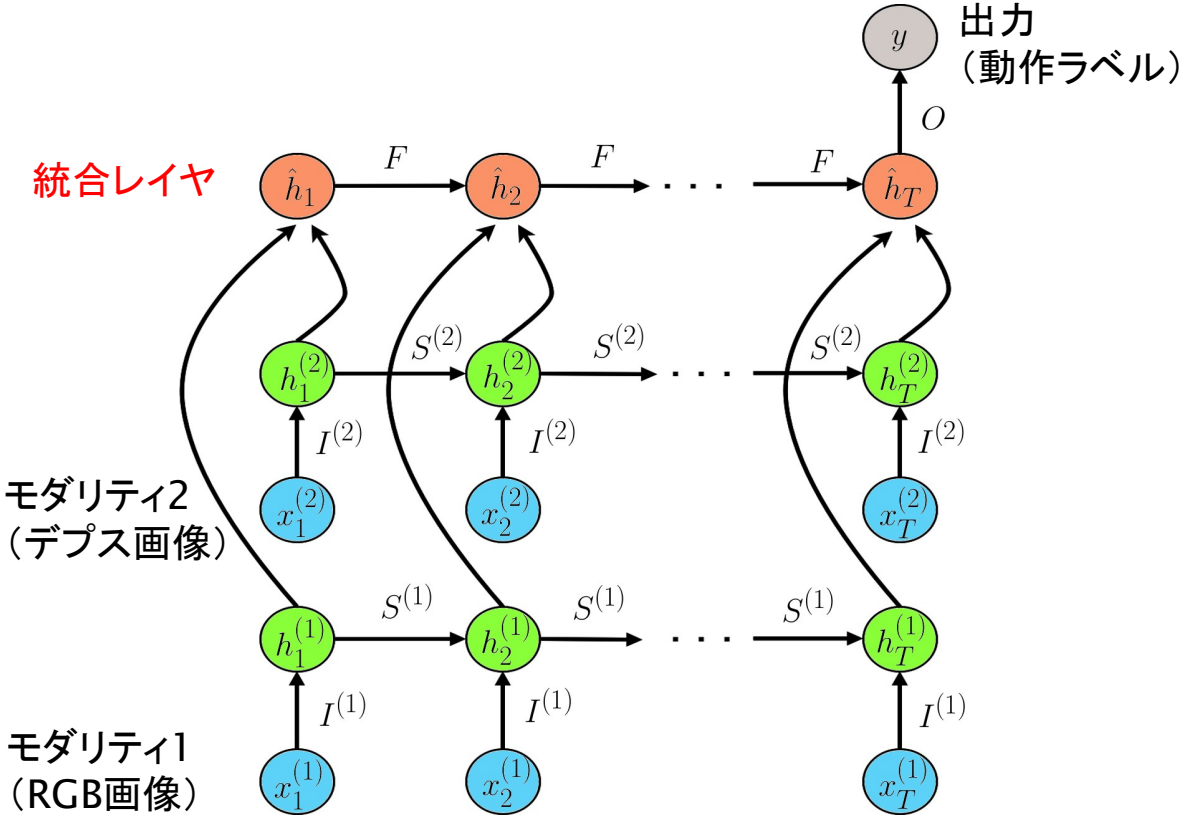
- Sheffield Kinect Gesture (SKIG) データセット [Liu et al., 2013]
  - キネクトで撮影された10クラスのジェスチャー動画
  - RGB画像とデプス画像の時系列データ
  - 本研究ではオプティカルフローも追加モダリティとして利用



- 興味
  - どのようにして複数のモダリティを束ねるべきか？
  - どのようにして時系列のダイナミクスを取り入れるべきか？

提案手法: Multi-stream RNN (MRNN)

- モダリティごとにRNNを用意し、各ステップで上位RNNに統合（＝**段階的に統合**）
  - 入力時点で結合するモデル（**Early fusion**)や出力で統合するモデル（**Late fusion**）よりも良好な性能



実験結果

- 複数モダリティにより認識精度が向上

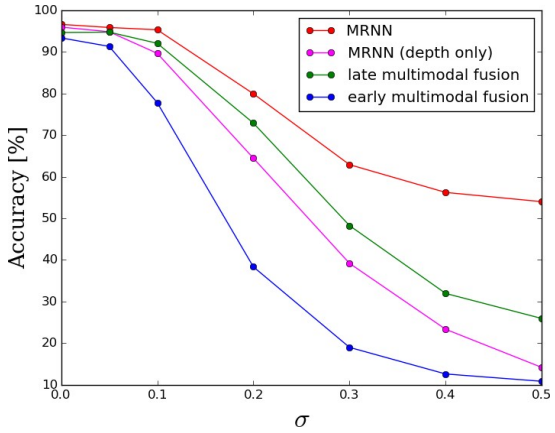
Table : Test accuracy (multiple modality vs. single modality)

Method	Accuracy (%)
MRNN (color)	91.6
MRNN (opt flow)	88.5
MRNN (depth)	95.9
MRNN (color + opt flow + depth)	**97.8**

