Unsupervised Methods for Learning Vector Representations of Sentences

Shuai Tang

Graduate Student in Cognitive Science Department

Committee Members

- Virginia R. de Sa, CogSci
- Benjamin K. Bergen, CogSci
- Eran A. Mukamel, CogSci
- Lawrence K. Saul, CSE
- Ndapandula Nakashole, CSE
- Richard Zemel, CS, University of Toronto

Outline

- Motivation
- Evaluation tasks
- Related work
- Our previous and ongoing work
- Future work

Outline

- **Motivation**
- Evaluation tasks
- Related work
- Our previous and ongoing work
- Future work

Why?

Motivation

- Sentences → Vectors
- Denotational vs. Distributional
- Localist vs. Distributed
- Supervised vs. Unsupervised

Sentences $\longrightarrow$ Vectors

We communicate in sentences,
and they convey our thoughts.

Sentences → Vectors

If we convert a sentence into a vector that **captures the meaning** of the sentence, then Google can do much better searches; they can search based on what's being said in a document. (Hinton, 2015)

Natural Reasoning

Motivation

- Sentences → Vectors
- Denotational vs. **Distributional**
- Localist vs. Distributed
- Supervised vs. Unsupervised

Distributional Hypothesis
Distributional Similarity
(Harris, 1954; Firth, 1957)

"You shall know a word by the company it keeps."

Motivation

- Sentences → Vectors
- Denotational vs. **Distributional**
- Localist vs. **Distributed**
- Supervised vs. Unsupervised

Localist Representations → Distributed Representations

The simplest way to represent things with neural networks is to dedicate **one** neuron to **each** thing.

One-hot Encoding

Clustering

Each concept is represented by **many** neurons, and **each** neuron participates in the representation of **many** concepts.

Continuous bag-of-words

Recurrent Neural Networks

Localist Representations → Distributed Representations

Efficient usage of space.
Better at capturing componential structure in data.

Motivation

- Sentences → Vectors
- Denotational vs. **Distributional**
- Localist vs. **Distributed**
- Supervised vs. **Unsupervised**

Outline

- Motivation
- **Evaluation tasks**
- Related work
- Our previous and ongoing work
- Future work

How to evaluate?

Evaluation Tasks

- Supervised Evaluation
  - A ***linear/non-linear model*** needs to be trained on top of the learnt vector representations.

- Unsupervised Evaluation
  - The similarity of two sentences is determined by the ***cosine similarity*** of two vector representations.

Evaluation Tasks

- Supervised Evaluation (13 tasks)
  - Sentiment Analysis
  - Paraphrase Detection
  - Caption-Image Retrieval
  - Semantic Relatedness Scoring
  - Natural Language Inference

- Unsupervised Evaluation (6 tasks)
  - Semantic Textual Similarity

Facebook AI Research, *https://github.com/facebookresearch/SentEval* ,2017

Our concerns …

- Coverage and Consistency
    - Coverage
    - Internal Bias
    - Internal Consistency
    - Linguistic Features
- Machine Learning Ethics
    - Overfitting

(Hill et al., NAACL2016; Conneau et al., ACL2018; Faruqui et al., RepEval@ACL2016)

Our concerns …

- Coverage and Consistency
  - Coverage
  - Internal Bias
  - Internal Consistency
  - Linguistic Features
- Generalisation
  - Choose the hyperparameters on the averaged performance on a small subset of the evaluation tasks
  - Choose the hyperparameters that lead to the best averaged performance across all tasks

(Hill et al., NAACL2016; Conneau et al., ACL2018; Faruqui et al., RepEval@ACL2016)

Outline

- Motivation
- Evaluation tasks
- **Related work**
- Our previous and ongoing work
- Future work

Related Work

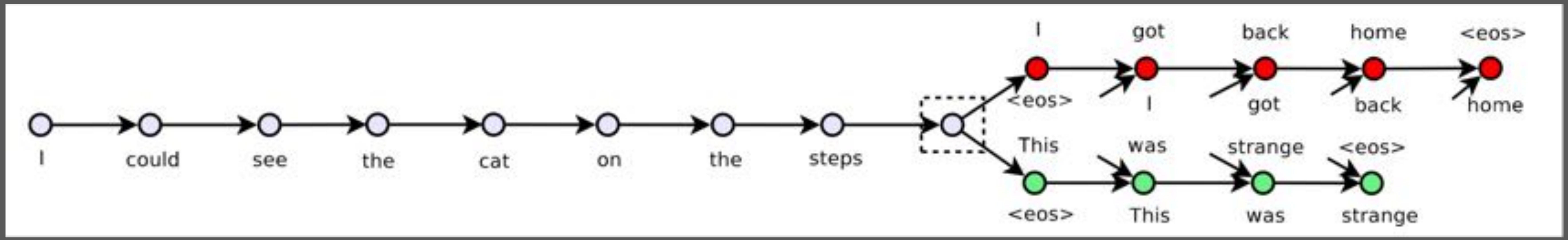
- Averaging word representations
- Learning with a generative objective
- Learning with a discriminative objective
- Supervised learning methods

Related Work

- Averaging word representations
  - Skip-gram & CBOW: Prediction-based models (Mikolov et al., NIPS2013)
  - GloVe: Count-based models (Pennington et al., EMNLP2014)
  - FastText: Skip-gram with character-level n-gram (Bojanowski et al., TACL2017)

Related Work

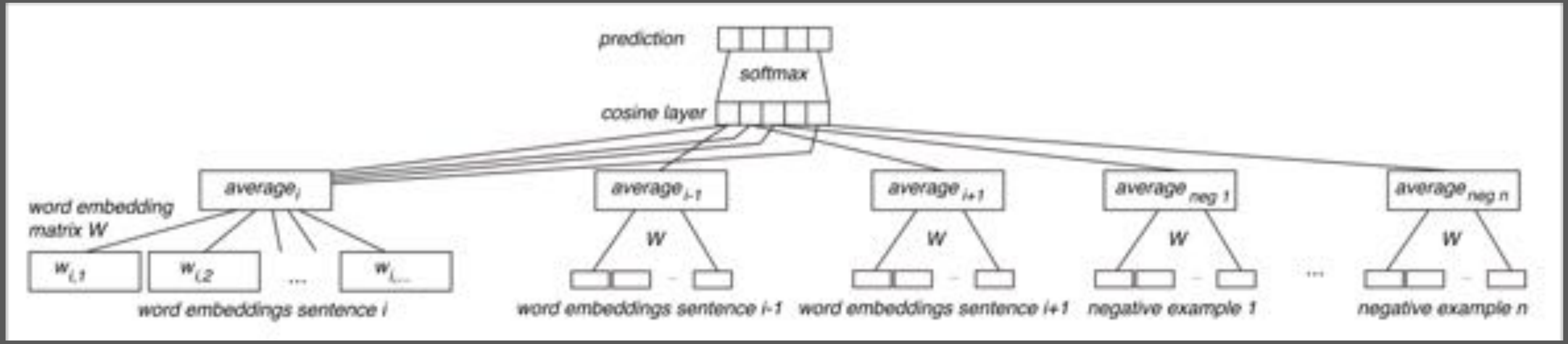
- Averaging word representations
- Learning with a generative objective
  - The encoder-decoder type of model
  - Skip-thoughts: predicting sentences in the context of the current one (Kiros et al., NIPS2015)
  - FastSent: (Hill et al., NAACL2016)

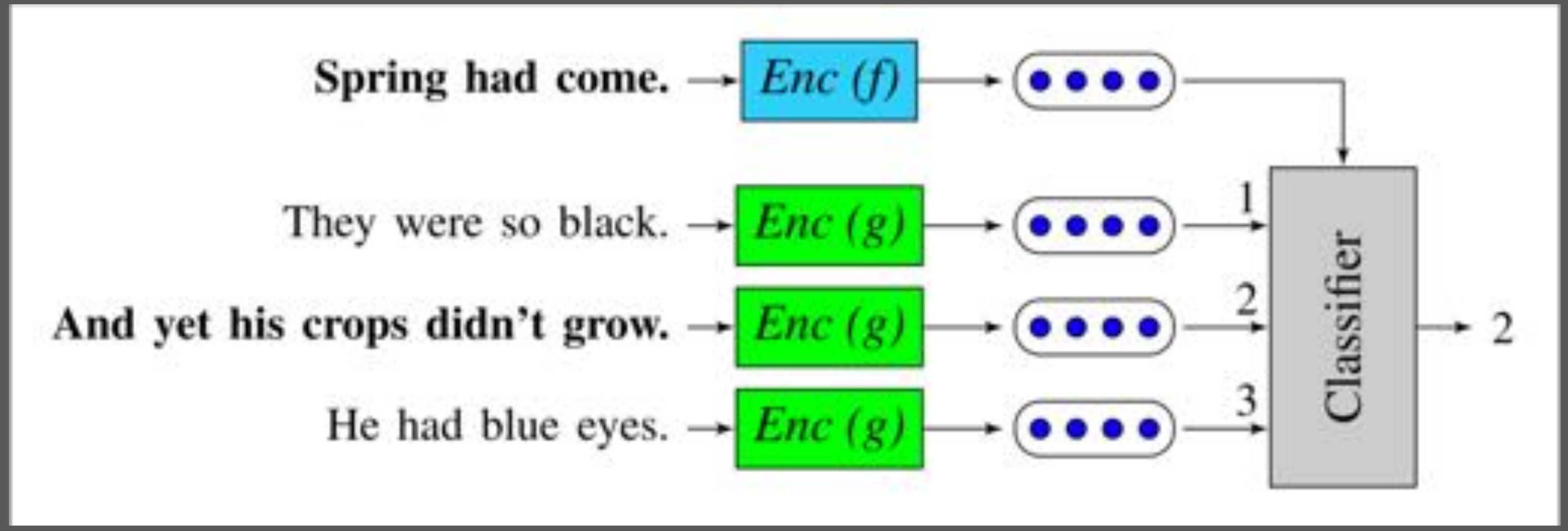


$$\mathbf{s_i} = \sum_{w \in S_i} u_w$$

Related Work

- Averaging word representations
- Learning with a generative objective
- Learning with a discriminative objective
  - Adjacent sentences should have more similar representations
  - Siamese CBOW: (Kenter et al., ACL2016)
  - Quick-thoughts: (Logeswaran & Lee, ICLR2018)





Related Work

- Averaging word representations
- Learning with a generative objective
- Learning with a discriminative objective
- Supervised learning methods (datasets)
  - Stanford Natural Language Inference (SNLI, Bowman et al., EMNLP2015)
  - Multi-genre Natural Language Inference (MultiNLI, Williams et al., 2017)
  - Machine Translation dataset
  - The Paraphrase Database (PPDB, Ganitkevitch et al., NAACL2013)

Related Work

- Averaging word representations
- Learning with a generative objective
- Learning with a discriminative objective
- Supervised learning methods (models)
  - InferSent (Conneau et al., EMNLP2017)
  - Context Vector (McCann et al., NIPS2017)
  - Paraphrastic Embeddings (Wieting & Gimpel, ACL2018)

Related Work

- Averaging word representations
- Learning with a generative objective
- Learning with a discriminative objective
- Supervised learning methods

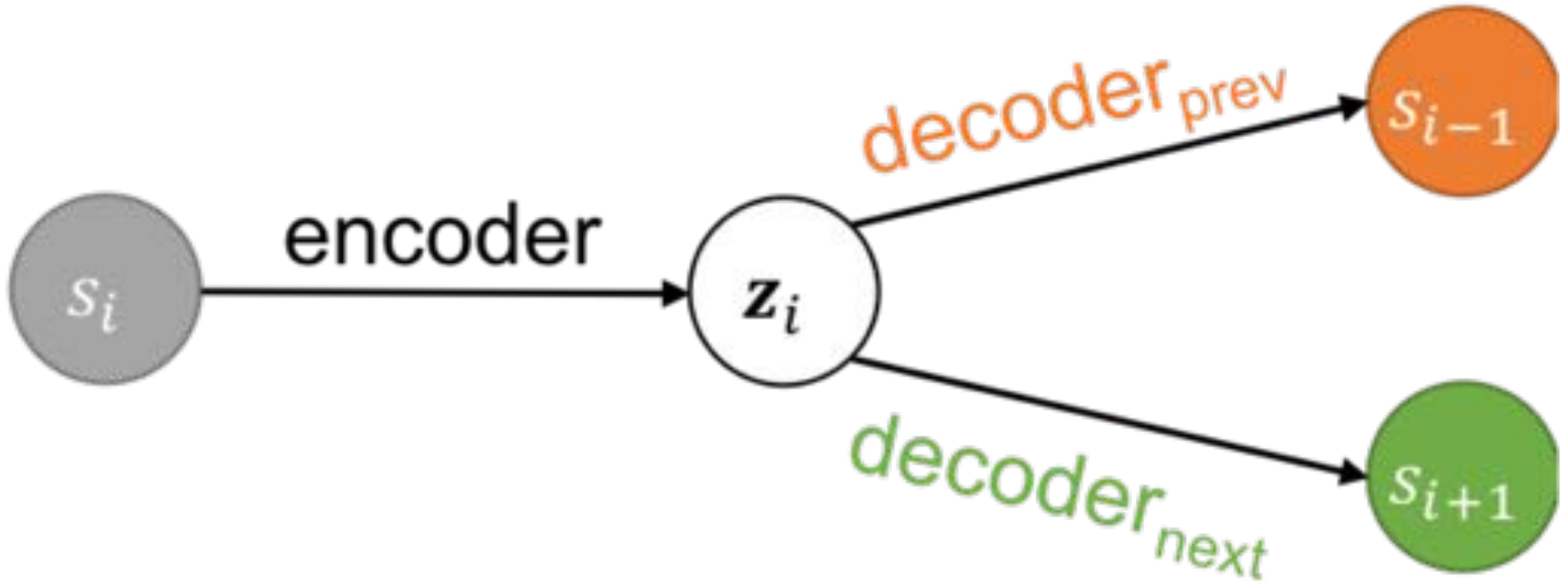
Outline

- Motivation
- Evaluation tasks
- Related work
- **Our previous and ongoing work**
- Future work

