Learning Image Embeddings using CNNs for Improved Multi-Modal Semantics

Douwe Kiela, Léon Bottou, 2014

Daub Haubenreisser

2.7.19

Autoren

Douwe Kiela

Léon Bottou

- 2014 war Kiela als Gastwissenschaftler bei Microsoft Research, wo Bottou angestellt war
- mittlerweile arbeiten beide bei Facebook AI unter Yann LeCun

Inhaltsverzeichnis

1. Frage

- Was ist die Hauptidee hinter dem multimodalen Ansatz und wie wird dieser zu menschlicher Kognition in Beziehung gesetzt?

[1] Douwe Kiela: TEDx talk

1. Frage

- Was ist die Hauptidee hinter dem multimodalen Ansatz und wie wird dieser zu menschlicher Kognition in Beziehung gesetzt?
- Man möchte die Bedeutung eines Wortes erweitern um zusätzliche Informationen **aus anderen Quellen (Modi)**
- ähnlich zum menschlichen Lernprozess: wahrnehmungsbasiert
- Ziel: Performanz der Wortrepräsentation steigern
- Allgemein: keine echte AI ohne Lösung des Symbol Grounding Problems (wie erhalten Zeichen ihre Bedeutung?) [1]

[1] Douwe Kiela: TEDx talk

Inhaltsverzeichnis

Modell

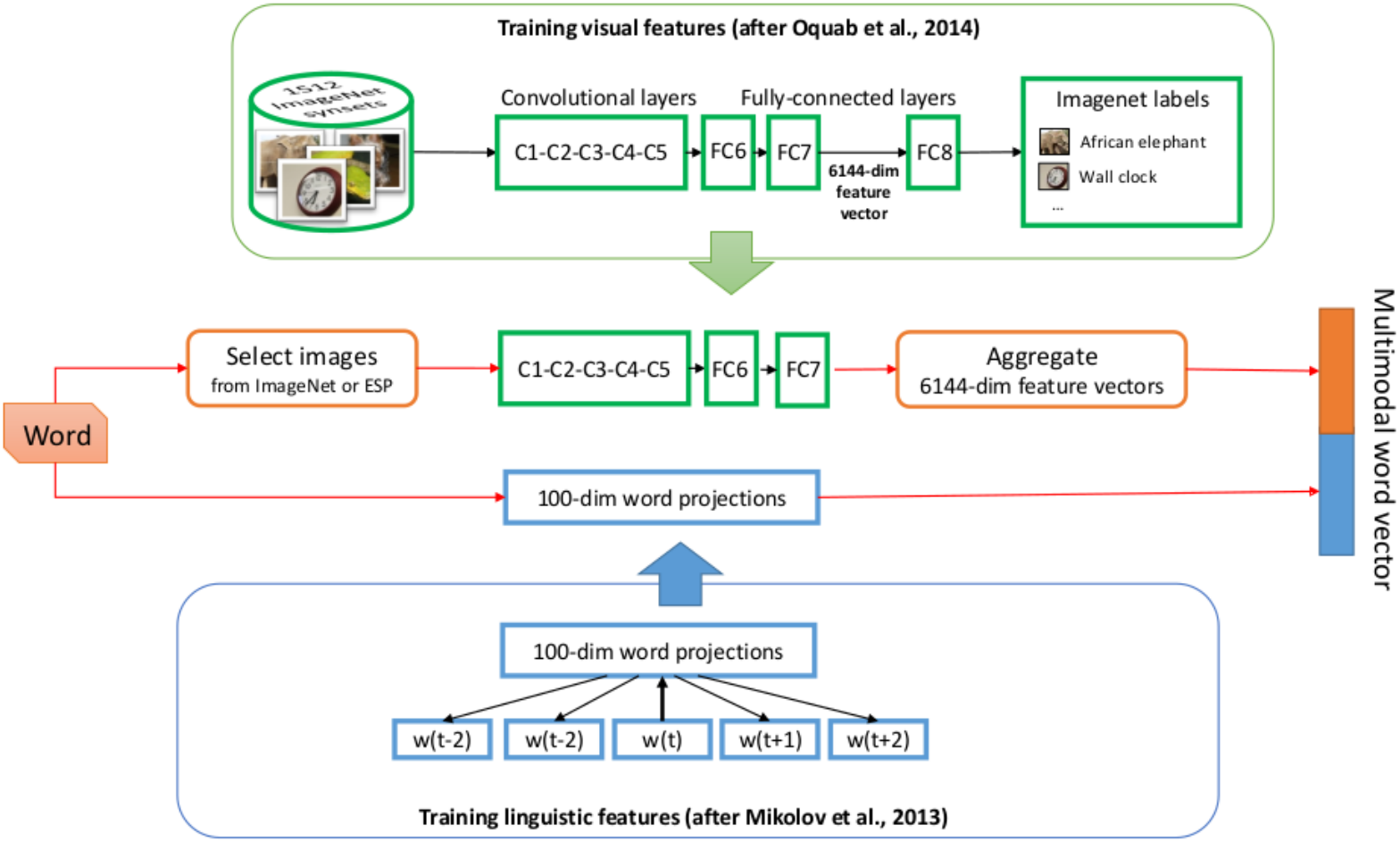


Figure 1: Computing word feature vectors.