Method	Accuracy (%)
Early multimodal fusion	94.1
Late multimodal fusion	94.6

かえって精度が落ちている…

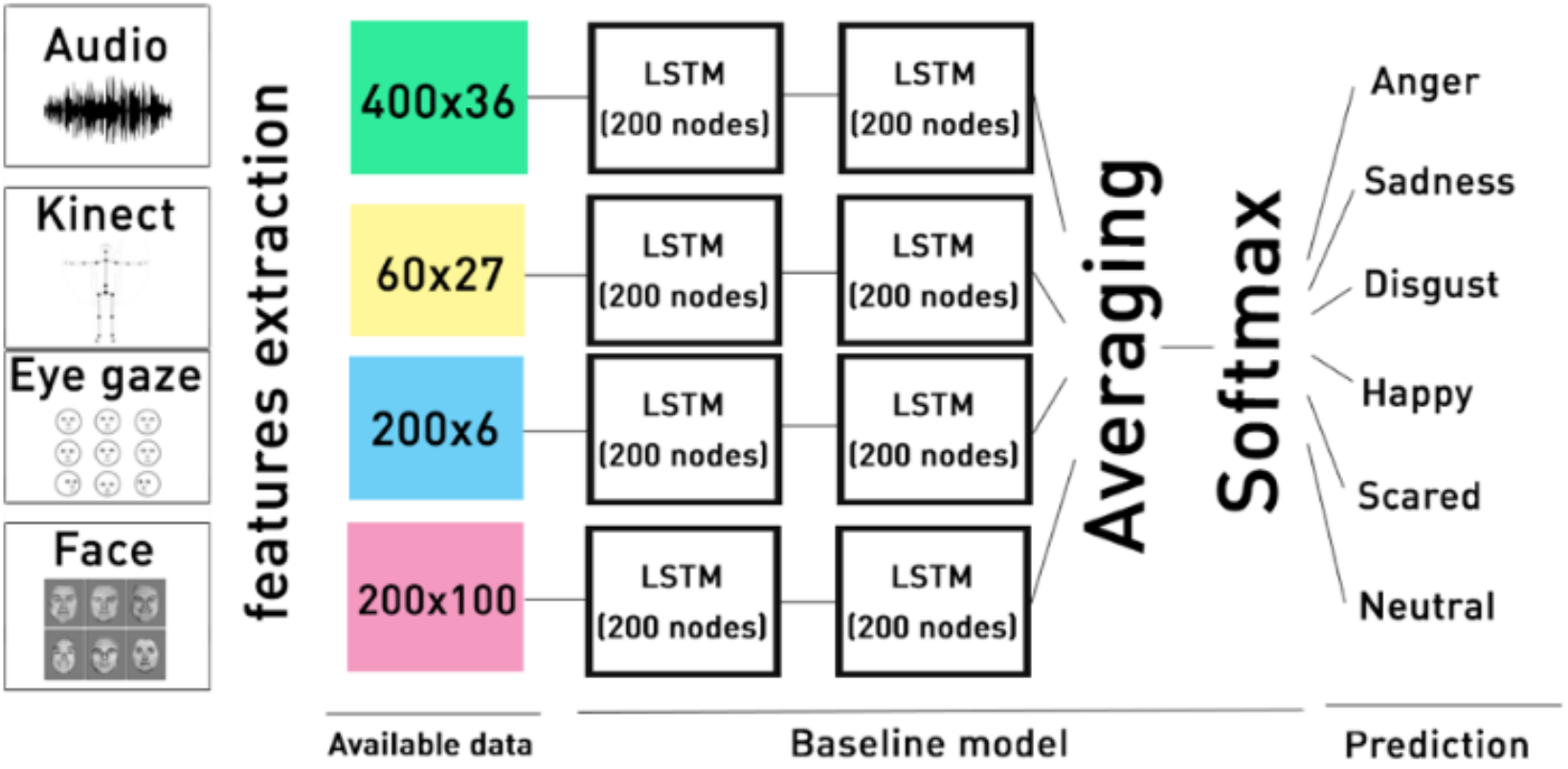
- ノイズ耐性
  - テスト時にデプス画像のみノイズを入れてみる
  - 提案手法は比較的頑健



教訓：モダリティの統合の仕方は重要！

マルチモーダル感情認識

- ユーザの感情状態を複数モダリティを活用して識別



Multimodal Emotion Recognition Challenge (MERC 2017)

http://www.datacombats.com/challenge/overview/

Visual Question Answering (VQA)

- RNN(LSTM)を用いた質問入力と回答の対応関係学習

H. Gao et al., “Are You Talking to a Machine? Dataset and Methods for Multilingual Image Question Answering”, In Proc. of NIPS, 2015.

Image					
Question	公共汽车是什么颜色的？ What is the color of the bus?	黄色的是什么？ What is there in yellow?	草地上除了人以外还有什么动物？ What is there on the grass, except the person?	猫咪在哪里？ Where is the kitty?	观察一下说出食物里任意一种蔬菜的名字？ Please look carefully and tell me what is the name of the vegetables in the plate?
Answer	公共汽车是红色的。 The bus is red.	香蕉。 Bananas.	羊。 Sheep.	在椅子上。 On the chair.	西兰花。 Broccoli.

M. Ren et al., “Exploring Models and Data for Image Question Answering”, In Proc. of NIPS, 2015.

CQ5429: What do two women hold with a picture on it?
Ground truth: cake
VIS+LSTM-2: cake (0.5611)
VIS+BOW: laptop (0.1443)
LSTM: umbrellas (0.1567)
BOW: phones (0.1447)