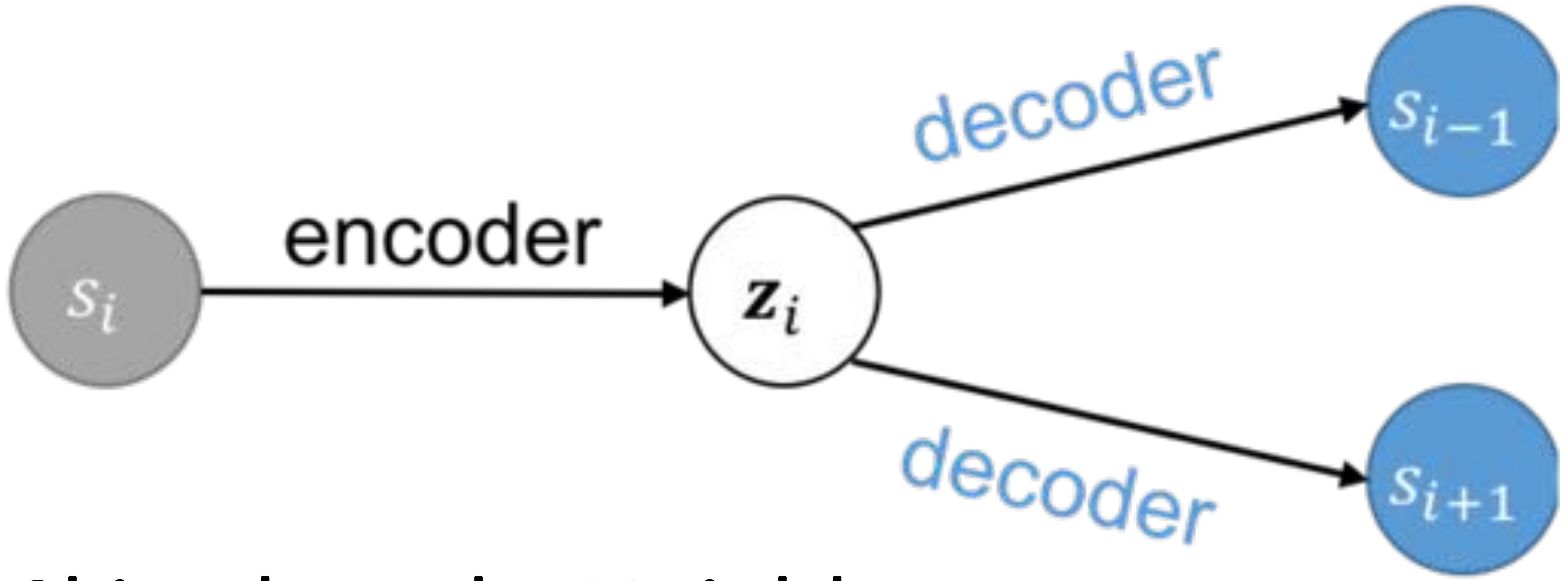
Our previous and ongoing work

- Part I: Skip-thought Neighbour Model
- Part II: Asymmetric RNN-CNN Model
- Part III: Multi-view Learning
- Part IV: Learning with Invertible Decoders