Modell

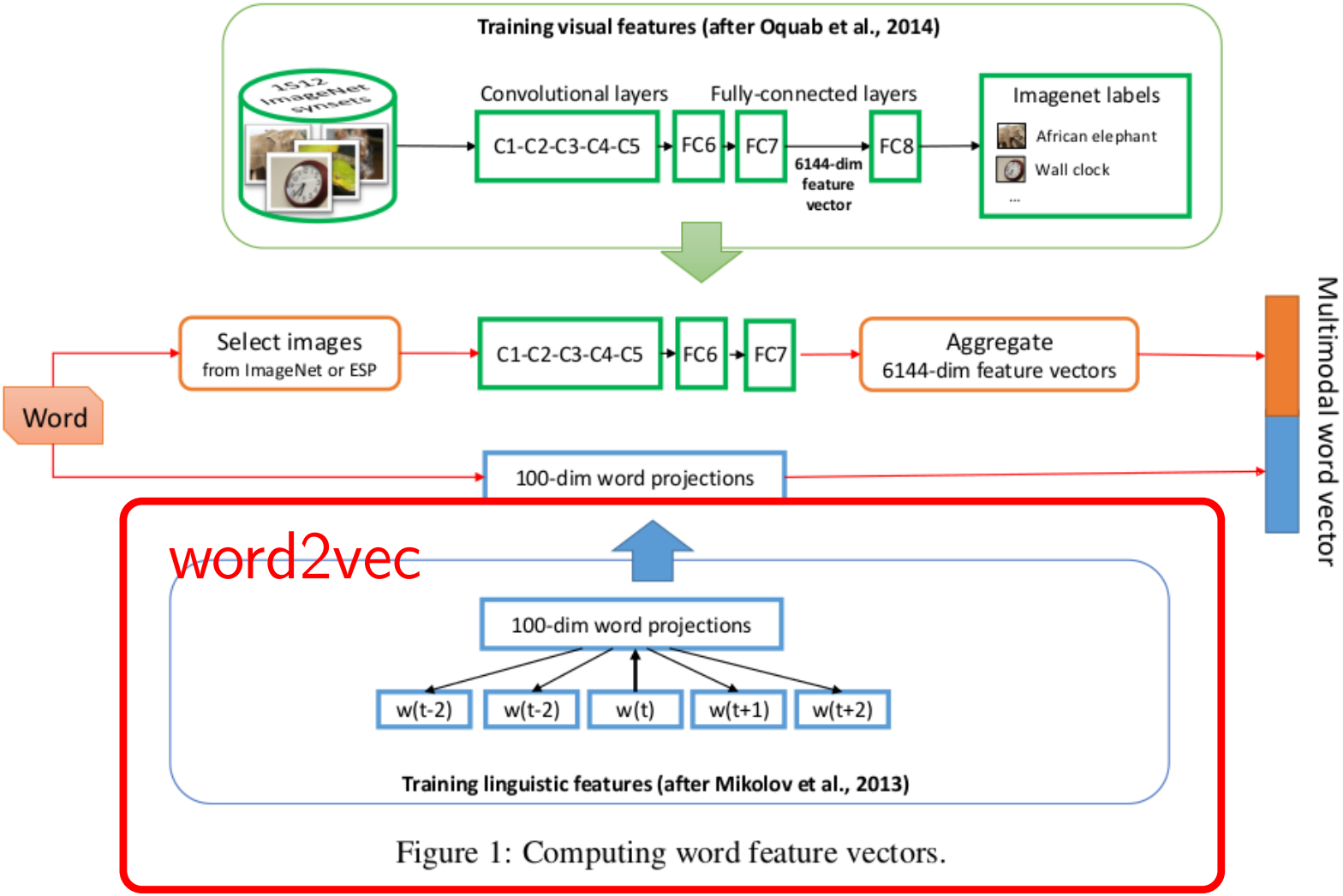


Figure 1: Computing word feature vectors.

Linguistische Repräsentation

- 100-dimensionale word2vec (skipgram) Vektoren
- Trainingsdaten:
  - Text8 Corpus (die ersten 10^8 Byte (100 MB) von wikipedia): 400 Millionen Worte [2]
  - BNC (100 Millionen Worte) [3]
- Mikolov et al. haben bereits 300 dim. Vektoren auf 783 Millionen Worten trainiert!

[2] http://www.mattmahoney.net/dc/textdata.html

[3] ota.ox.ac.uk/desc/2554

Modell

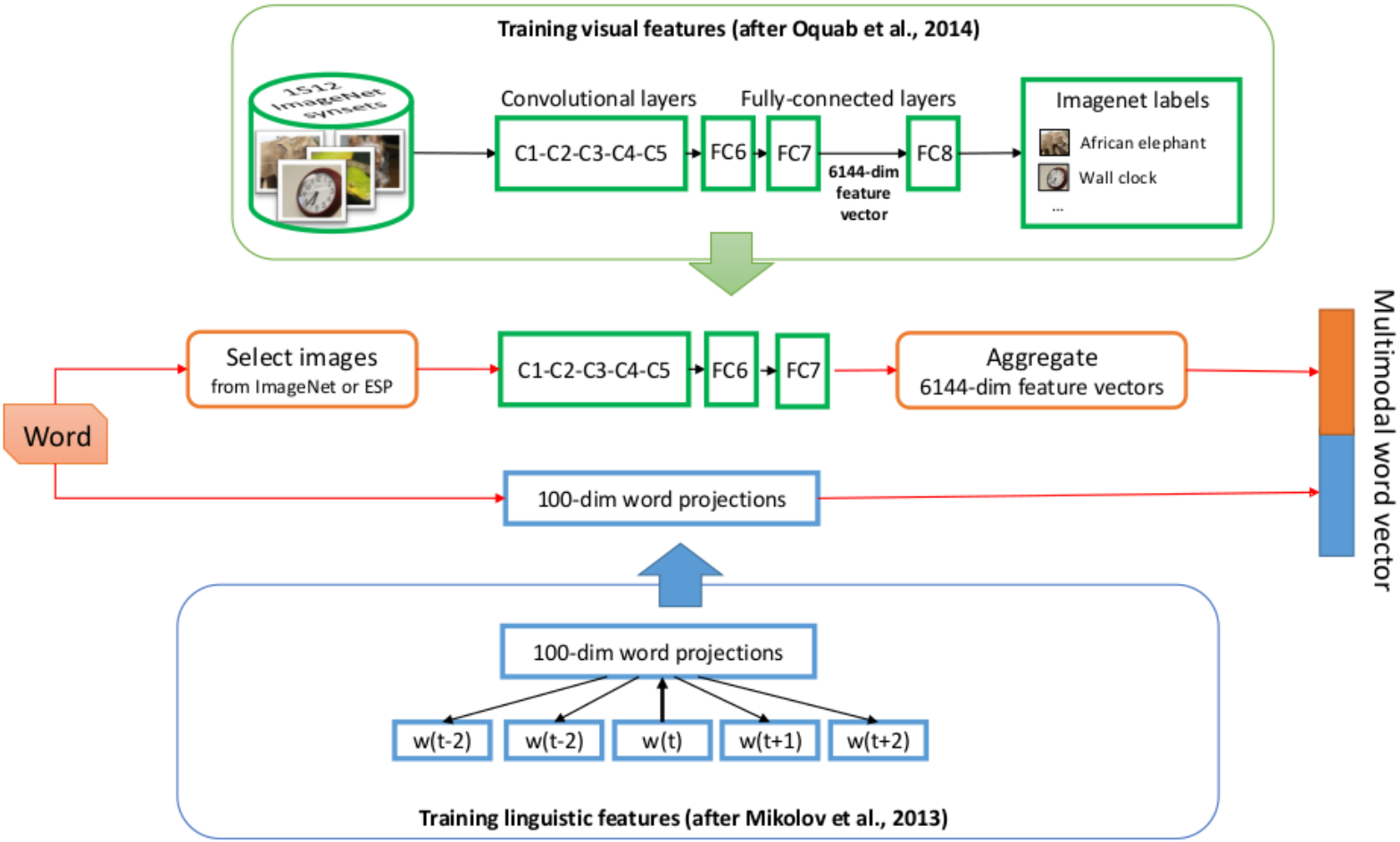


Figure 1: Computing word feature vectors.