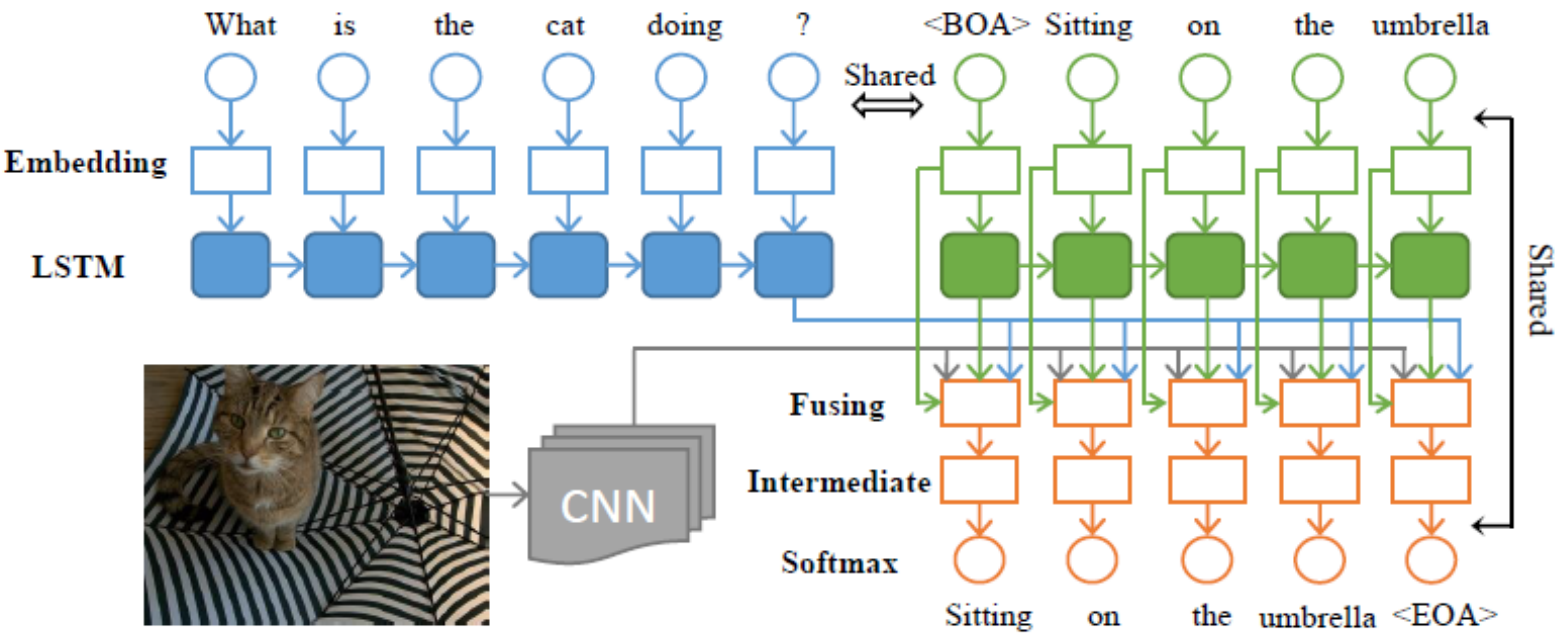
CQ24952: What is the black and white cat wearing?
Ground truth: hat
VIS+LSTM-2: hat (0.6349)
LSTM: tie (0.5821)

CQ25218: Where are the ripe bananas sitting?
Ground truth: basket
VIS+LSTM-2: basket (0.4965)
LSTM: bowl (0.6415)

CQ25218a: What are in the basket?
Ground truth: bananas
VIS+LSTM-2: bananas (0.6443)
LSTM: bears (0.0956)

Visual Question Answering (VQA)

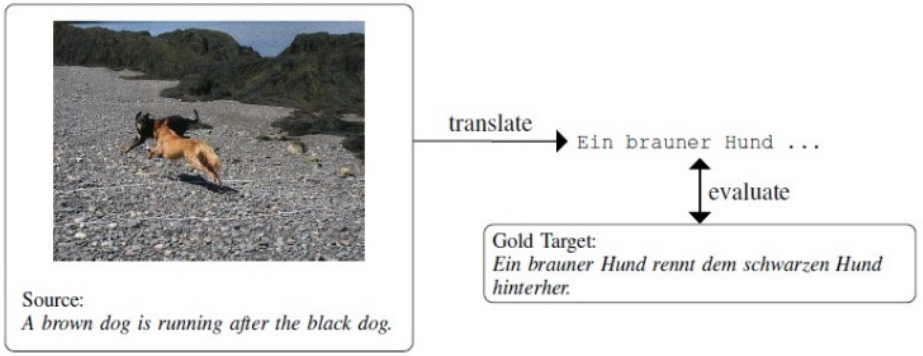
- NNを使った機械翻訳モデルの応用
- 質問文に加え、CNN対象画像の特徴抽出を行い、回答文生成のRNNへ入力



H. Gao et al., "Are You Talking to a Machine? Dataset and Methods for Multilingual Image Question Answering", In Proc. of NIPS, 2015.

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.

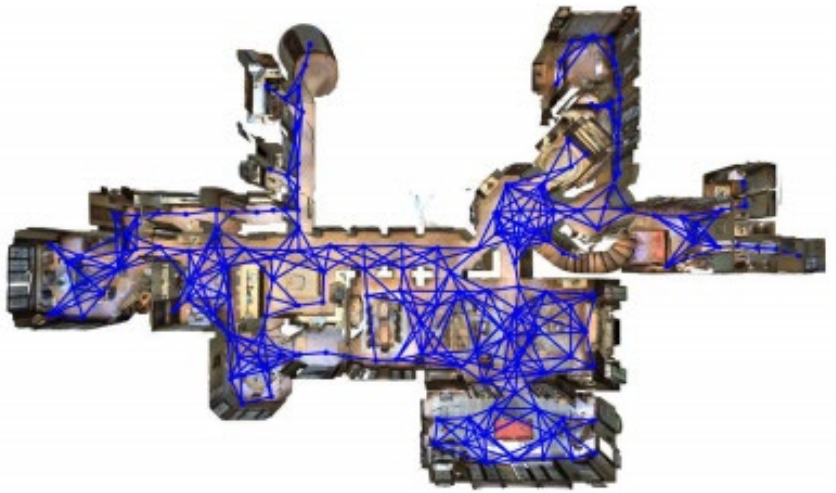
Mostafazadeh et al., "Image-Grounded Conversations: Multimodal Context for Natural Question and Response Generation", In Proc. of IJCNLP, 2017.