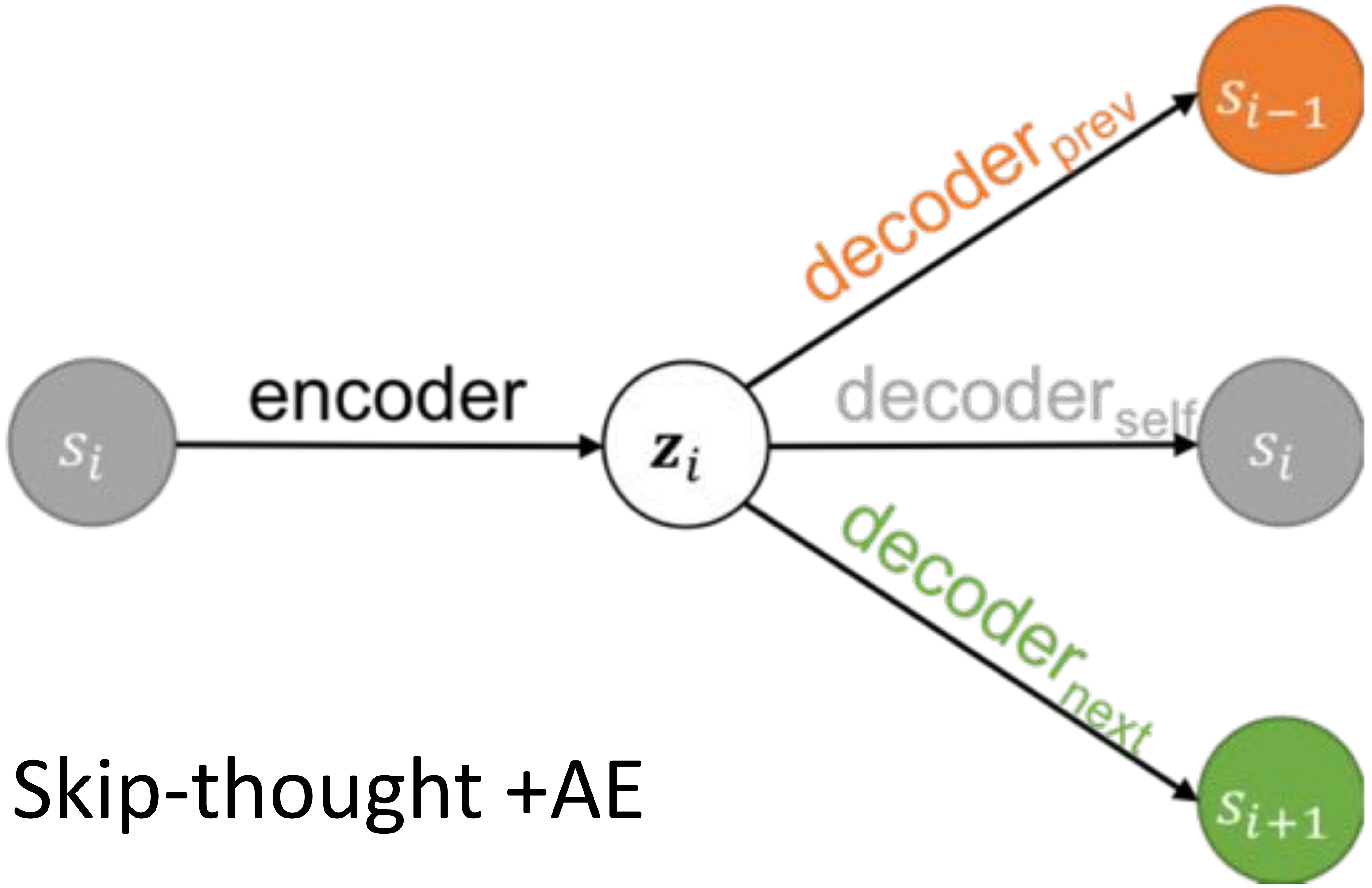
Part I: Skip-thought Neighbour Model



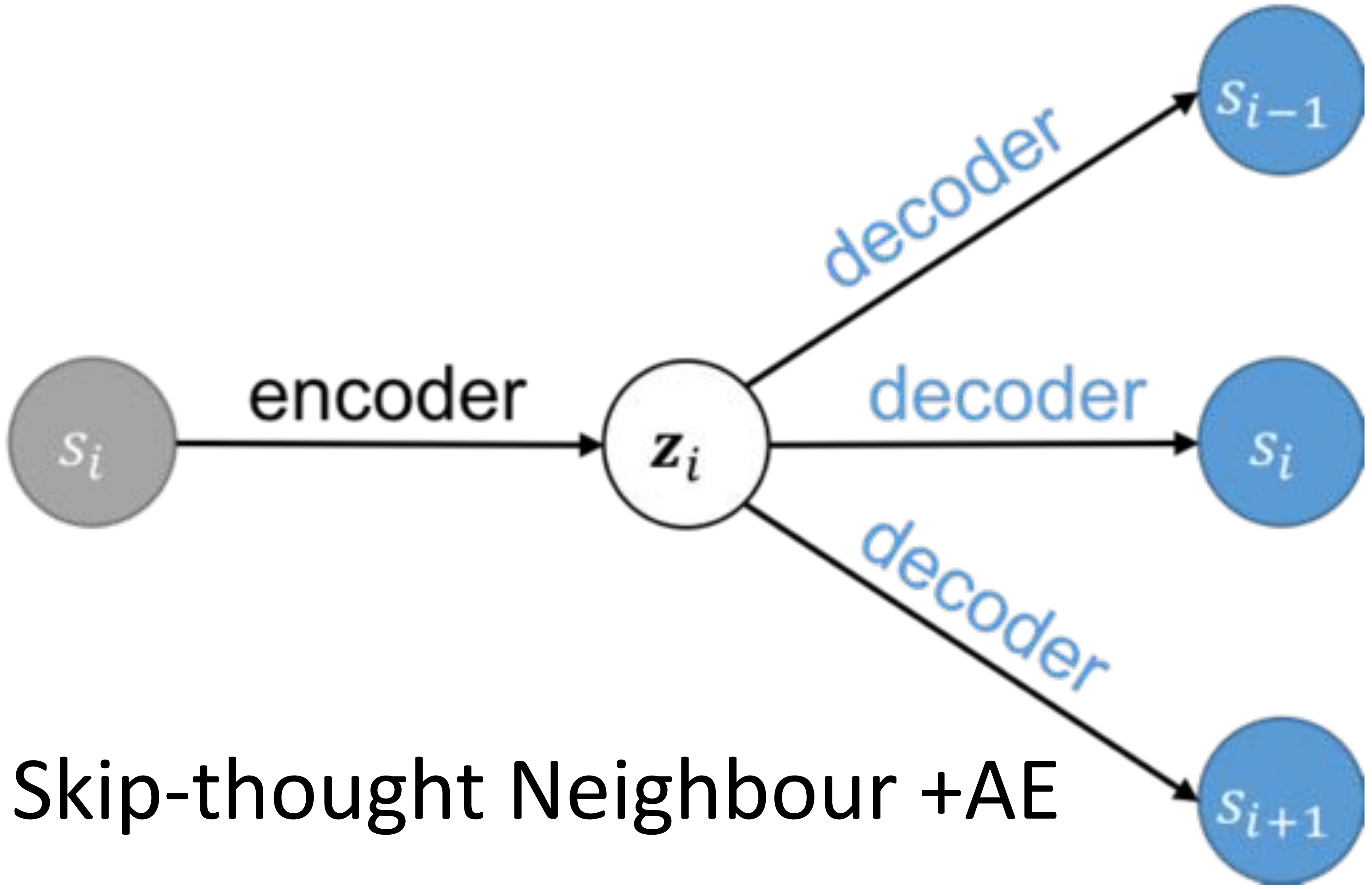
Skip-thought



Skip-thought Neighbour

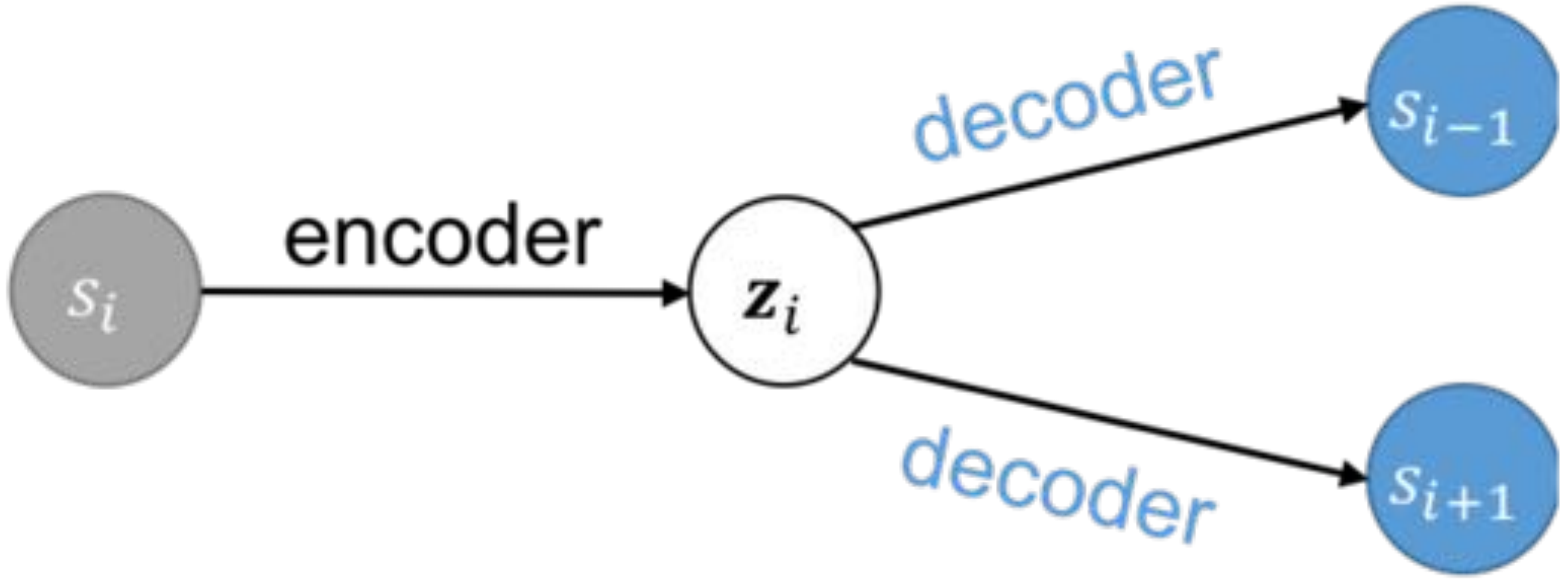


Skip-thought +AE

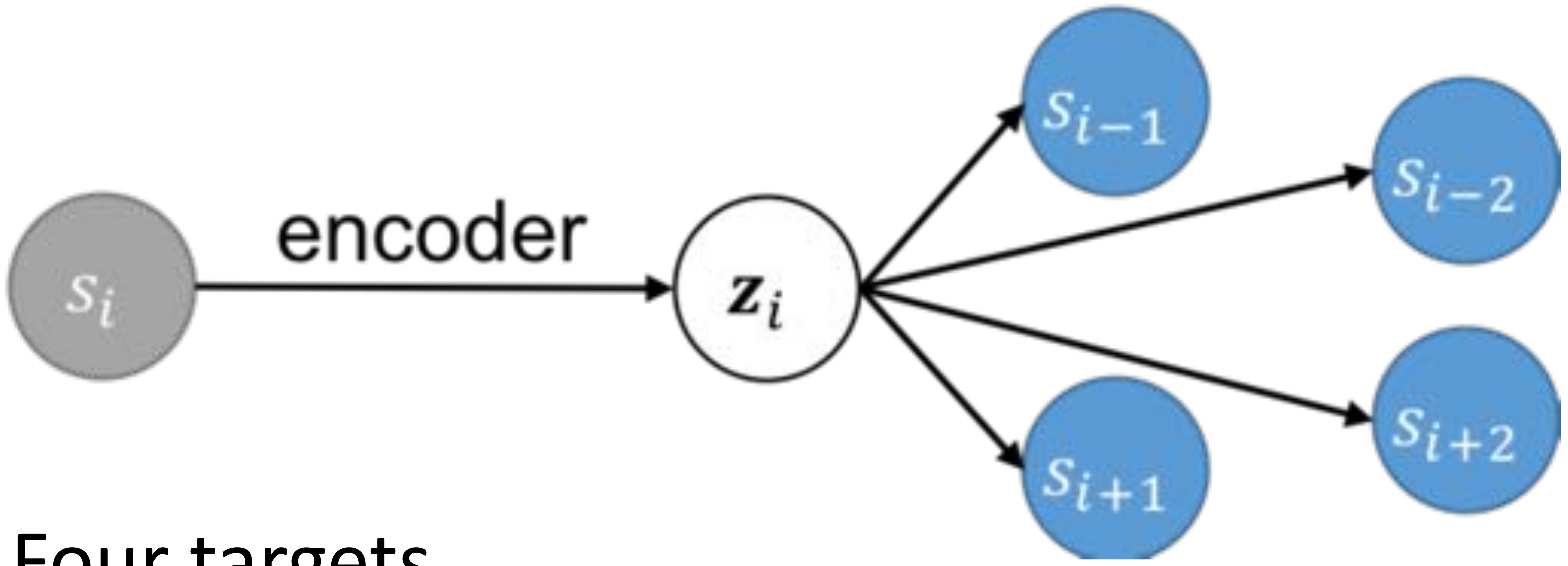


Skip-thought Neighbour +AE

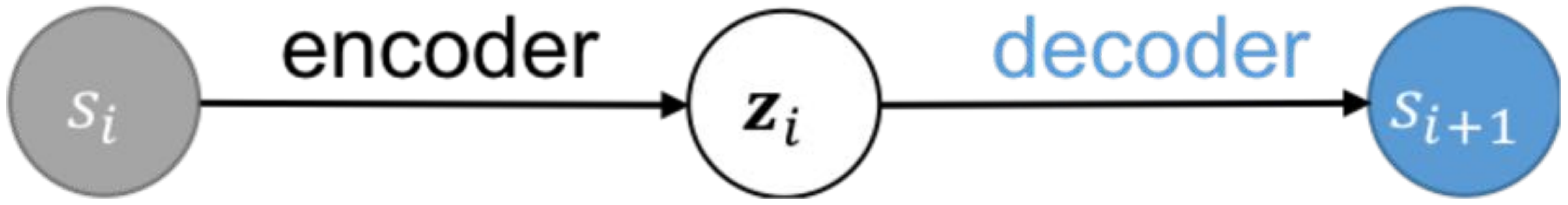
Part I: Skip-thought Neighbour Model



Two targets

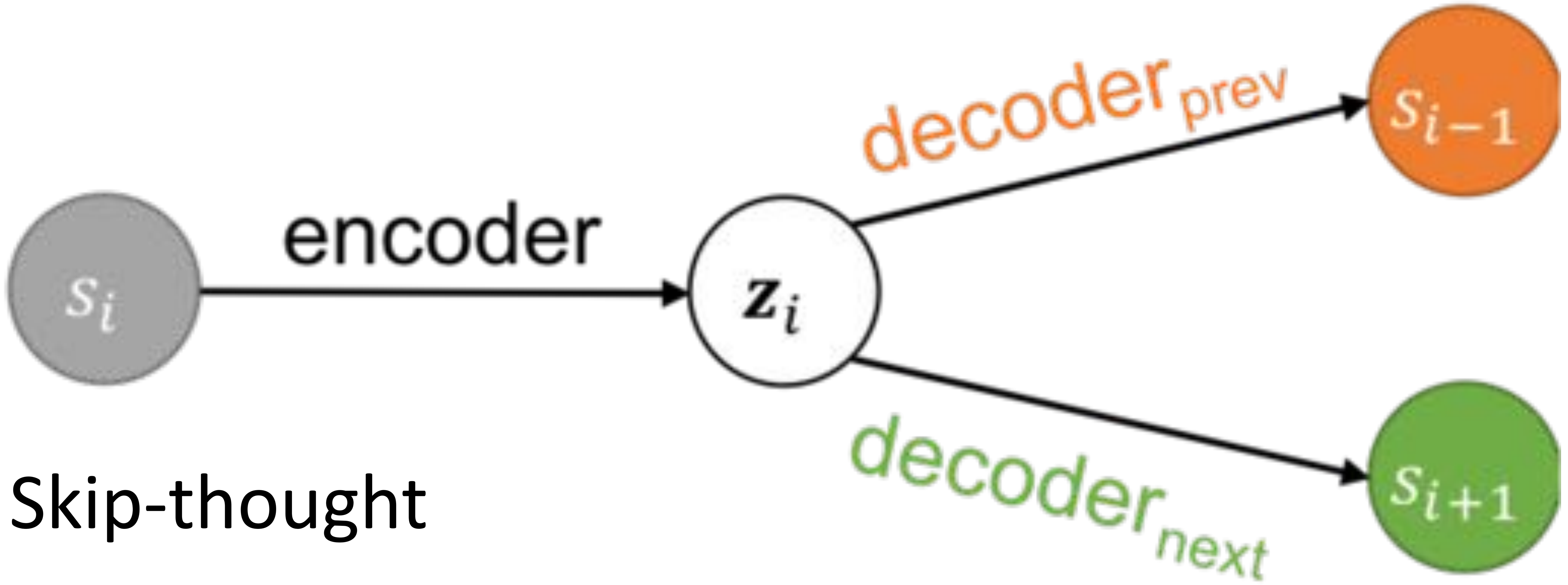


Four targets