Modell

ImageNet

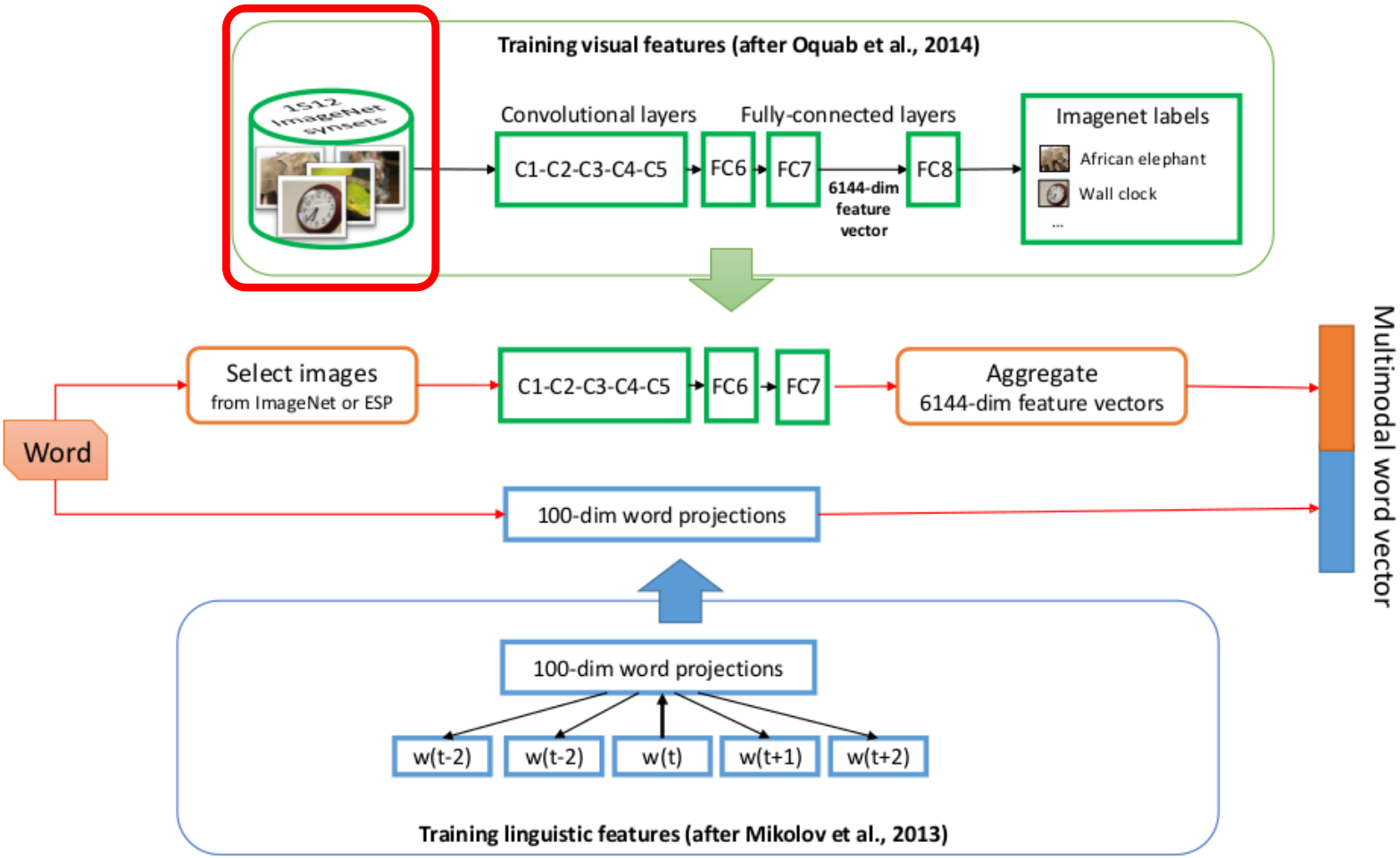


Figure 1: Computing word feature vectors.

ImageNet I

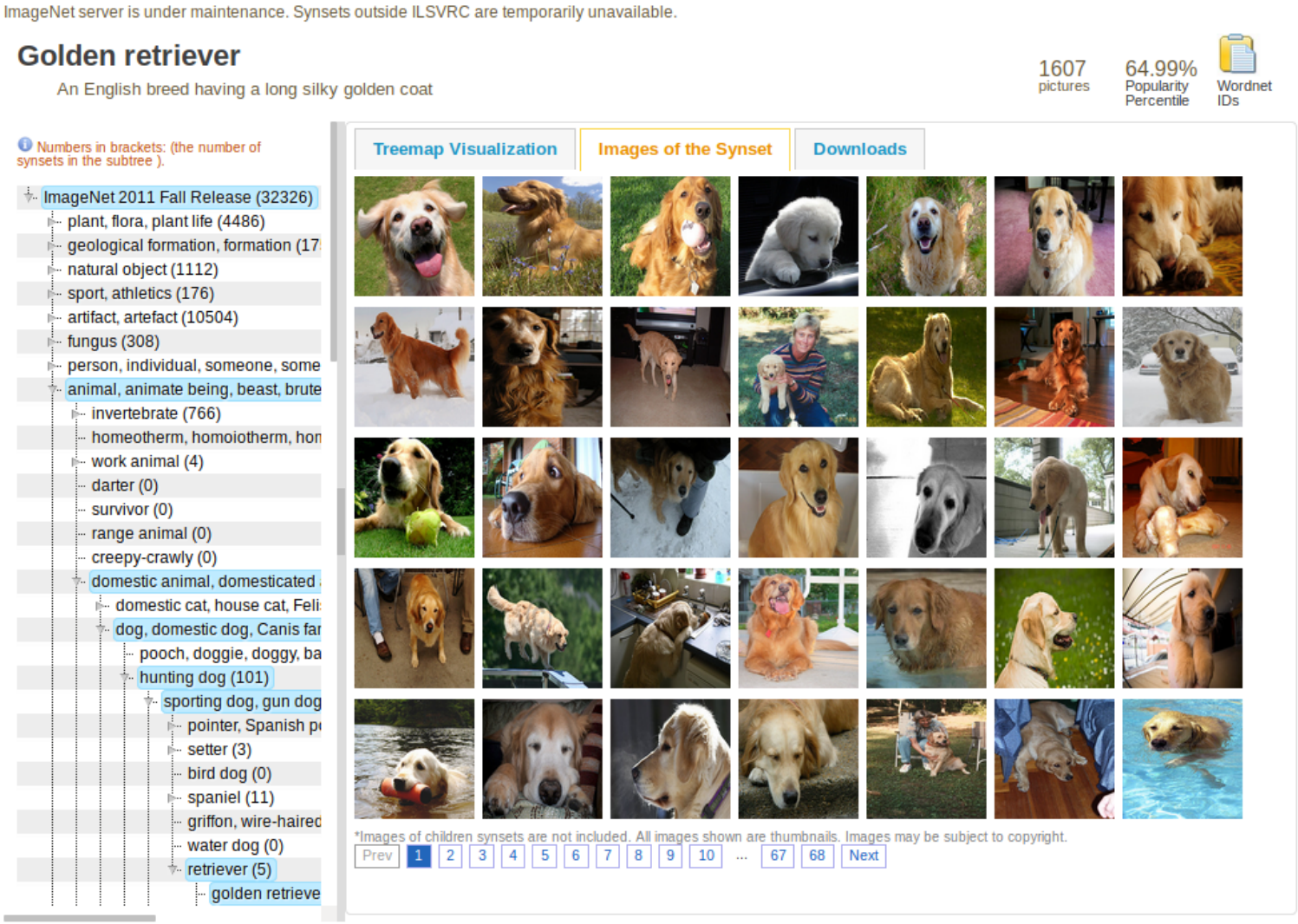
- entwickelt in Stanford, erste Version aus dem Jahr 2009
- Vorbild: WordNet $\rightarrow$ identische Struktur (synsets), Hierarchie
- sämtliche synsets sind Nomen
- Ziel für ImageNet: 500-1000 Bilder pro synset (50 Millionen insgesamt) [4]