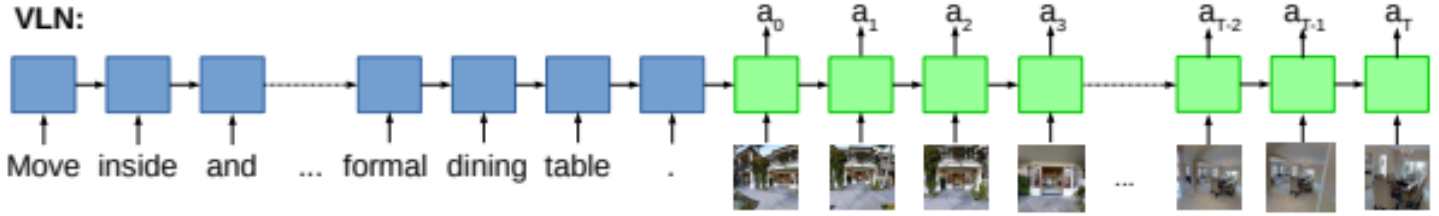
その他の言語+画像タスクの例

- Vision-and-Language Navigation [Anderson+, 2018]
  - 自然言語でロボットを目的地へ誘導
  - とるべき行動の系列を強化学習で生成

Anderson et al., "Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real environments", In Proc. of CVPR, 2018.

Instruction: Head upstairs and walk past the piano through an archway directly in front. Turn right when the hallway ends at pictures and table. Wait by the moose antlers hanging on the wall.



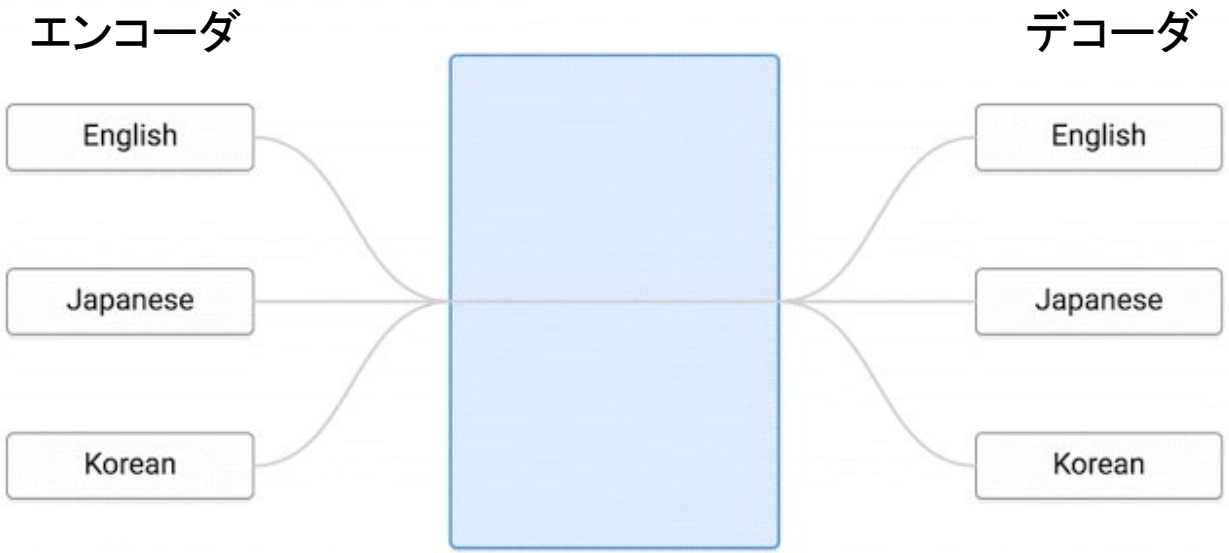


個人的雑感

- マルチモーダルの本質的な面白さはなんだろうか？
  - 入力が増えているのだから性能向上は当たり前？
- それなりに新しいことができるようになったが、
  結局は従来的な教師付き機械学習（が多い）
  - 学習時・推論時に、常に全てのモダリティが揃っていることを前提
  - 解けそうなタスクを見つけて、データセットを作るルーチンワーク