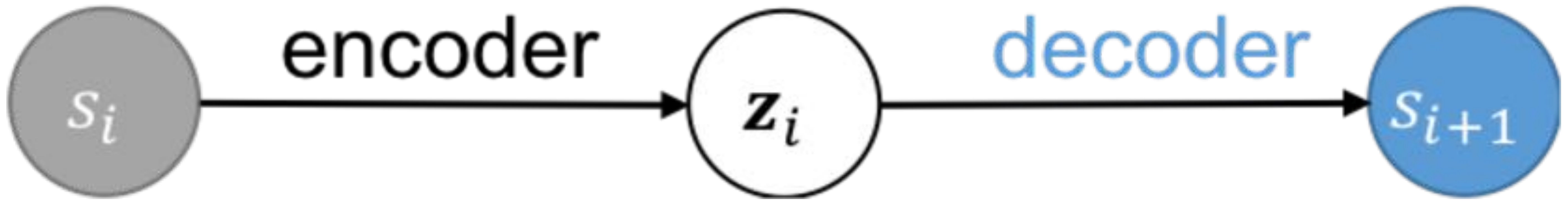


One target

Part I: Skip-thought Neighbour Model

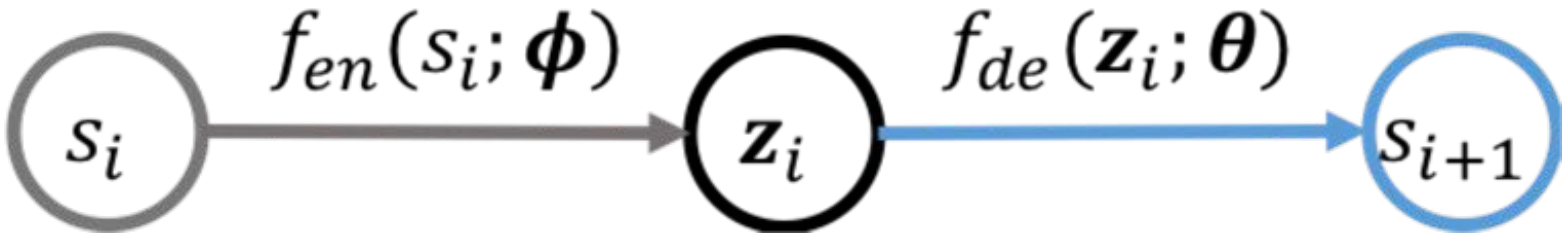


Skip-thought



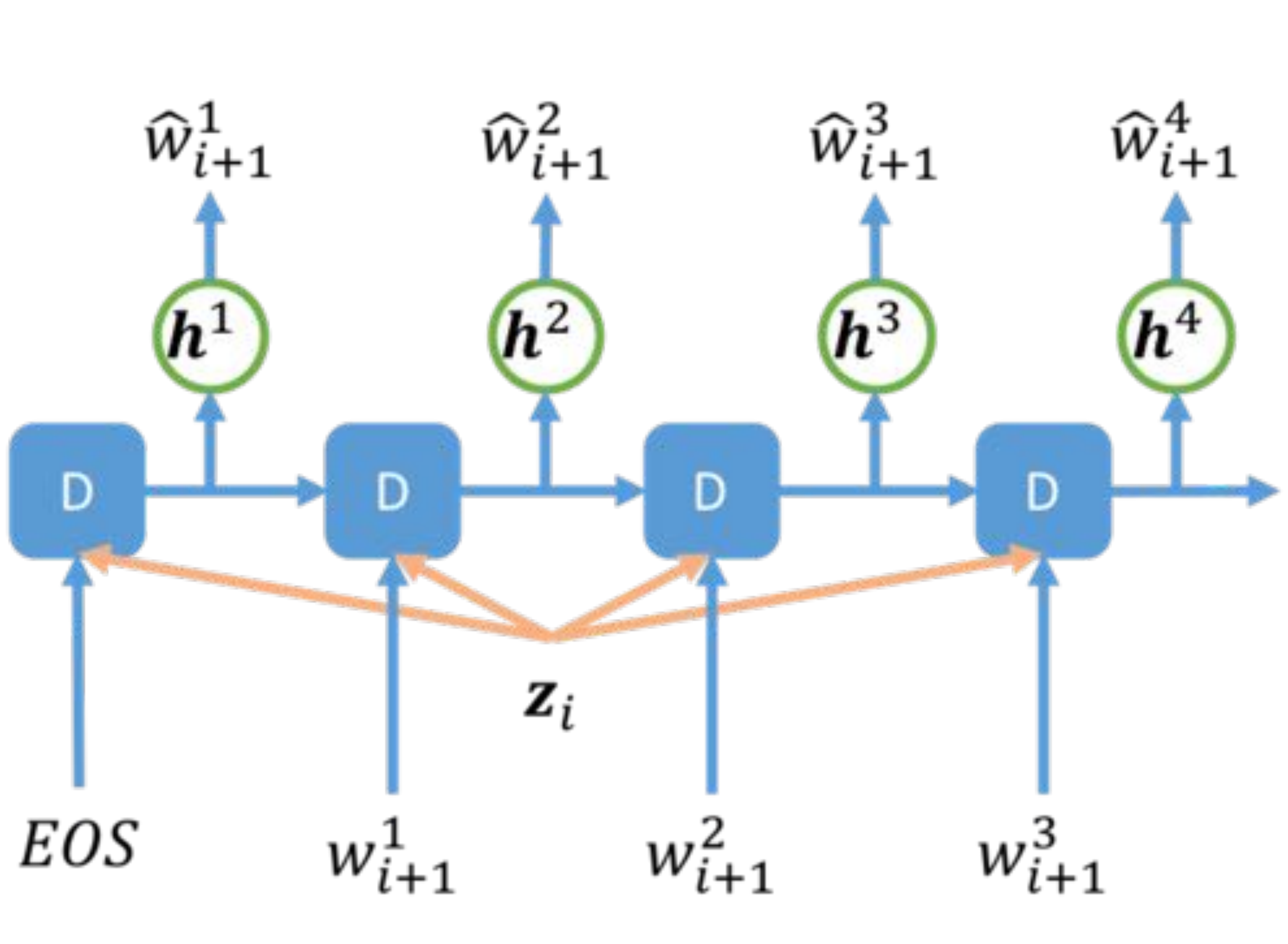
Skip-thought Neighbour with one target

Part II: Non-autoregressive CNN Decoding

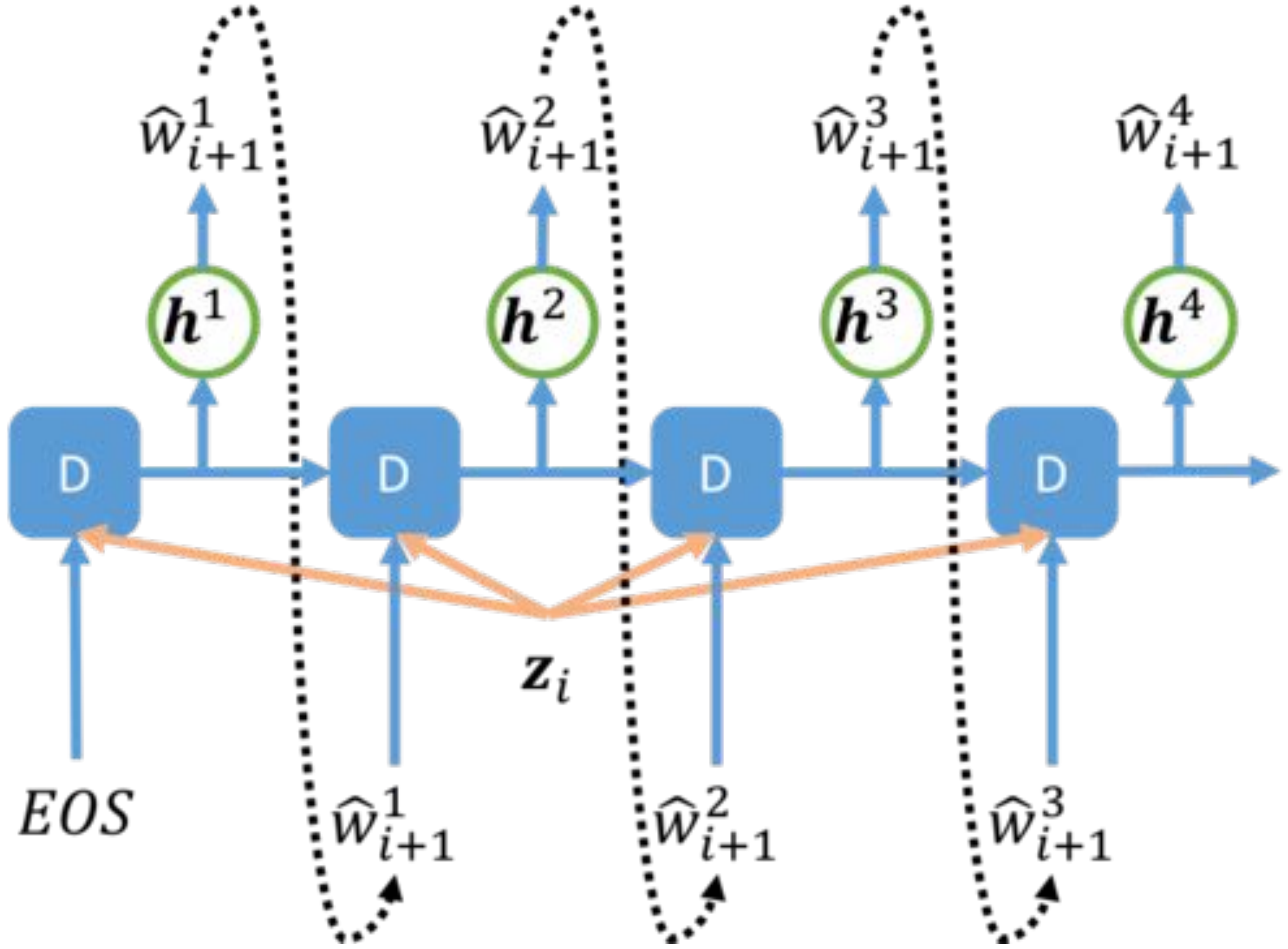


- Autoregressive Decoding?
- RNN Decoder?

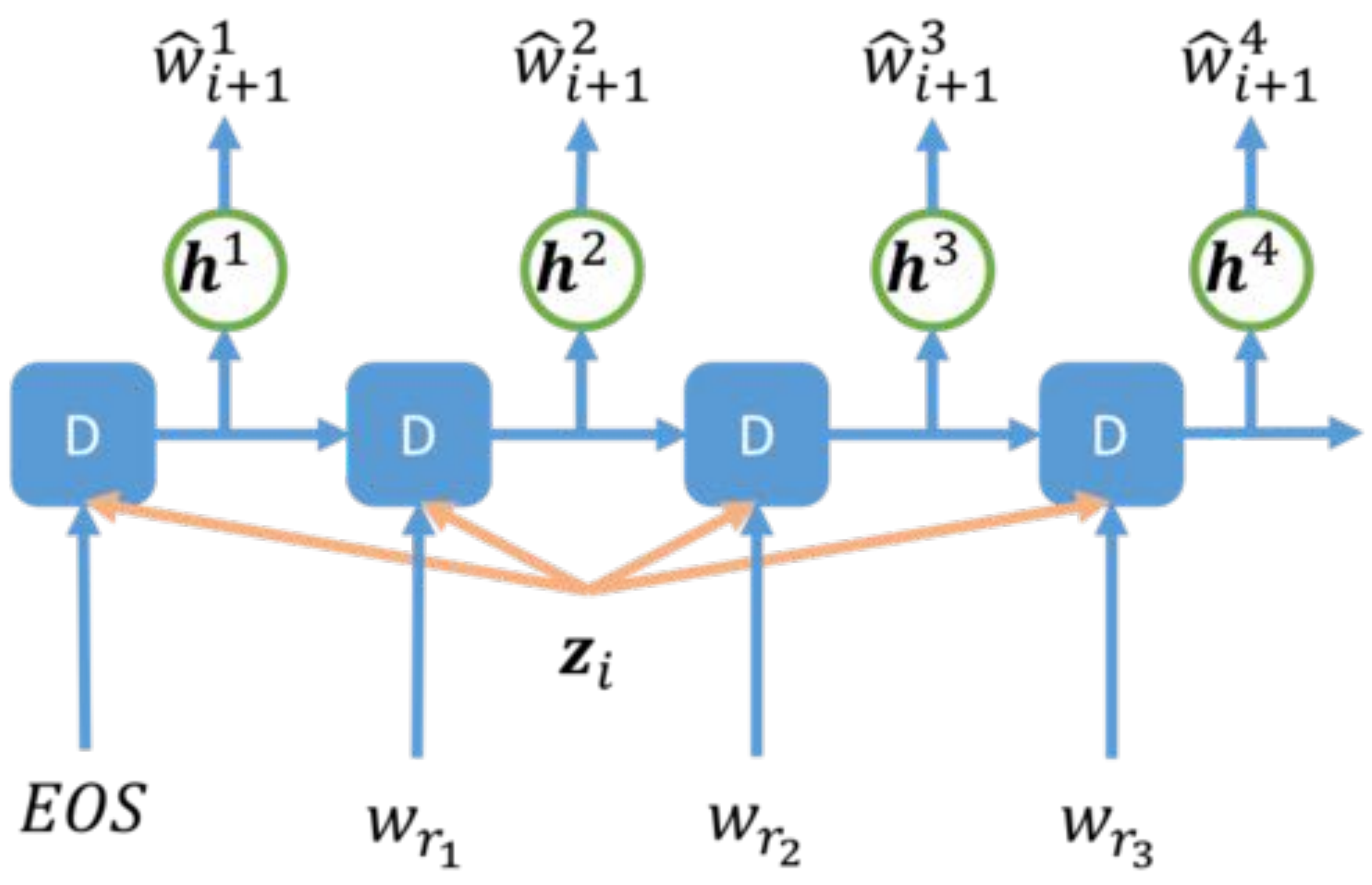
Part II: Non-autoregressive CNN Decoding