[4] http://image-net.org/papers/imagenet_cvpr09.pdf

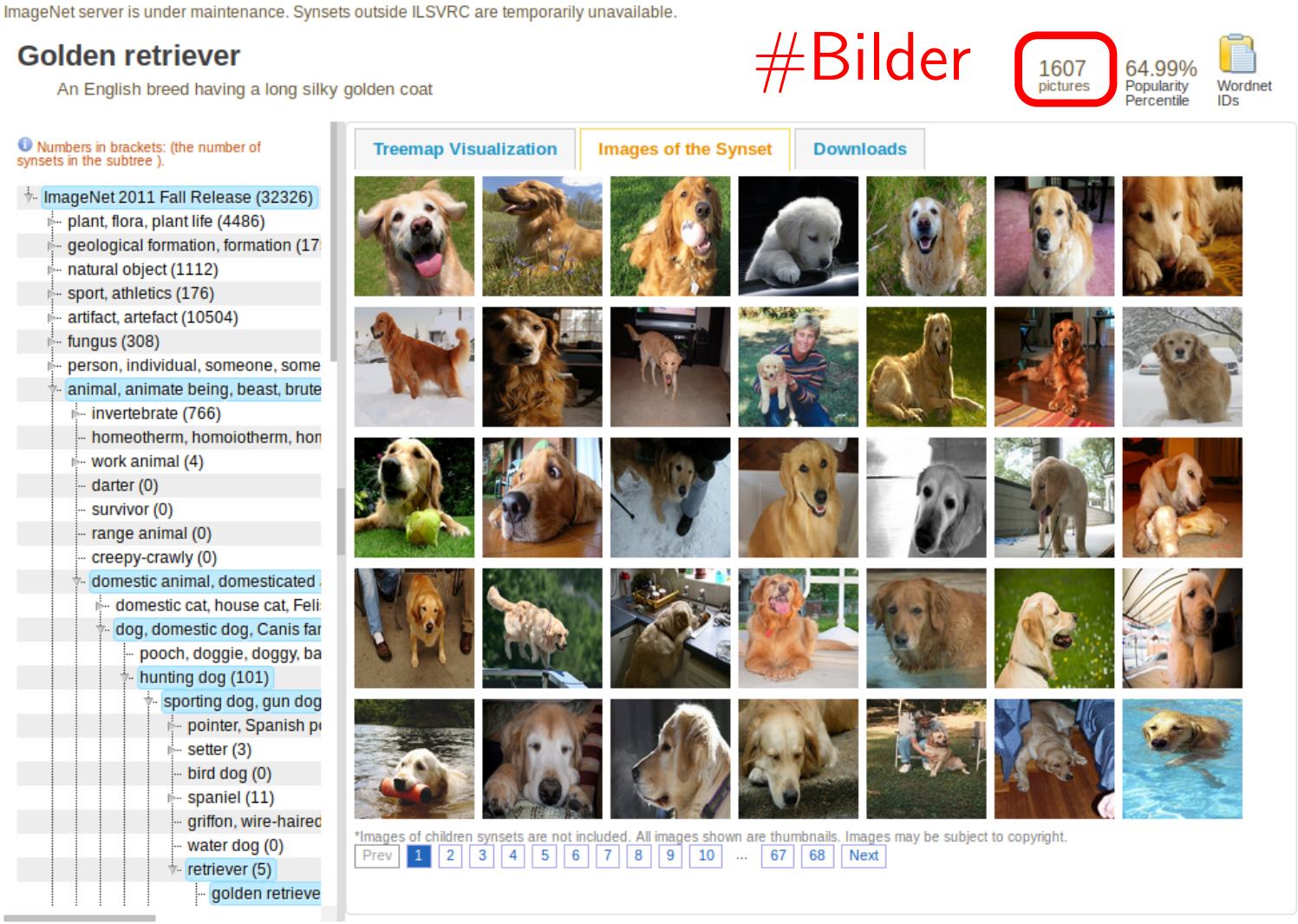
ImageNet II

- Stand Juni 2019: $\sim$14 Millionen Bilder für $\sim$22.000 synsets
- hier: $\sim$12.5 Millionen Bilder für $\sim$22.000 synsets
- manuelle Annotation mittels Amazon Mechanical Turk
- Webseite hostet nur Thumbnails, die Bilder müssen von anderen Seiten geladen werden!

ImageNet III



ImageNet III



ImageNet III

ImageNet III

ImageNet IV

Beispiel: Golden Retriever

- 500 x 483 Pixel
- mehr Mensch als Hund zu sehen

- 500 x 458 Pixel
- man sieht nur den Kopf

ESP Game Datensatz

- entwickelt an der CMU, erste Veröffentlichung 2004
- Idee: Menschen labeln Bilder in einem Spiel

Unterschiede ImageNet - ESP Game

ImageNet:

- 22.000 Worte (synsets)
- 12.5 Millionen Bilder
- 1 tag pro Bild(?)
- ordentliche Qualität
- Objekt normalerweise zentriert und gut zu erkennen

ESG Game:

- 20515 Worte
- 100.000 Bilder
- $\sim$14 tags pro Bild
- deutlich schlechtere Qualität
- Objekt kann auch im Hintergrund auftauchen

Modell

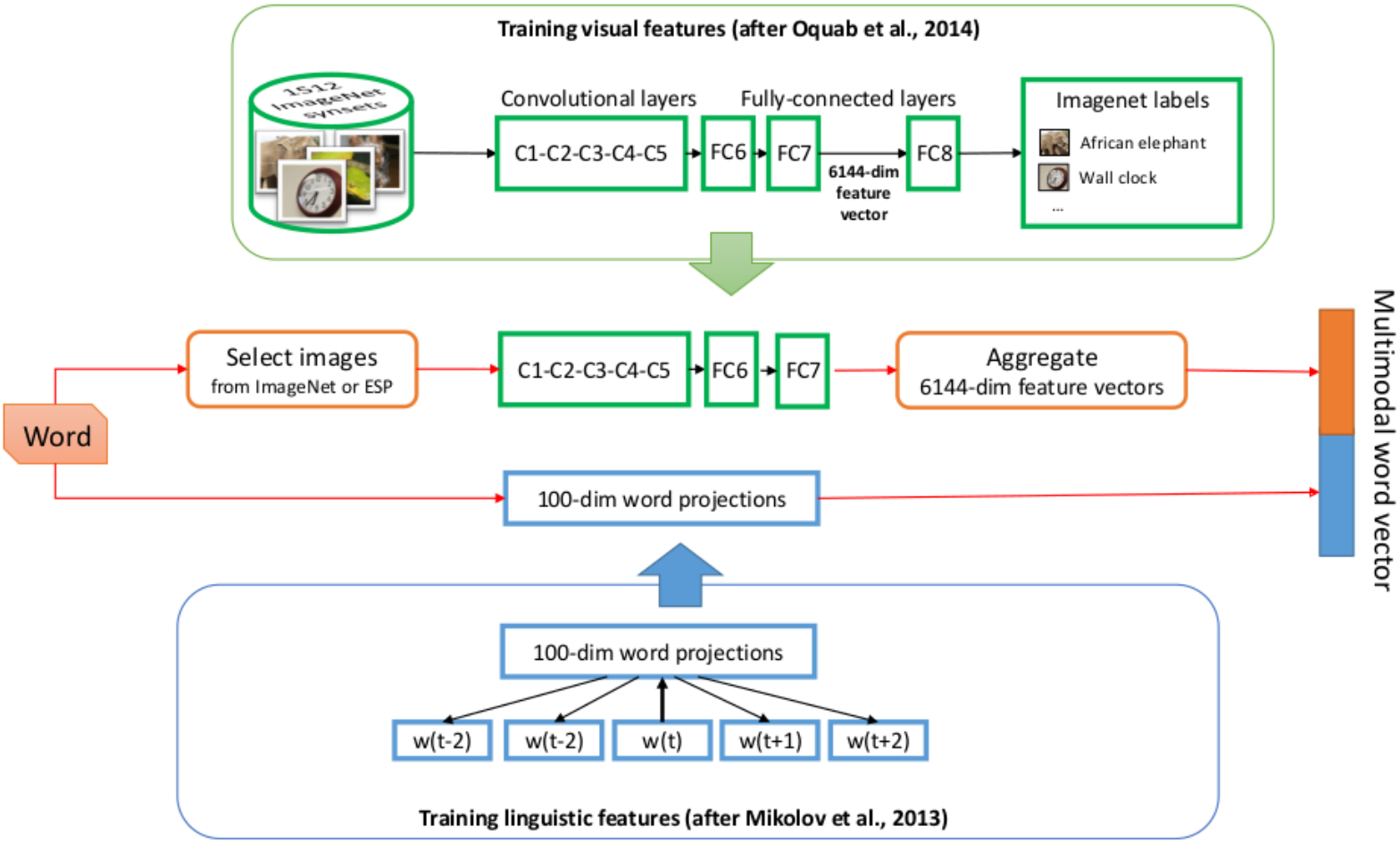


Figure 1: Computing word feature vectors.