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

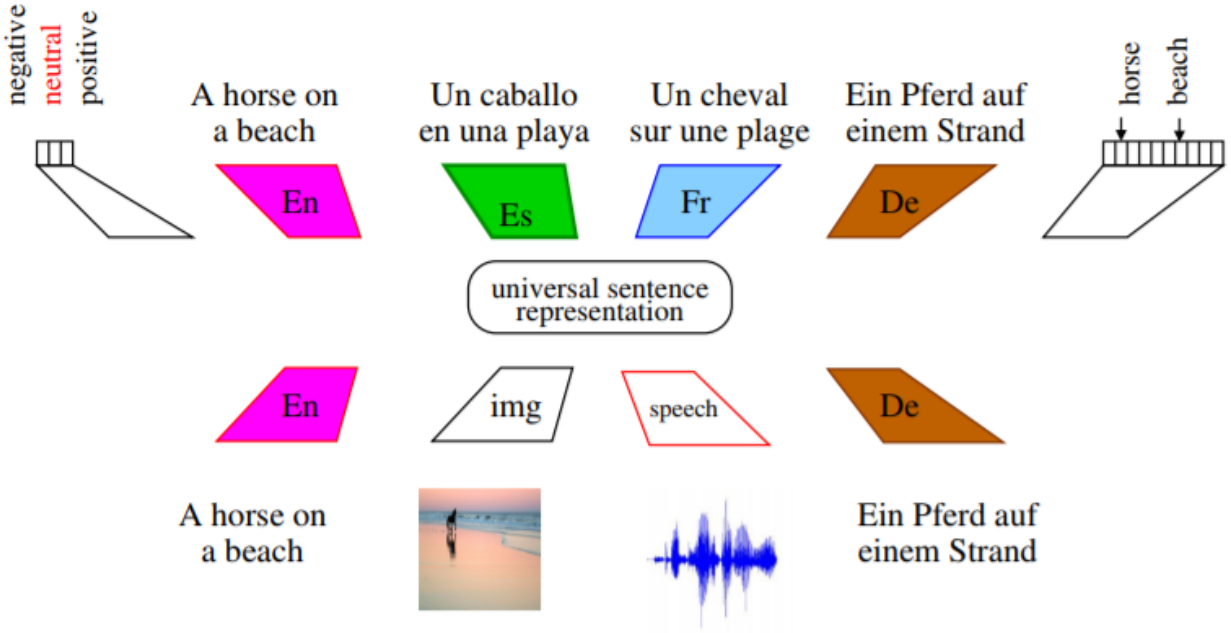


https://ai.googleblog.com/2016/11/zero-shot-translation-with-googles.html

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



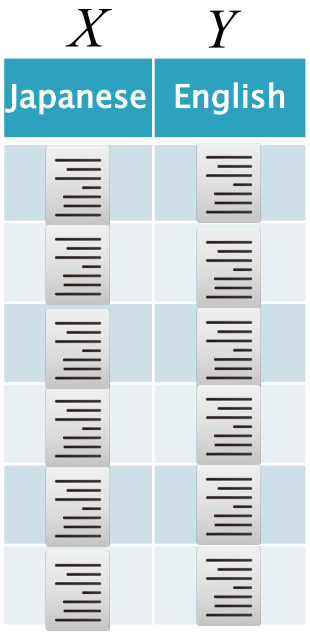
Schwenk and Douze, “Learning Joint Multilingual Sentence Representations with Neural Machine Translation”, In Proc. of RepL4NLP, 2017.

目次

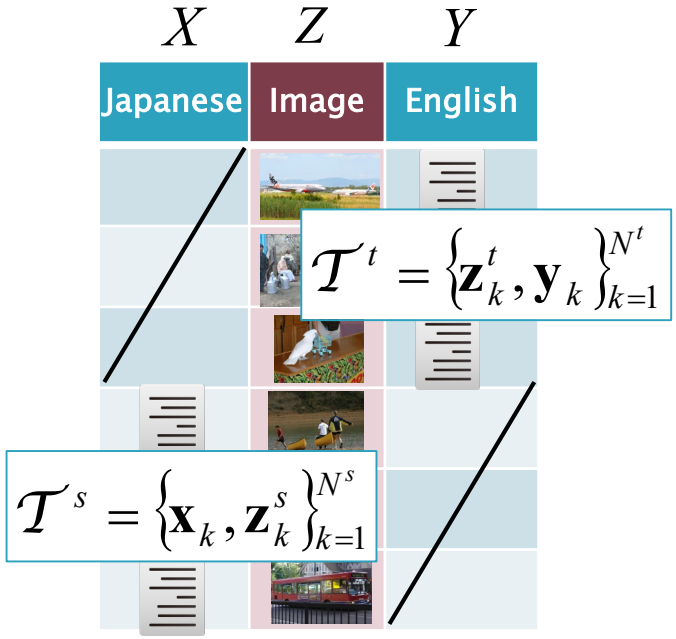
- １．各分野における定番ネットワークの進化
- ２．マルチモーダル（クロスモーダル）深層学習
  - エンコーダ・デコーダモデルとマルチモーダル表現
  - One-to-one タスク
  - Many-to-one タスク
  - Many-to-many タスク
- **３．研究紹介**
  - **画像を媒介としたゼロショット機械翻訳**

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）
  - 大規模なパラレルコーパス
    が必要



- 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



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

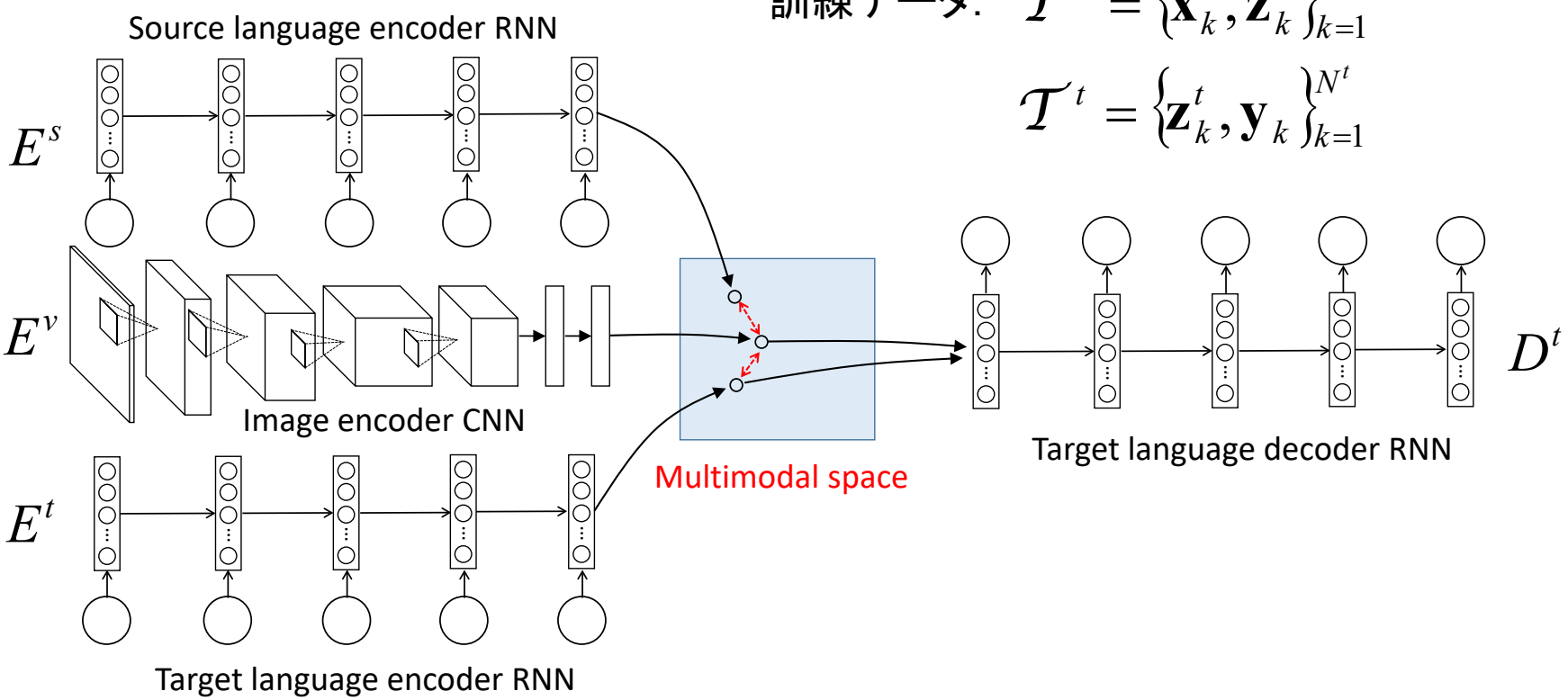
Funaki and Nakayama, “Image-mediated learning for zero-shot cross-lingual document retrieval”, In Proc. of EMNLP, 2015.