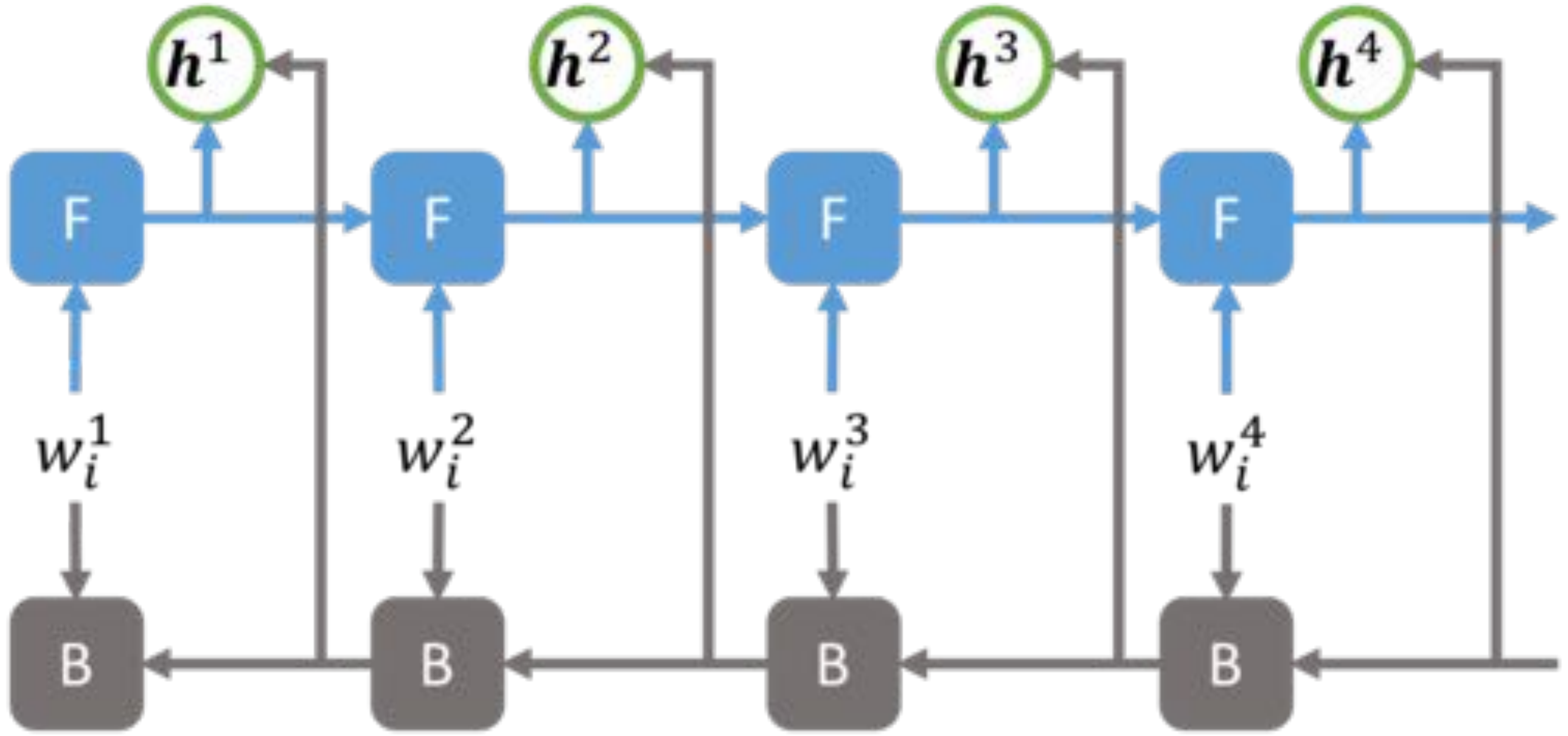
Teacher-forcing



Always Sampling

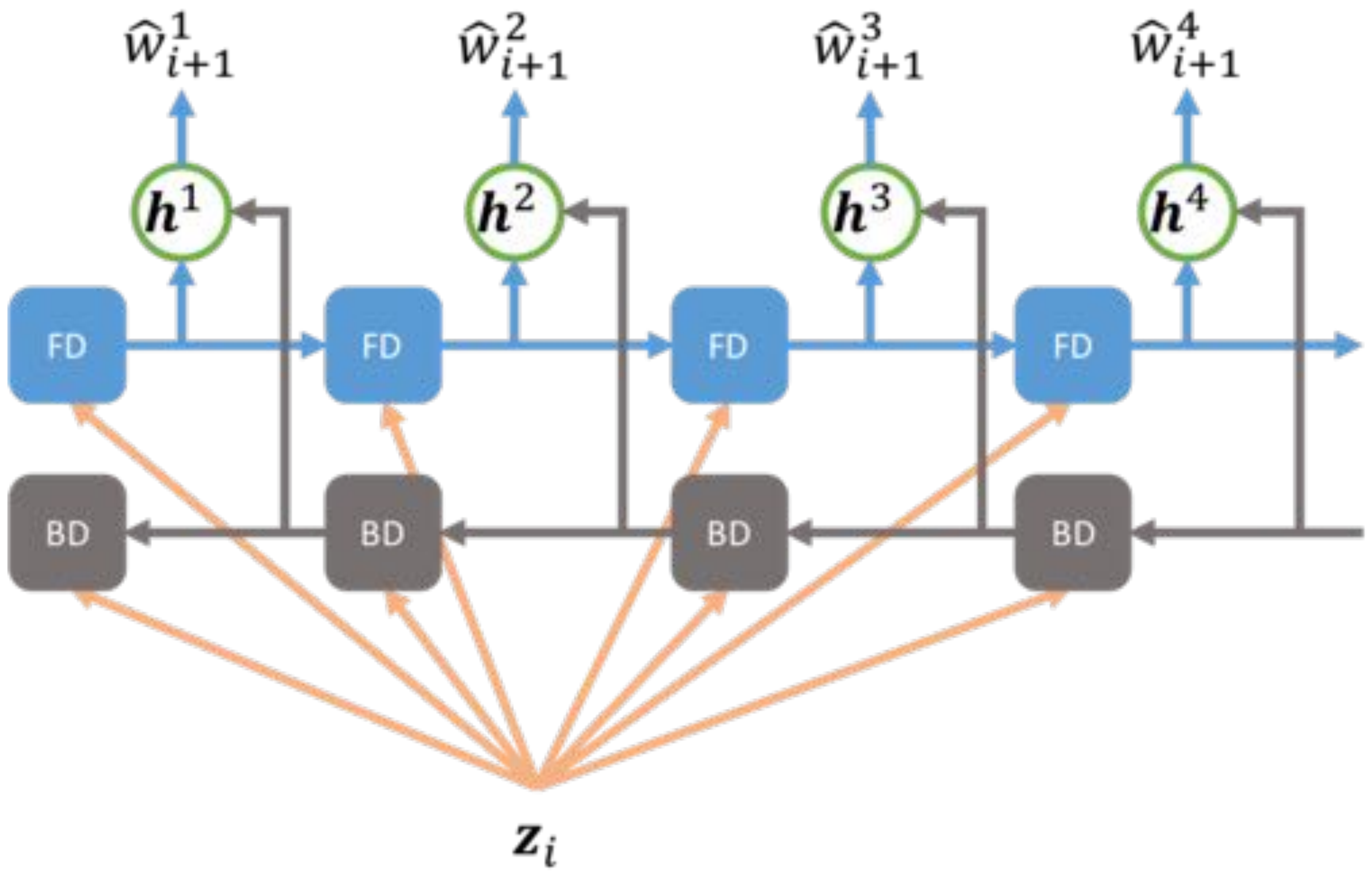


Uniform Sampling

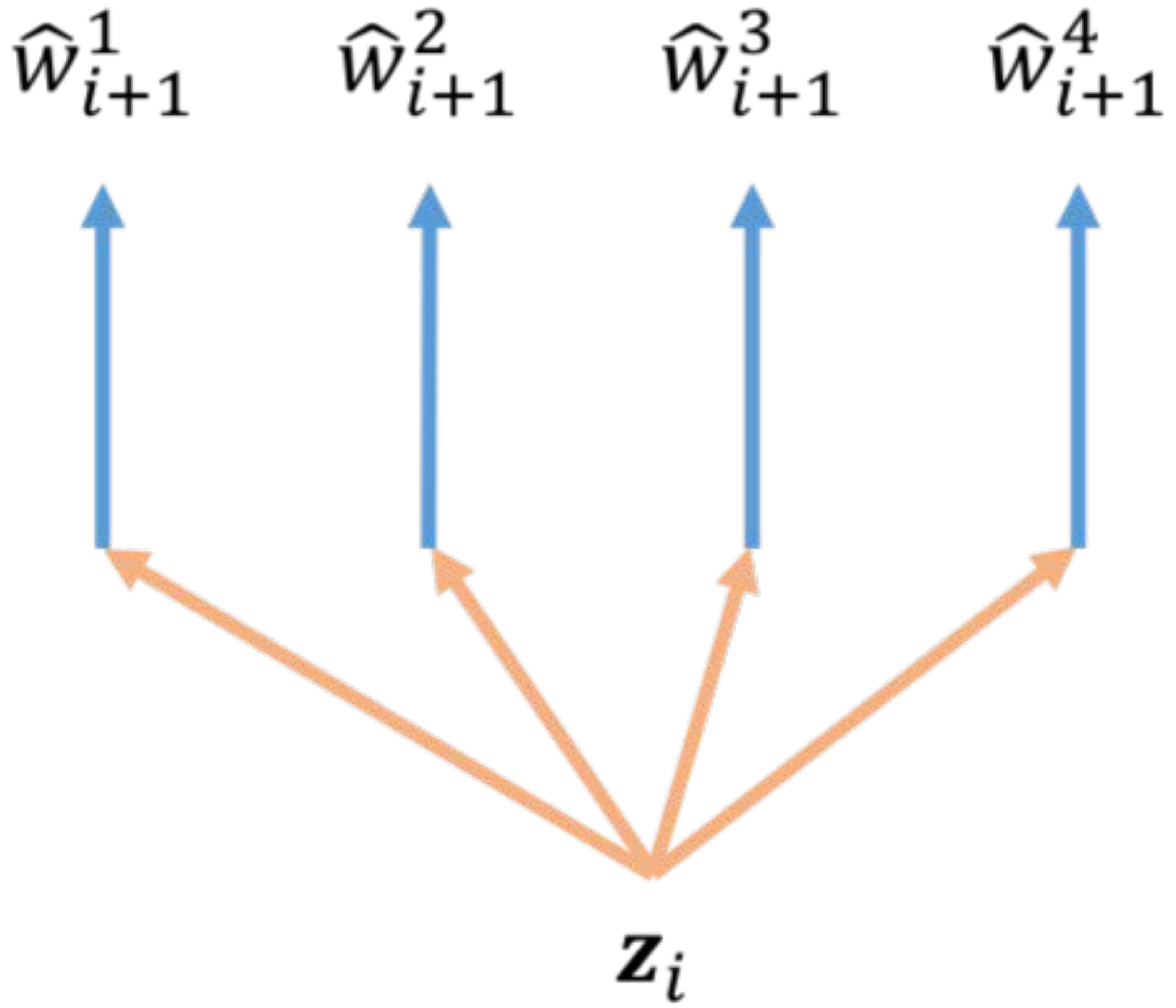


(Bengio et al., NIPS2015)

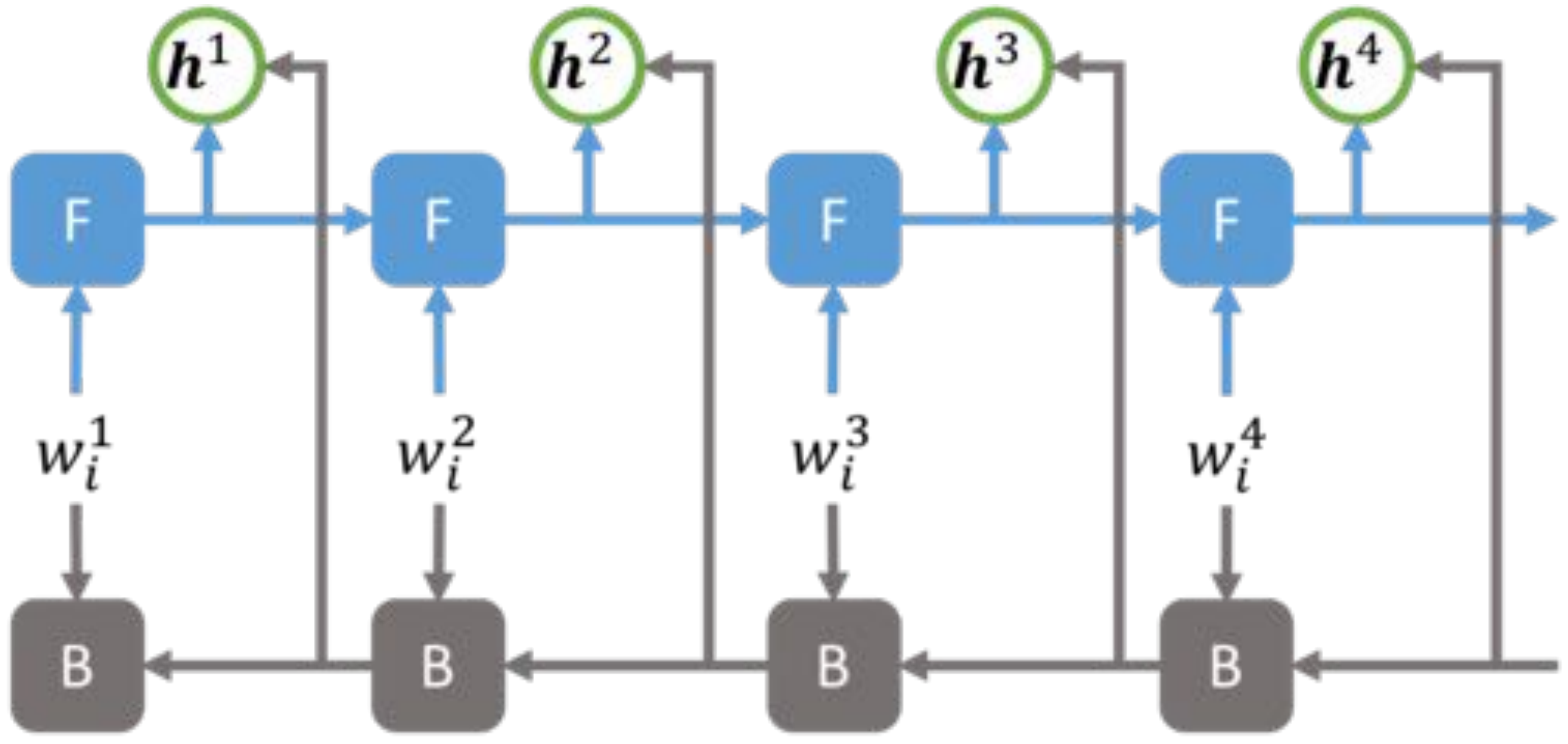
Part II: Non-autoregressive CNN Decoding