Modell

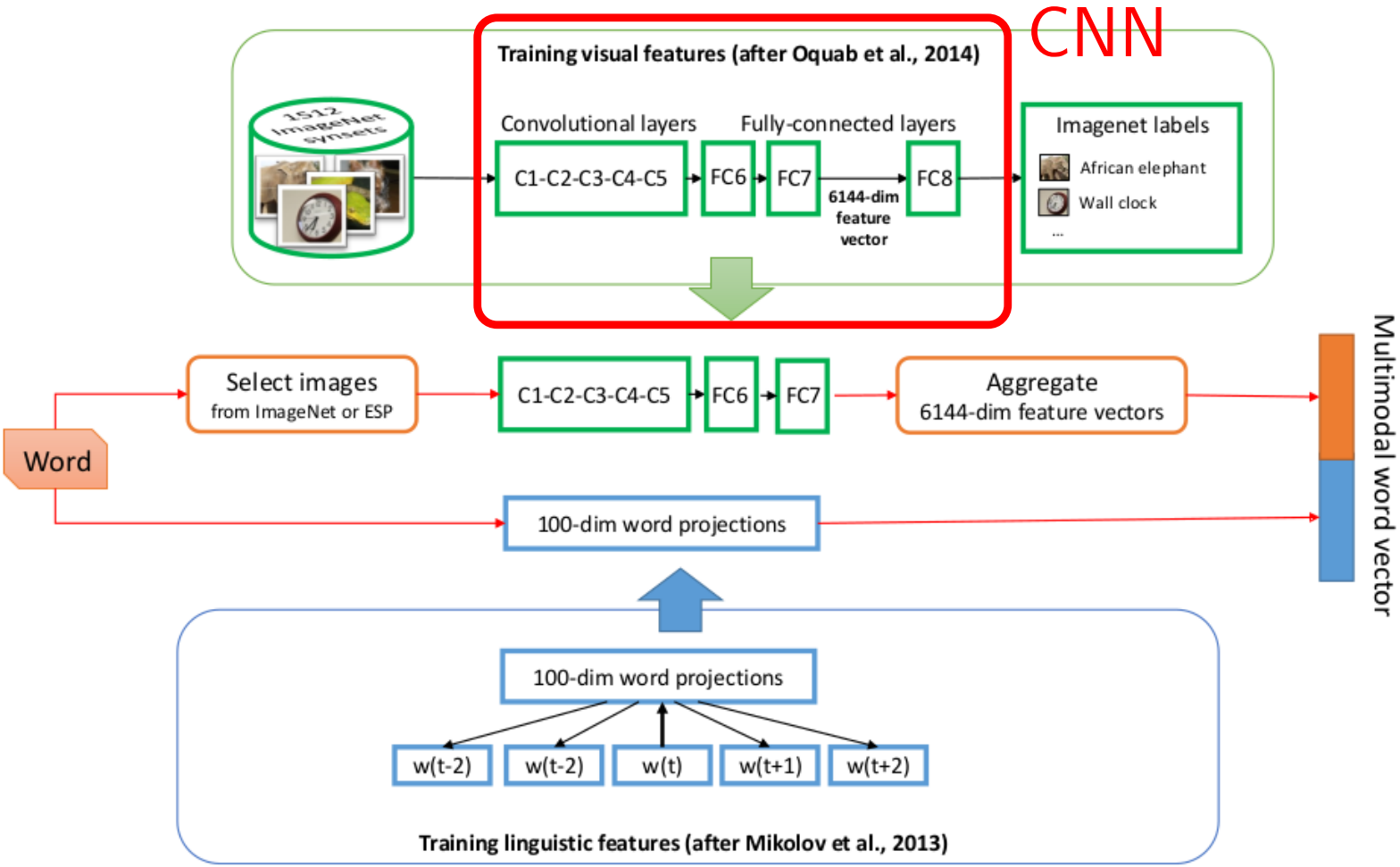


Figure 1: Computing word feature vectors.

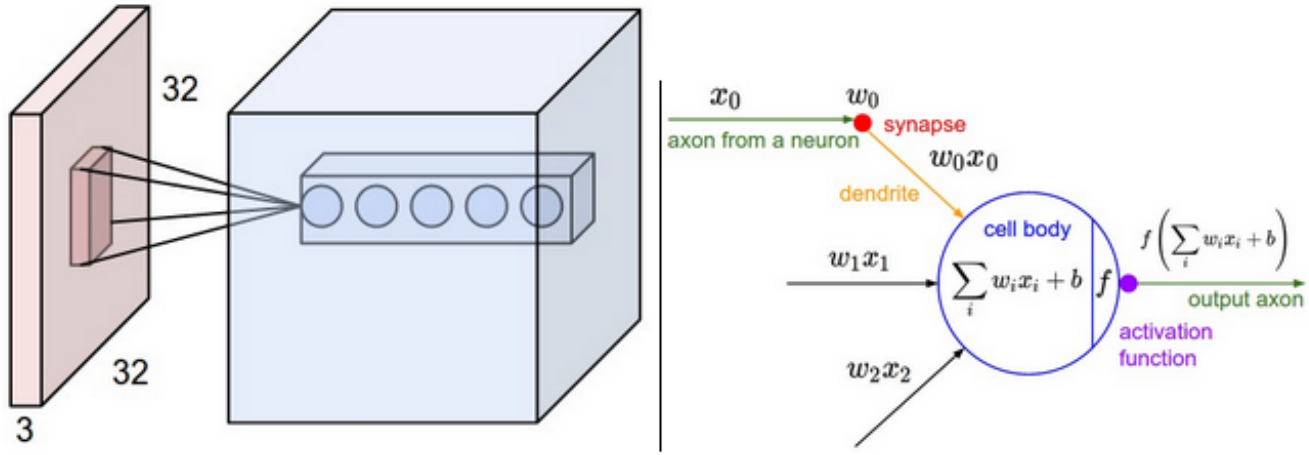
CNN

- basiert auf AlexNet
  - entwickelt von Alex Krizhevsky, veröffentlicht im Jahr 2012
  - laut google scholar über 42.000 mal zitiert [5]
- 5 convolution layer (Cx), gefolgt von 3 fully connected (FCx) layern
- das vorletzte layer (FC7) mit 6144 Gewichten wird benutzt zur Konstruktion der multi-modalen Repräsentation (*transfer learning*)

[5] Google scholar

CNN - convolution layer

Standford CD 231n class



Left: An example input volume in red (e.g. a 32x32x3 CIFAR-10 image), and an example volume of neurons in the first Convolutional layer. Each neuron in the convolutional layer is connected only to a local region in the input volume spatially, but to the full depth (i.e. all color channels). Note, there are multiple neurons (5 in this example) along the depth, all looking at the same region in the input - see discussion of depth columns in text below. **Right:** The neurons from the Neural Network chapter remain unchanged: They still compute a dot product of their weights with the input followed by a non-linearity, but their connectivity is now restricted to be local spatially.

CNN - kurzer Überblick

- zwei Teile:
  - der vordere Teil besteht prinzipiell nur aus convolution und max-pooling layern
  - der hintere Teil ist ein fully connected MLP
- in den convolution layern läuft ein Filter zeilenweise über das Bild, dann wird das dot product aus den Filtergewichten und Pixelwerten berechnet
- Filter (fast) immer quadratisch, Aktivierungsfunktion in den convolutional layern (fast) immer ReLU