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

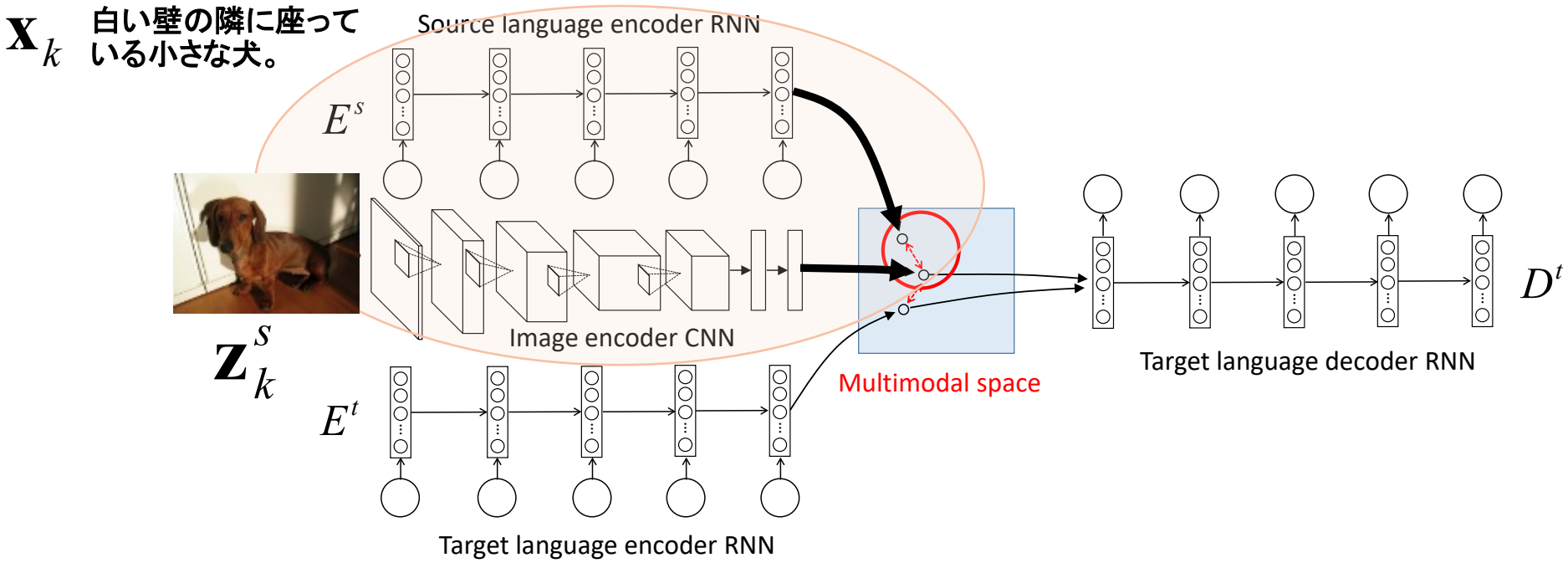
訓練データ: $\mathcal{T}^s = \{\mathbf{x}_k, \mathbf{z}_k^s\}_{k=1}^{N^s}$

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



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。

$\mathbf{z}_k^s$

Source language en

E^s

Image enc

E^t

Target language encoder RNN

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

Pair-wise Rank Loss
[Frome+, NIPS'13]

$s(\ )$: Similarity score function

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

Margin (Hyper parameter)