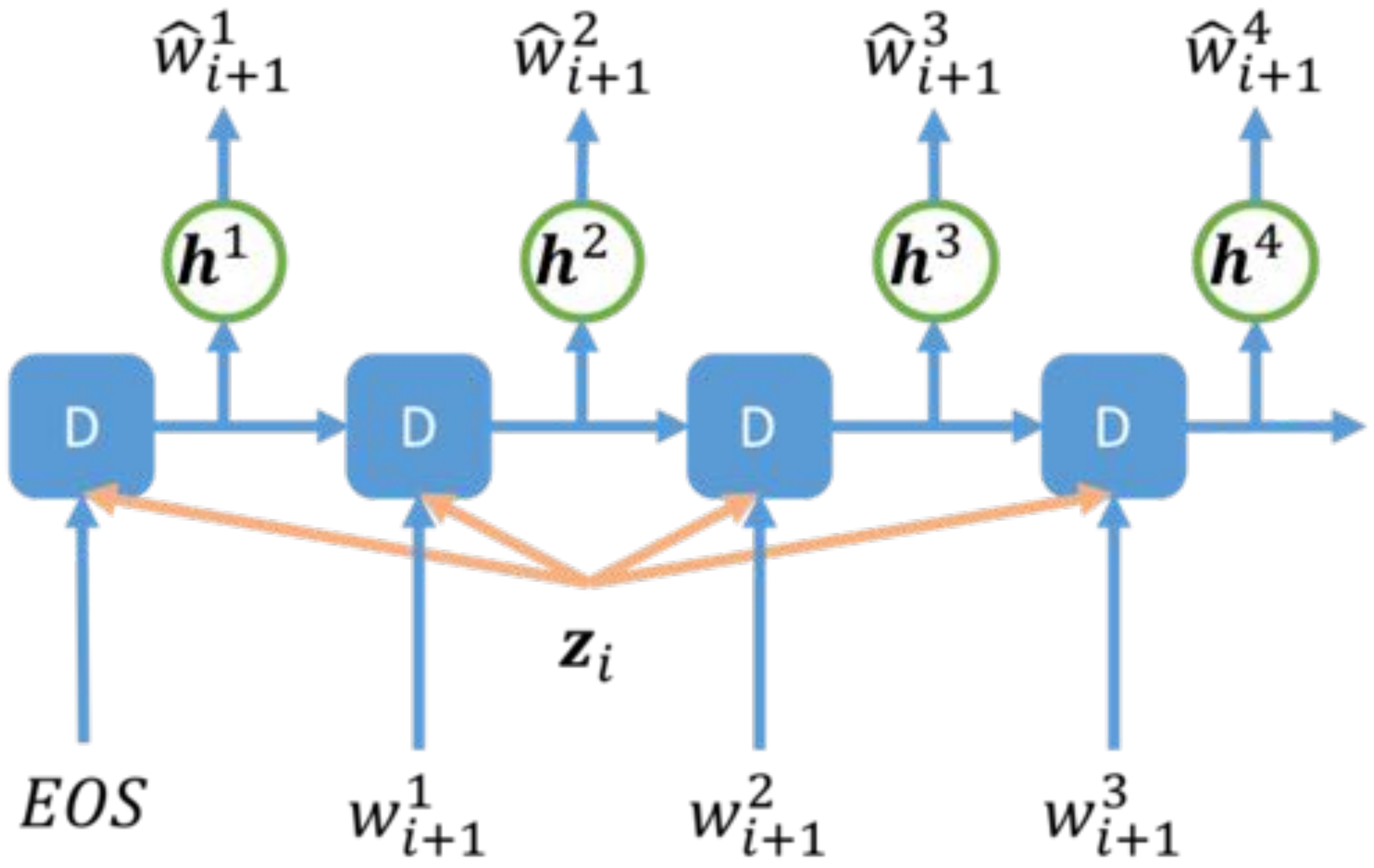
Non-autoregressive
RNN Decoding

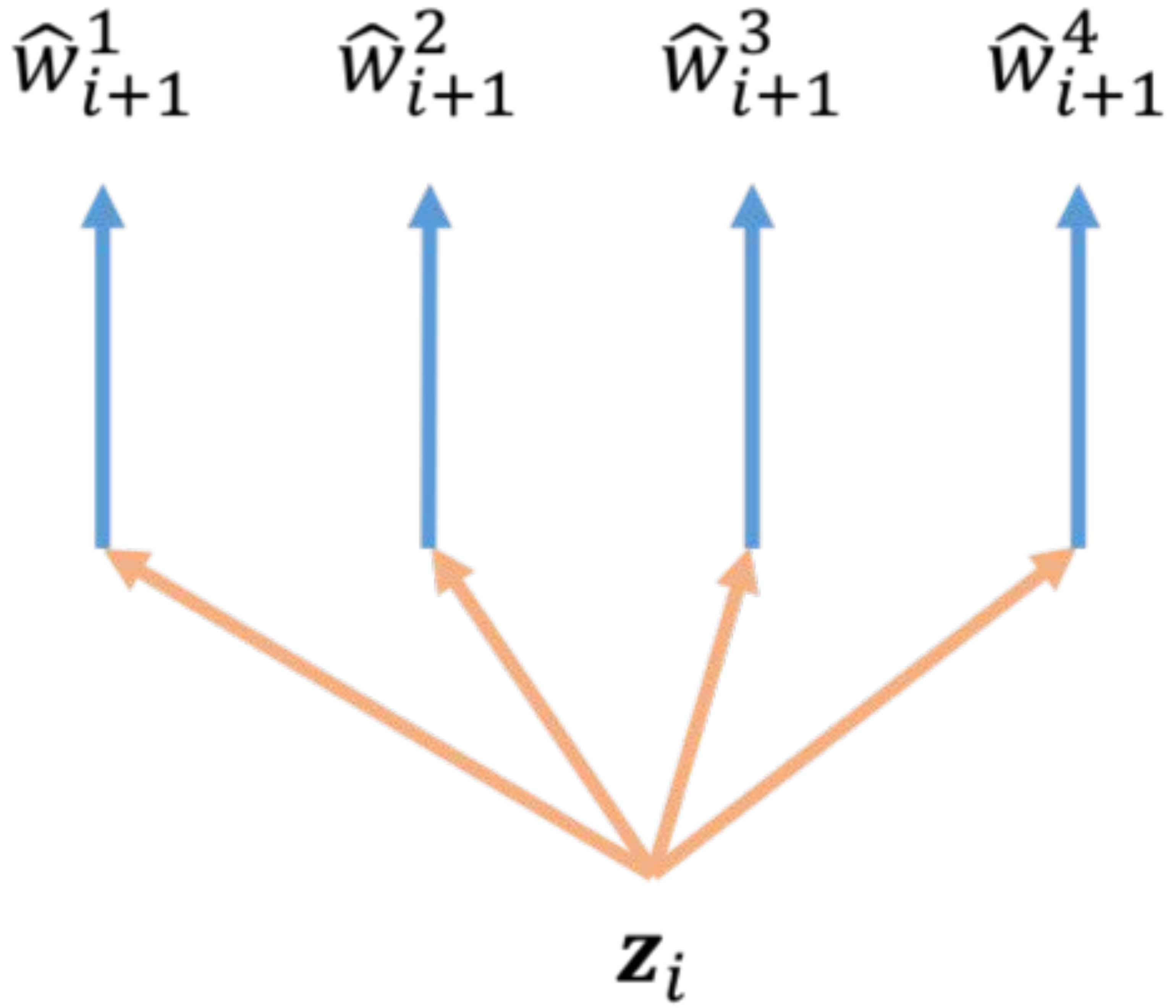


Non-autoregressive
CNN Decoding

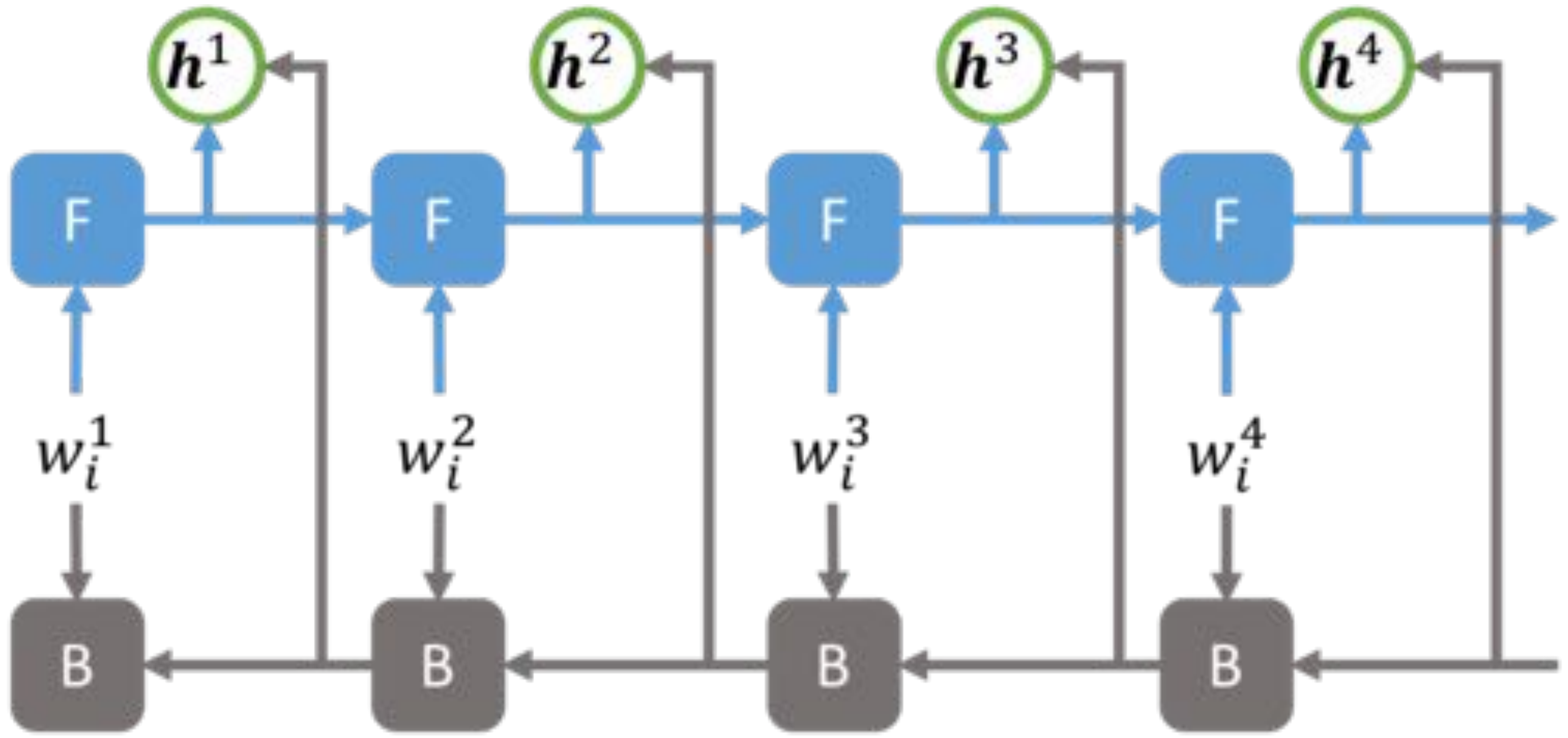


Part II: Non-autoregressive CNN Decoding





Skip-thought Neighbour



RNN-CNN Model

Part III: Multi-view Learning

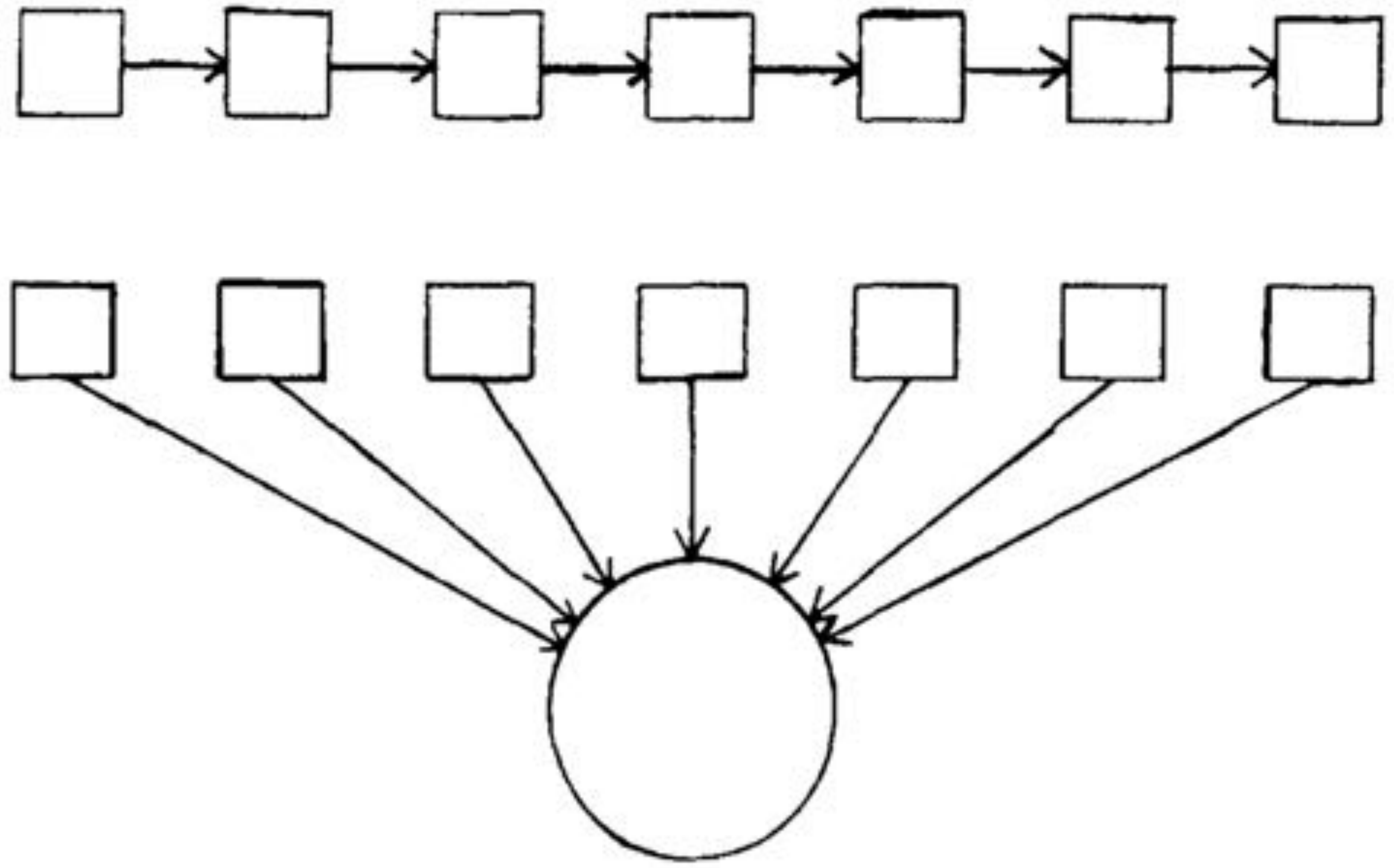
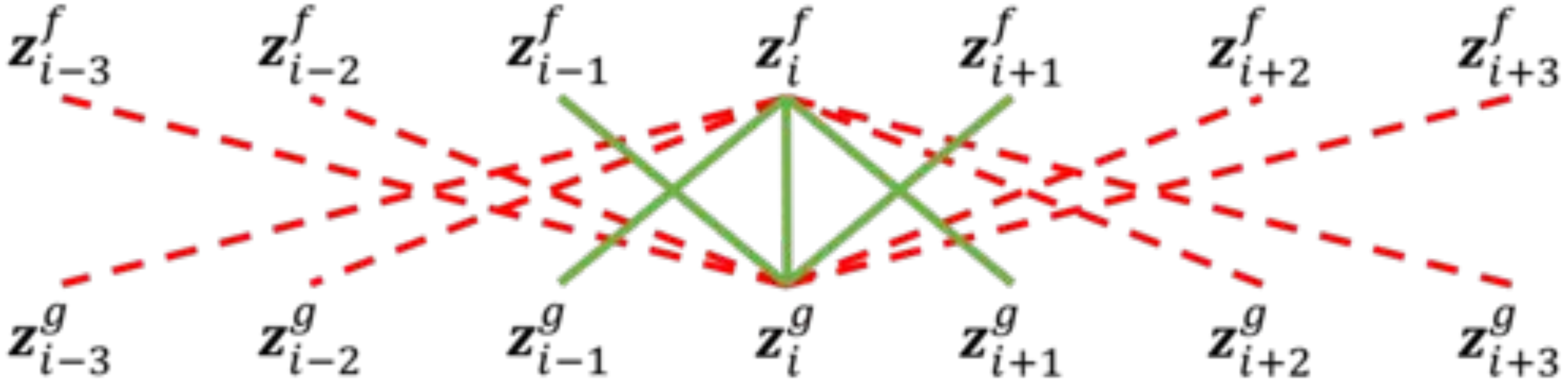


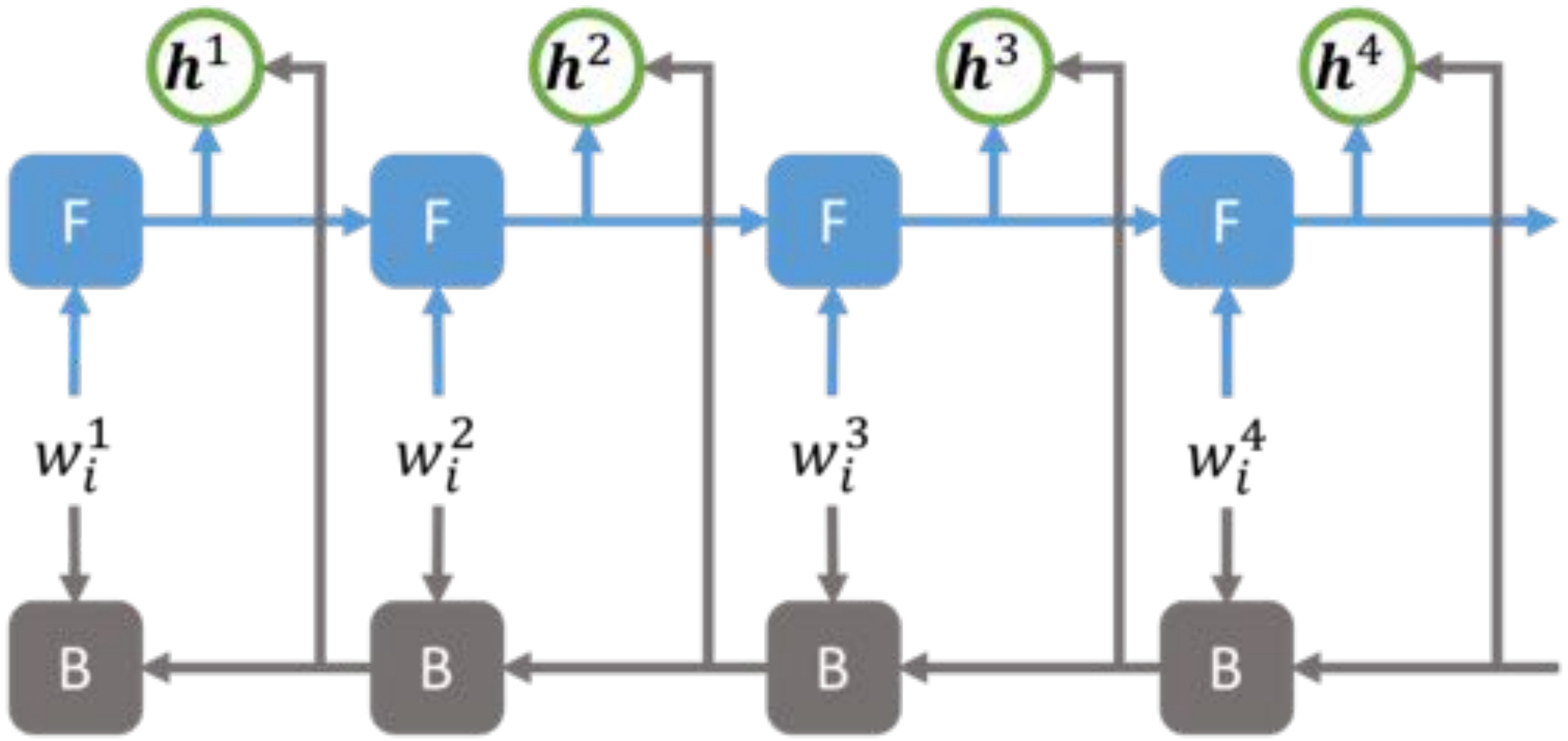