Modell

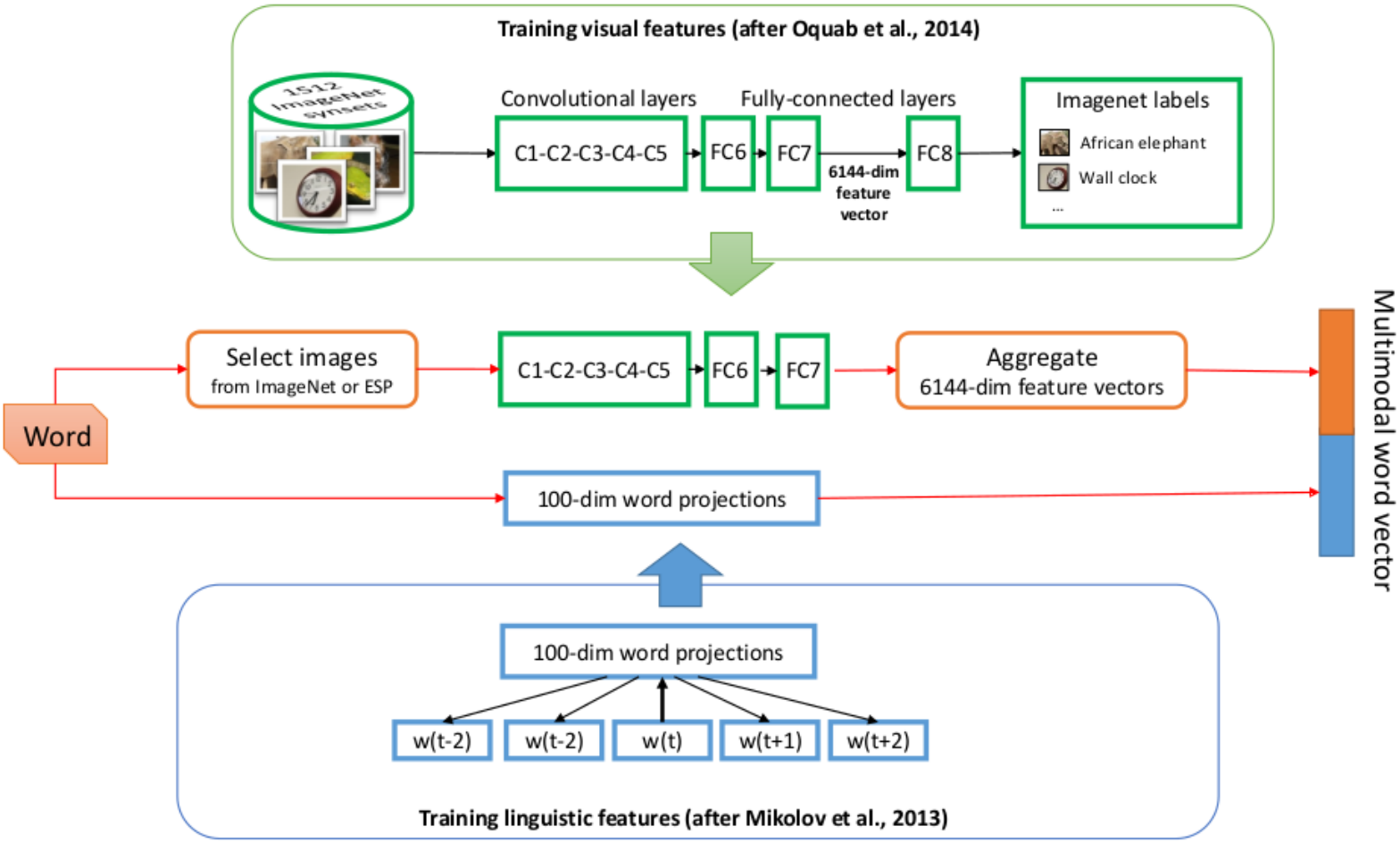


Figure 1: Computing word feature vectors.

Modell

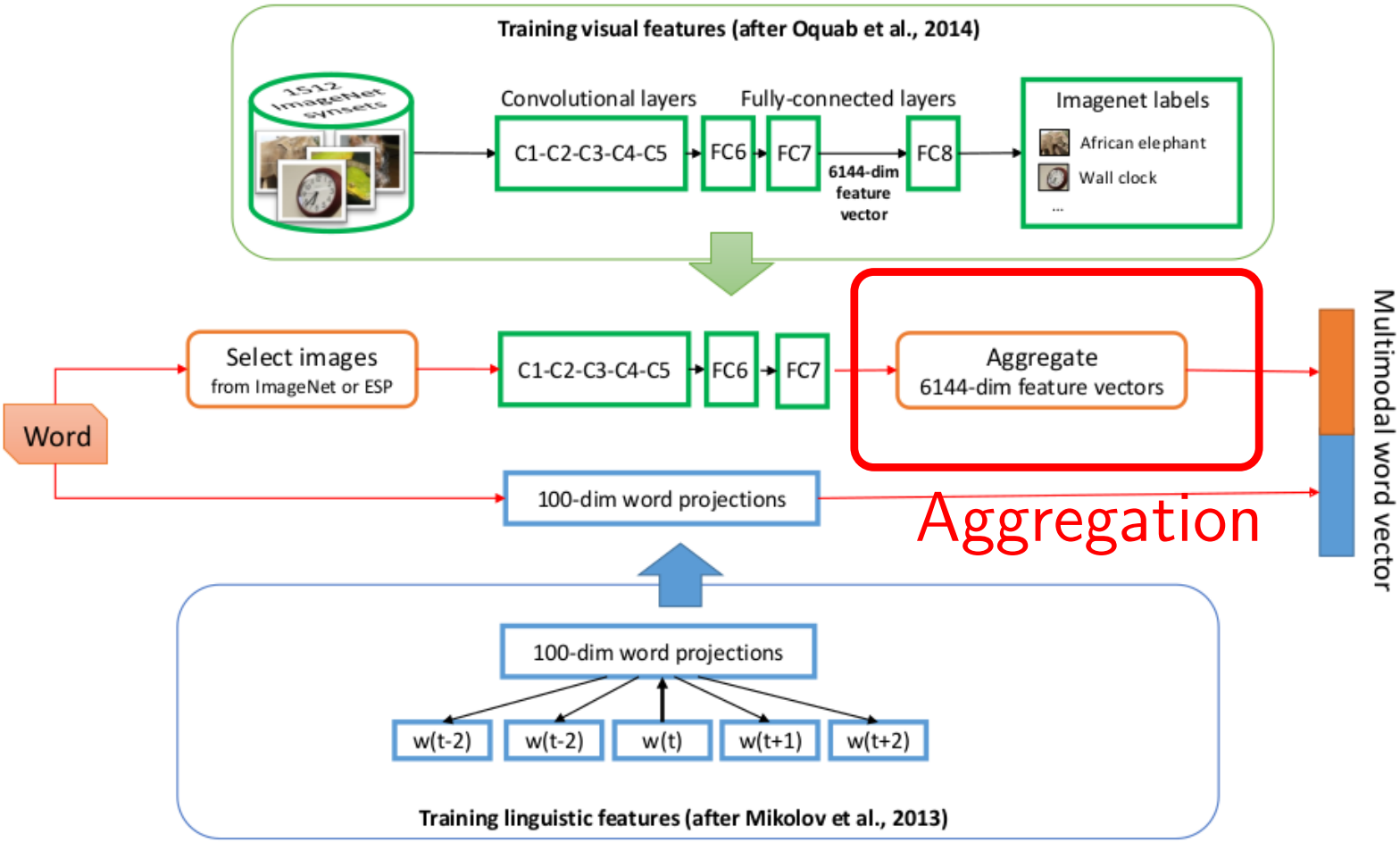


Figure 1: Computing word feature vectors.