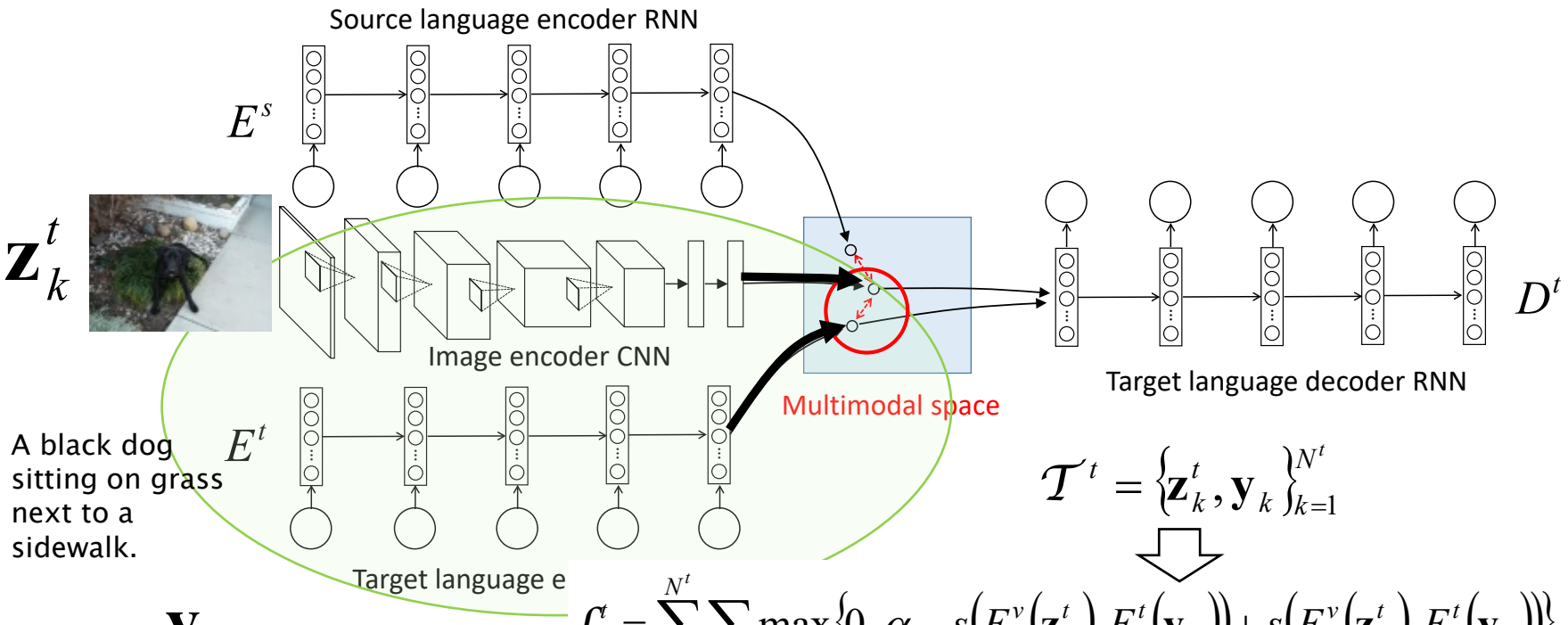
An image

Paired text

Negative (not paired) text

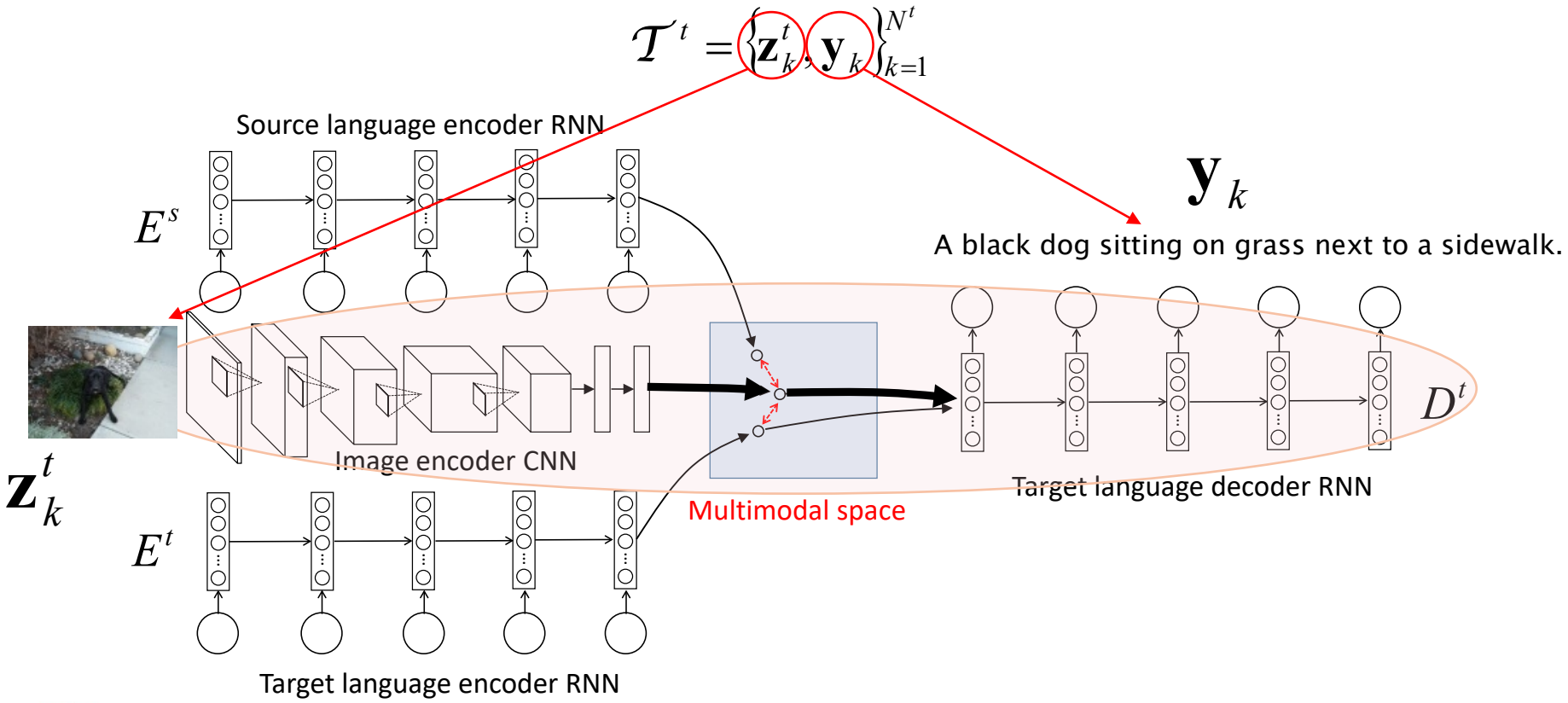
目的関数 2

- 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



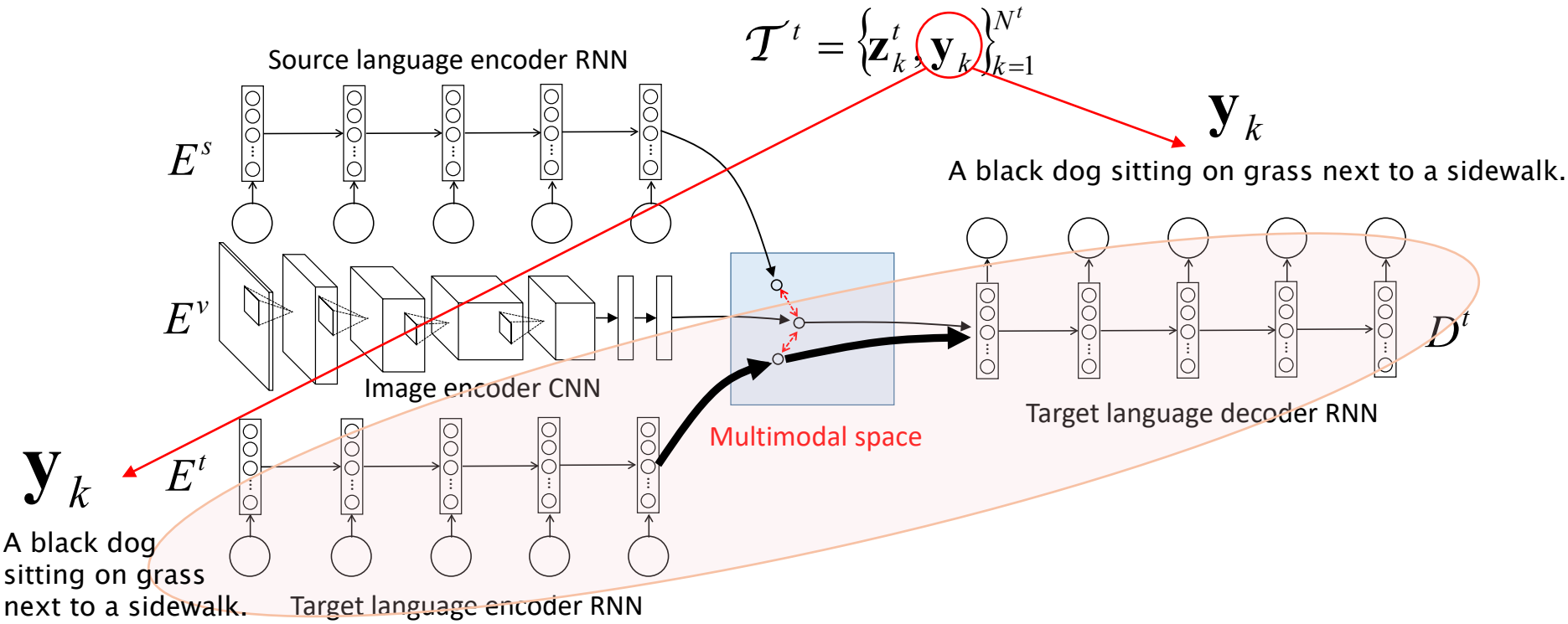
目的関数３

- 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



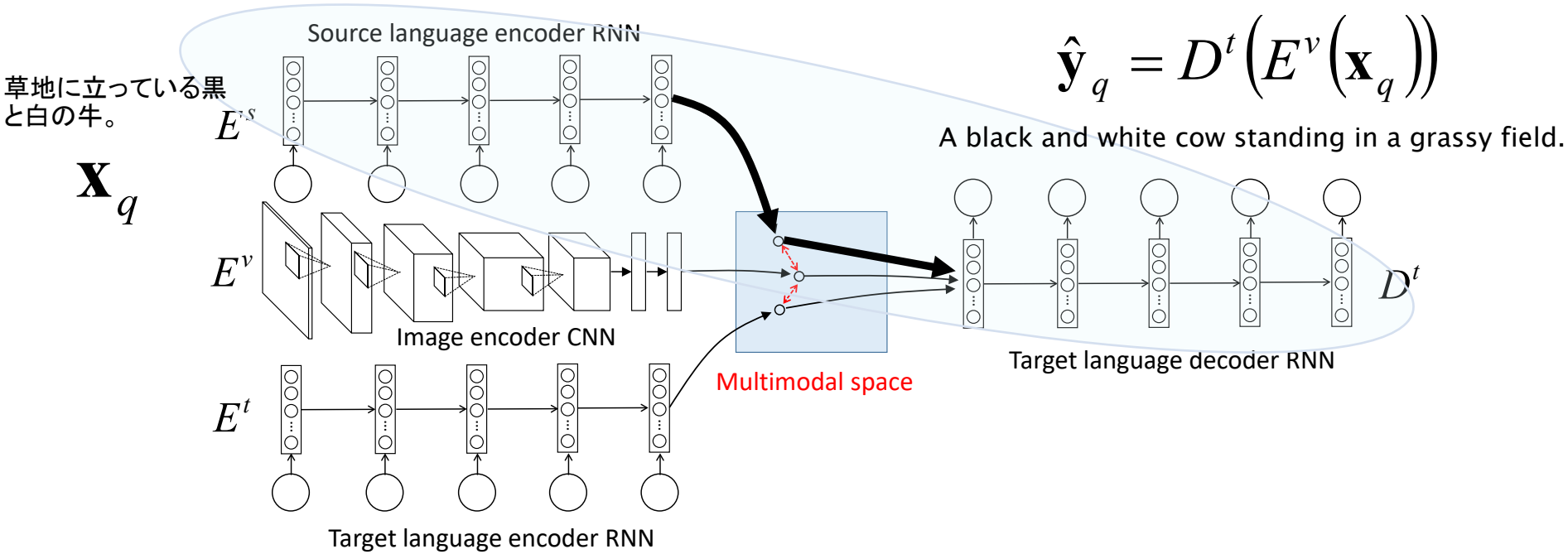
目的関数４

- 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)	18.9 (3.8)
2000	29.2 (22.6)	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;)

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
  - 分野間の障壁がなくなり、さまざまなタスクやアプローチが登場
- マルチモーダルのご利益
  - 精度・頑健性の向上
  - 新しいアプリケーション
  - 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