Figure 1. Linear processing (left hemisphere) and simultaneous processing (right hemisphere)

Tovey, Design Studies 1984

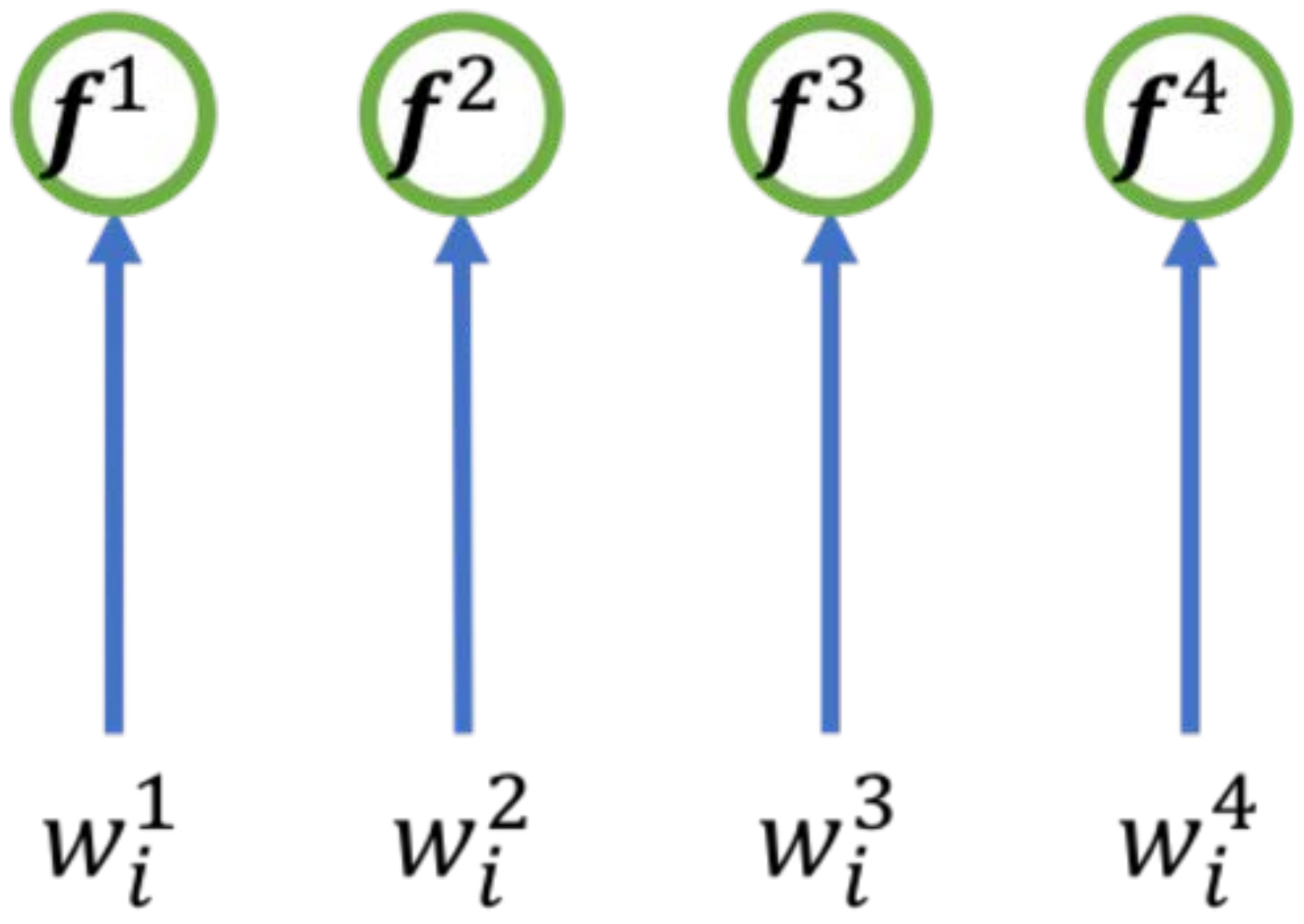
Lateralisation and asymmetry in information processing of the two hemispheres of the human brain. (Bryden, 2012)
For most adults, sequential processing dominates the left hemisphere, and the right hemisphere has a focus on parallel processing.