Aggregation

- die trainierten perzeptuellen Vektoren werden auf zwei Arten aggregiert:
  1. **CNN-Mean** berechnet für jedes x_i den Durchschnitt aller Vektoren
  2. **CNN-Max** nimmt für jedes x_i das Maximum aller Vektoren $\rightarrow$ *bag of visual properties*

Modell

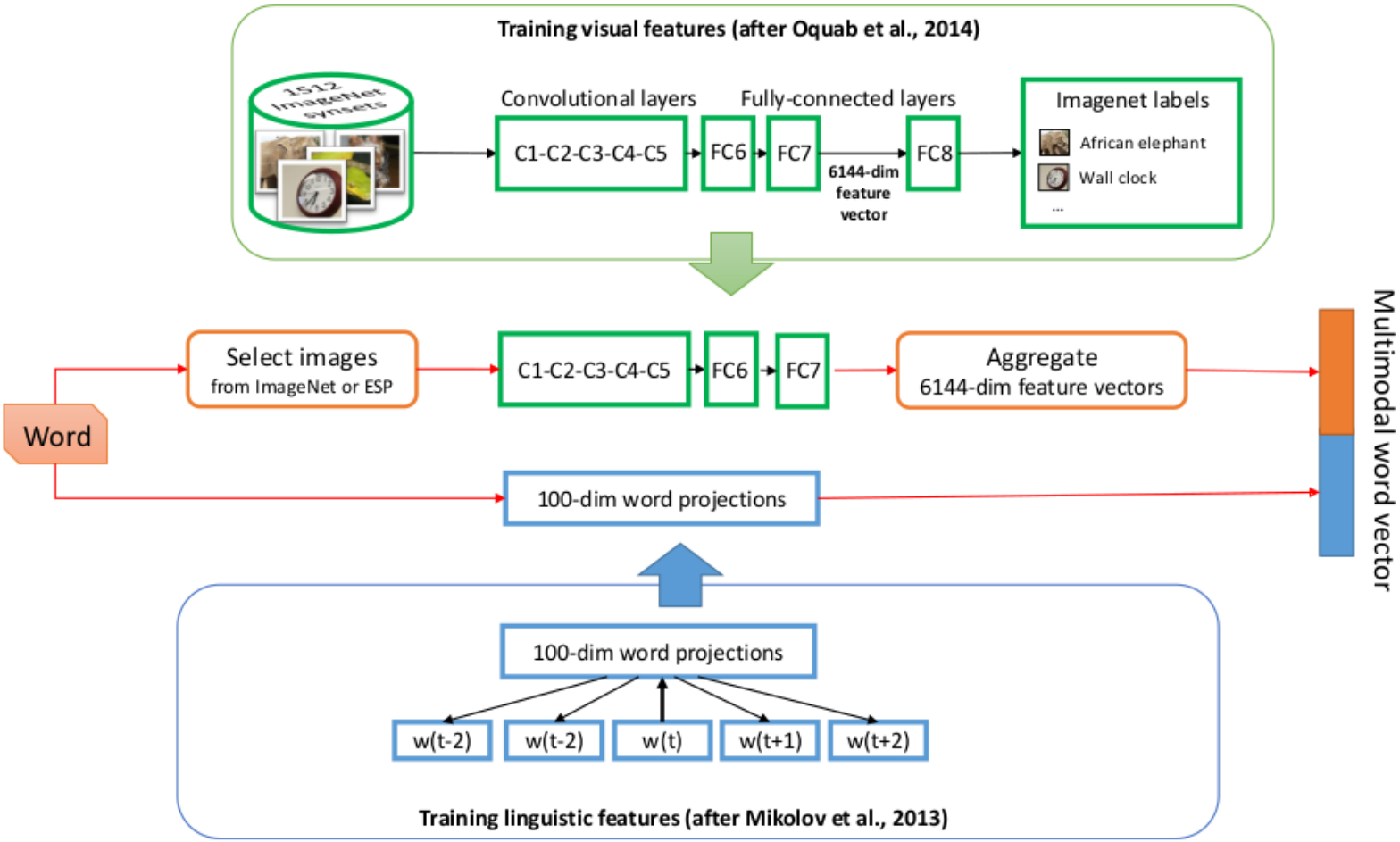


Figure 1: Computing word feature vectors.

Modell

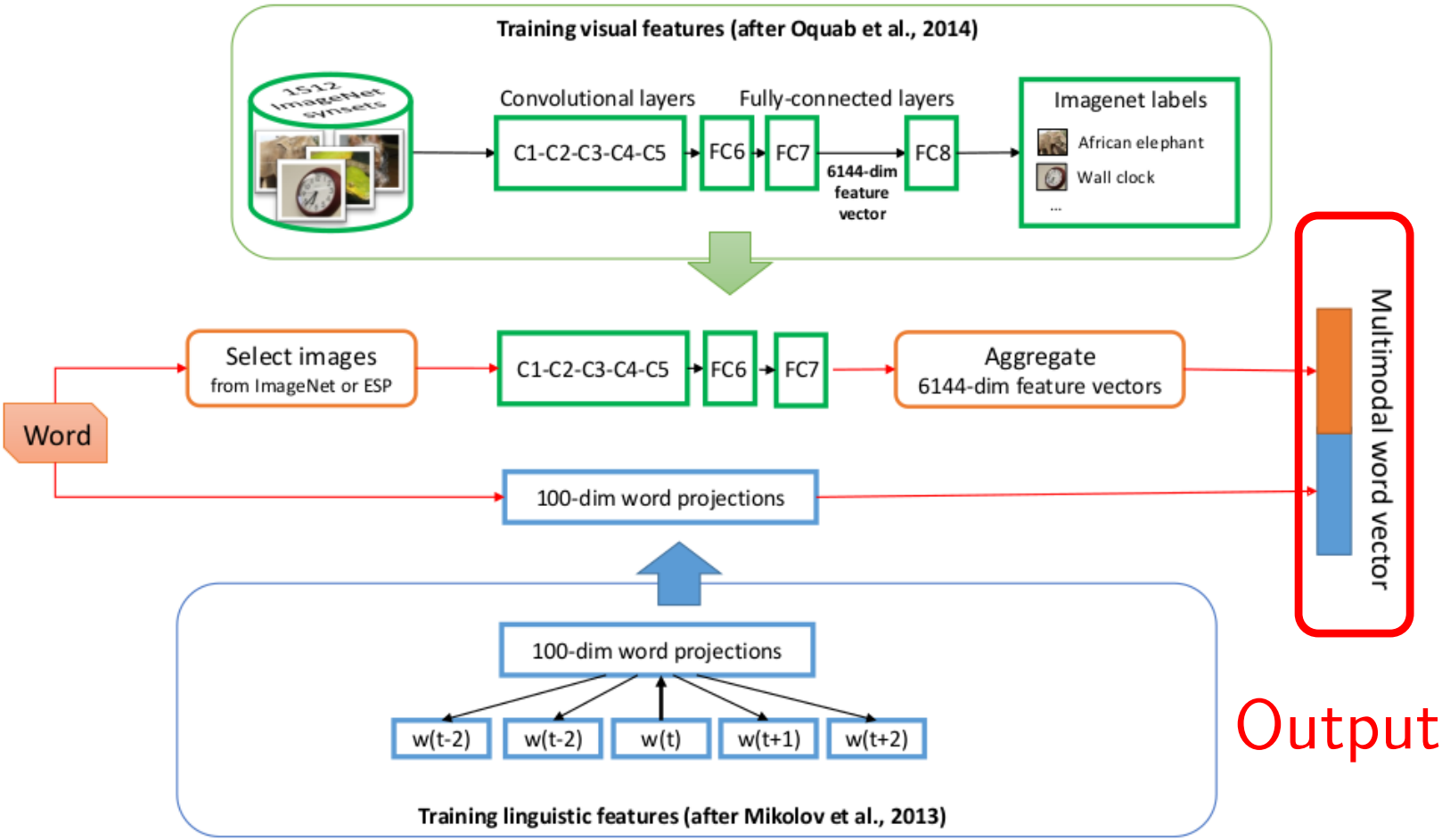


Figure 1: Computing word feature vectors.