Part III: Multi-view Learning





f: RNN Encoder



g: Linear Encoder

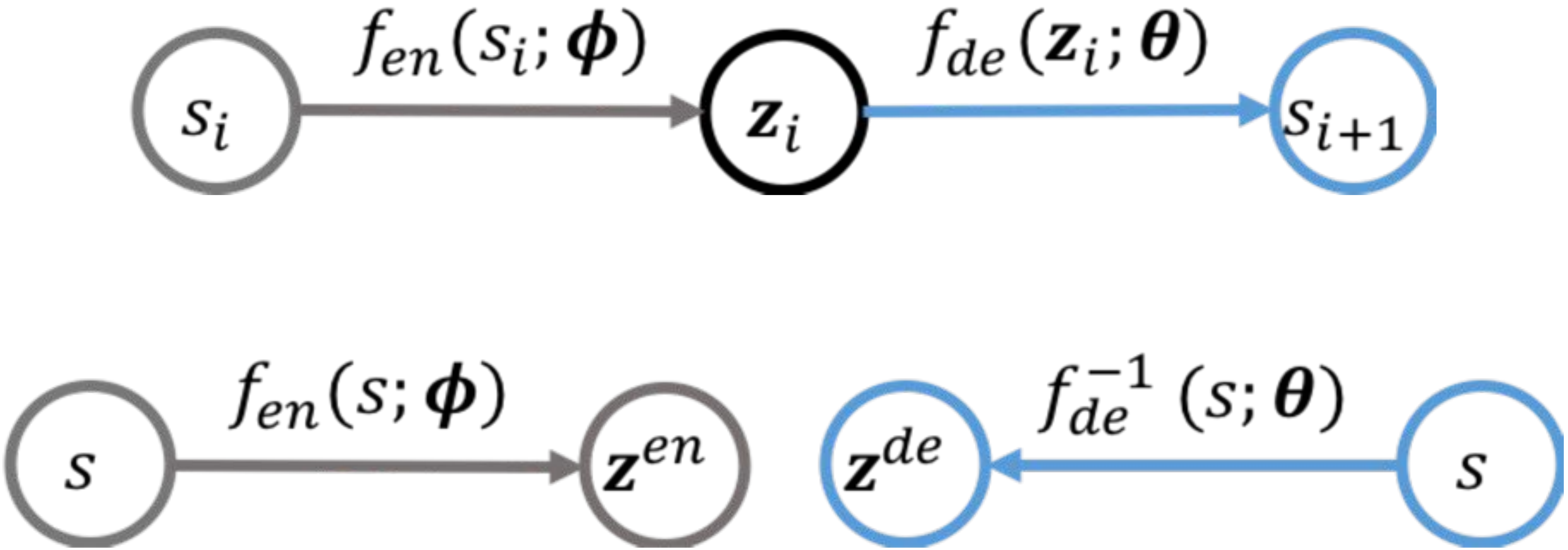
Part III: Multi-view Learning

$$f \longleftrightarrow g$$

$$f \qquad\qquad g$$

$$f_1 \longleftrightarrow f_2 \qquad\qquad g_1 \longleftrightarrow g_2$$

Part IV: Learning with Invertible Decoders



Part IV: Learning with Invertible Decoders

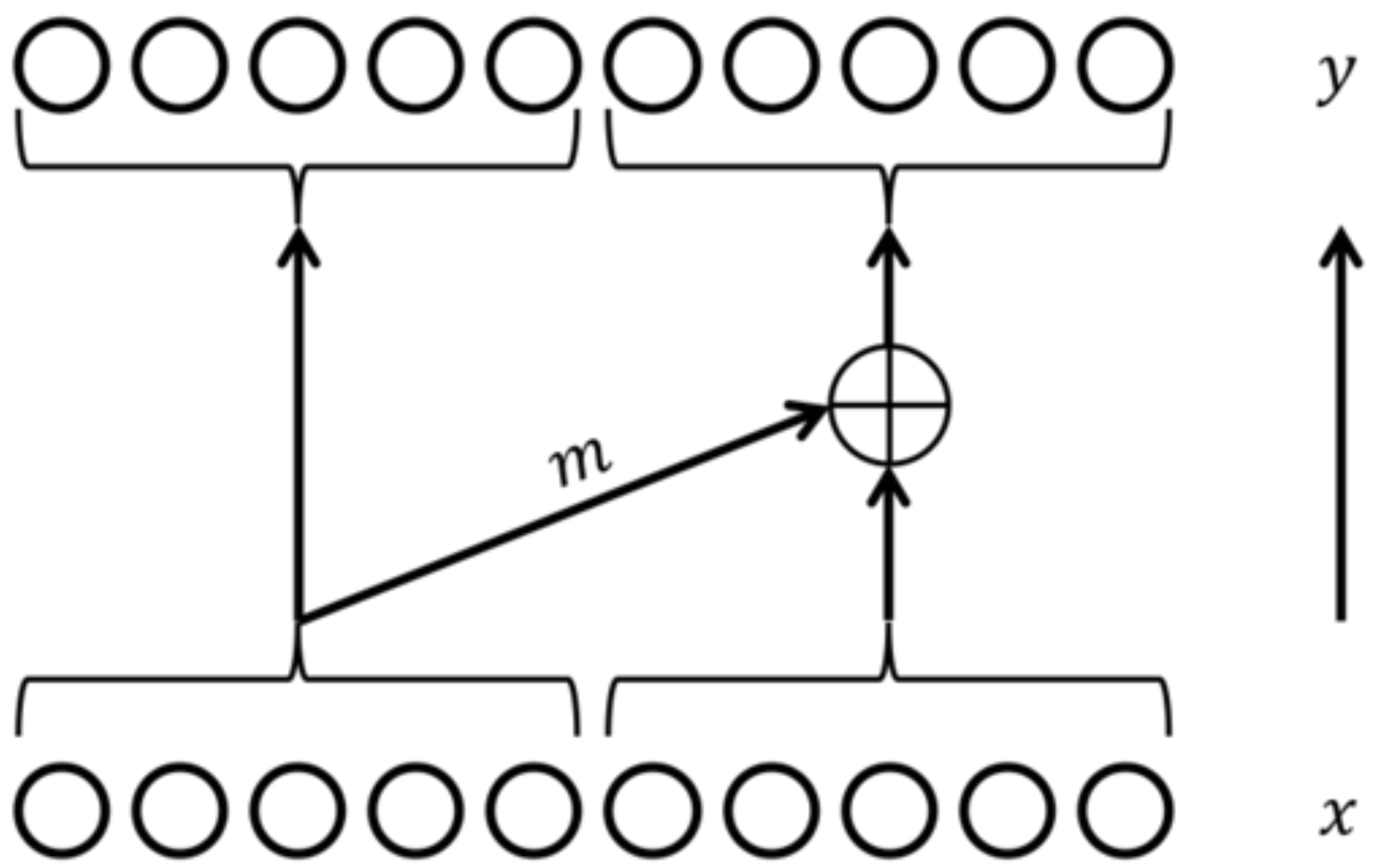
- Linear Projection
  - (Cissé et al., ICML2017)

$$f_{\text{de}}(\mathbf{z}) = \mathbf{W}\mathbf{z}$$

$$f_{\text{de}}^{-1}(\mathbf{x}) = \mathbf{W}^\top(\mathbf{W}\mathbf{W}^\top)^{-1}\mathbf{x} \qquad \mathbf{W}\mathbf{W}^\top = \mathbf{I}$$

$$f_{\text{de}}^{-1}(\mathbf{x}) = \mathbf{W}^\top\mathbf{x}$$

- Bijective Transformations
  - (Dinh et al., ICLR Workshop 2014)



Our previous and ongoing work

- Part I: Skip-thought Neighbour Model
  - Tang et al., RepL4NLP@ACL2017
- Part II: Non-autoregressive CNN Decoding
  - Tang et al., RepL4NLP@ACL2018
- Part III: Multi-view Learning
  - Tang & de Sa, submitted to NIPS2018
- Part IV: Learning with Invertible Decoders
  - Tang & de Sa, submitted to EMNLP2018

Outline

- Motivation
- Evaluation tasks
- Related work
- Our previous and ongoing work
- **Future work**

Future Work

- On unifying the generative objective and discriminative objective
- Curse and Blessings of the Dimensionality
- Representation Space

Generative & Discriminative Objective

Generative Objective
Encoder-decoder

Multi-view Learning with a Discriminative Objective

Inverse of the decoder

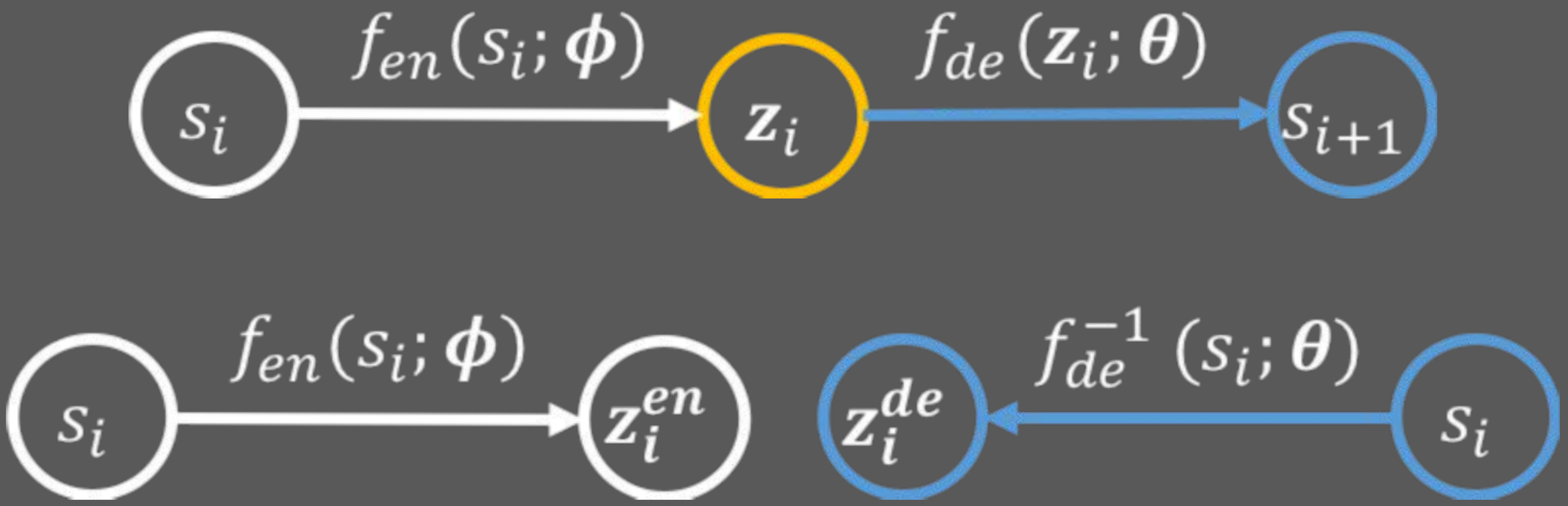
Encoder f

Encoder

Encoder g

Generative & Discriminative Objective

Multi-task Learning



Curse and Blessings of the Dimensionality

- Curse of the dimensionality
  - Unsupervised evaluation tasks

- Blessings of the dimensionality
  - Supervised evaluation tasks

- Leverage both principles into a unified hierarchical model

(Donoho, 2000)

Representation Space

- Euclidean Space (Osgood et al., 1957)
  - Frequently appeared words have representations with longer lengths
- Unit Sphere (cosine similarity)
  - Curse of the dimensionality
- Hyberbolic Geometry
  - n-dimensional Poincaré ball

Outline

- Motivation
- Evaluation tasks
- Related work
- Our previous and ongoing work
- Future work

Acknowledgements

- All the committee members

Adobe
RESEARCH

- Sam Bowman at NYU
- All my friends

Thank you!