Multimodale Repräsentation

$$\vec{v}_{concept} = \alpha \times \vec{v}_{ling} \parallel (1 - \alpha) \times \vec{v}_{vis} \quad (1)$$

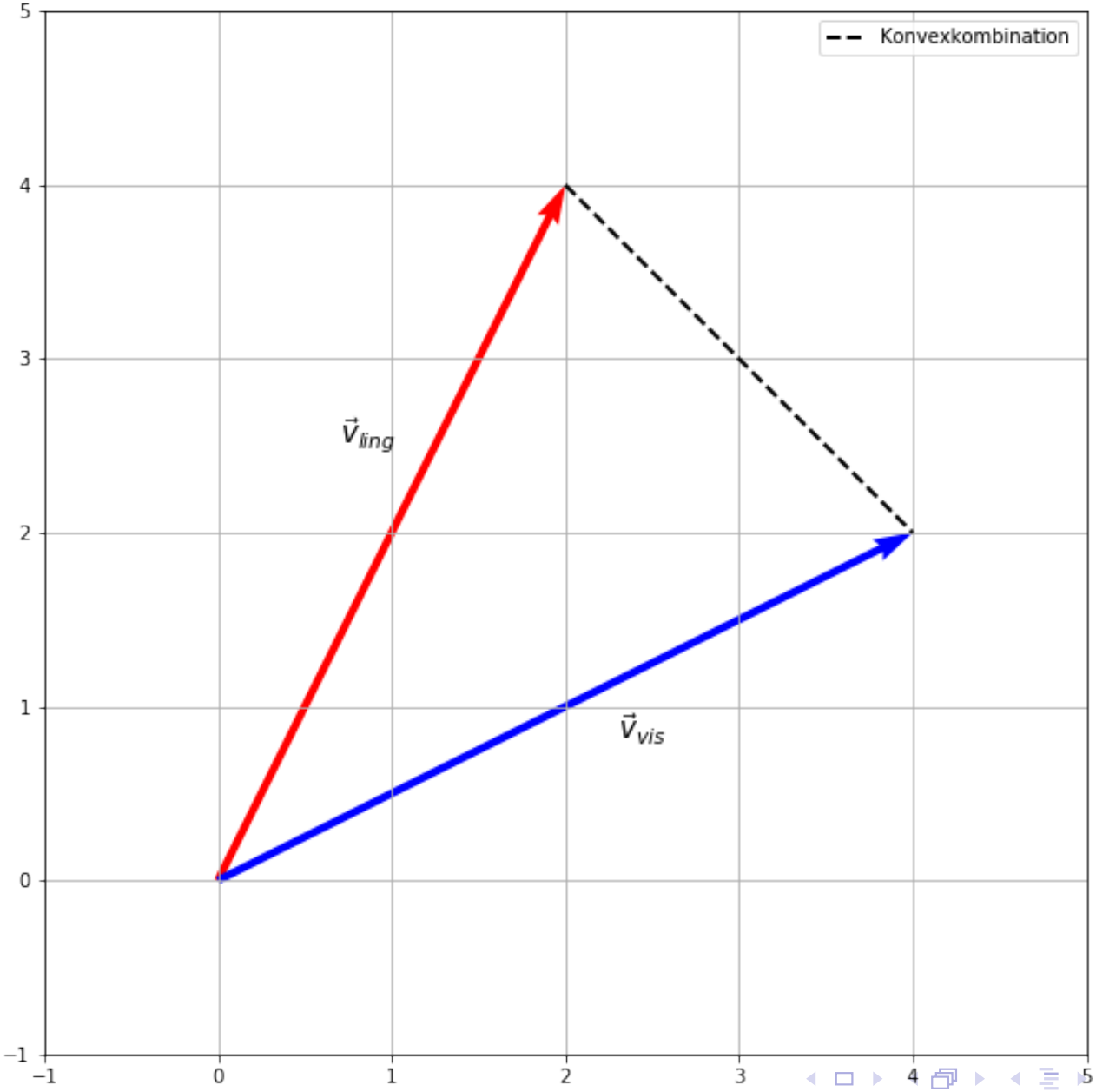
Multimodale Repräsentation

$$\vec{v}_{concept} = \alpha \times \vec{v}_{ling} \parallel (1 - \alpha) \times \vec{v}_{vis} \tag{1}$$

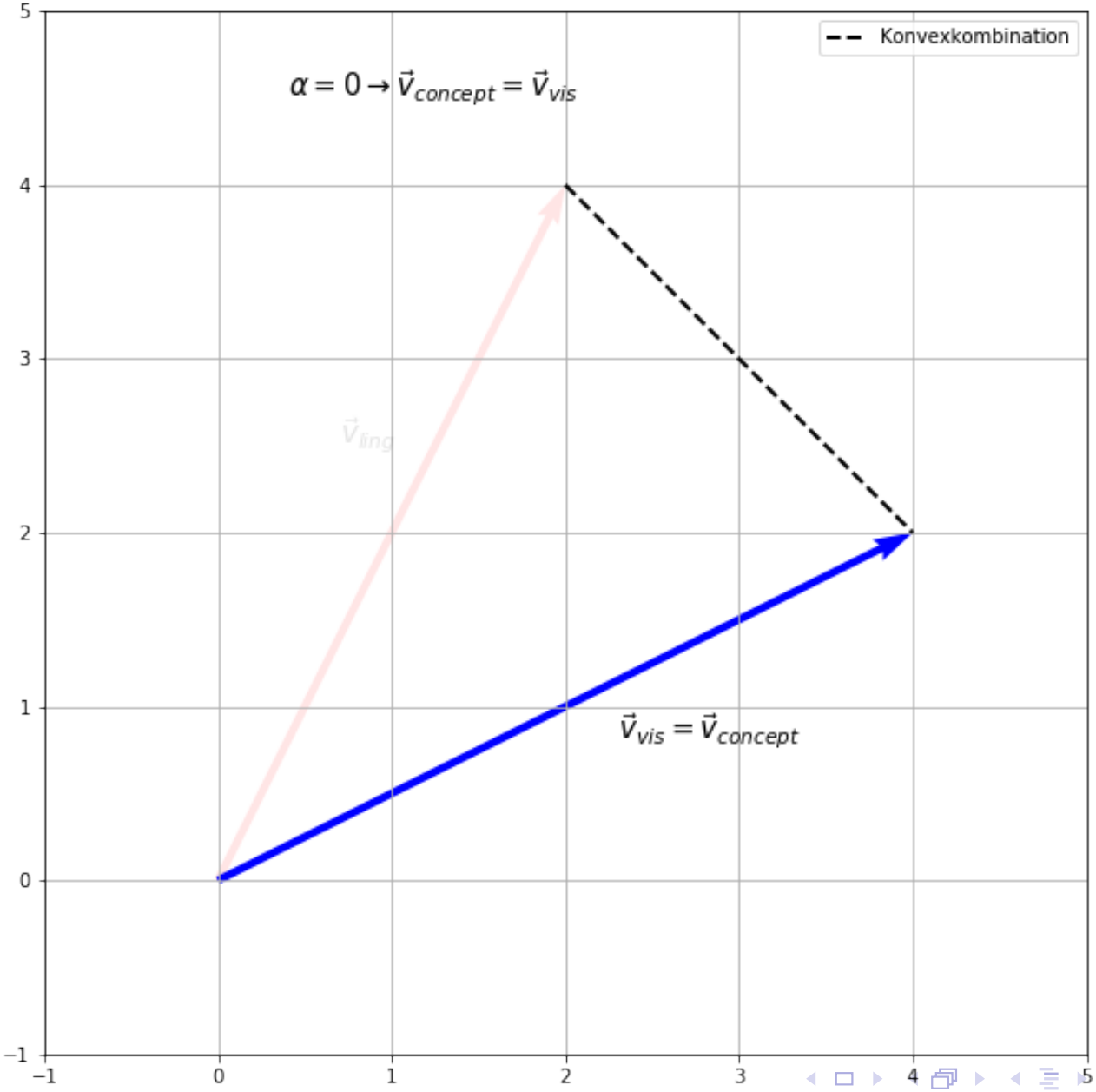
- man kann die Formel als eine Konvexkombination der beiden Vektoren $\vec{v}_{ling}, \vec{v}_{vis}$ auffassen:

$$\vec{v}_{ling} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{100} \\ x_{101} = 0 \\ \vdots \\ x_{6244} = 0 \end{bmatrix} \quad \vec{v}_{vis} = \begin{bmatrix} x_1 = 0 \\ x_2 = 0 \\ \vdots \\ x_{100} = 0 \\ x_{101} \\ \vdots \\ x_{6244} \end{bmatrix}$$

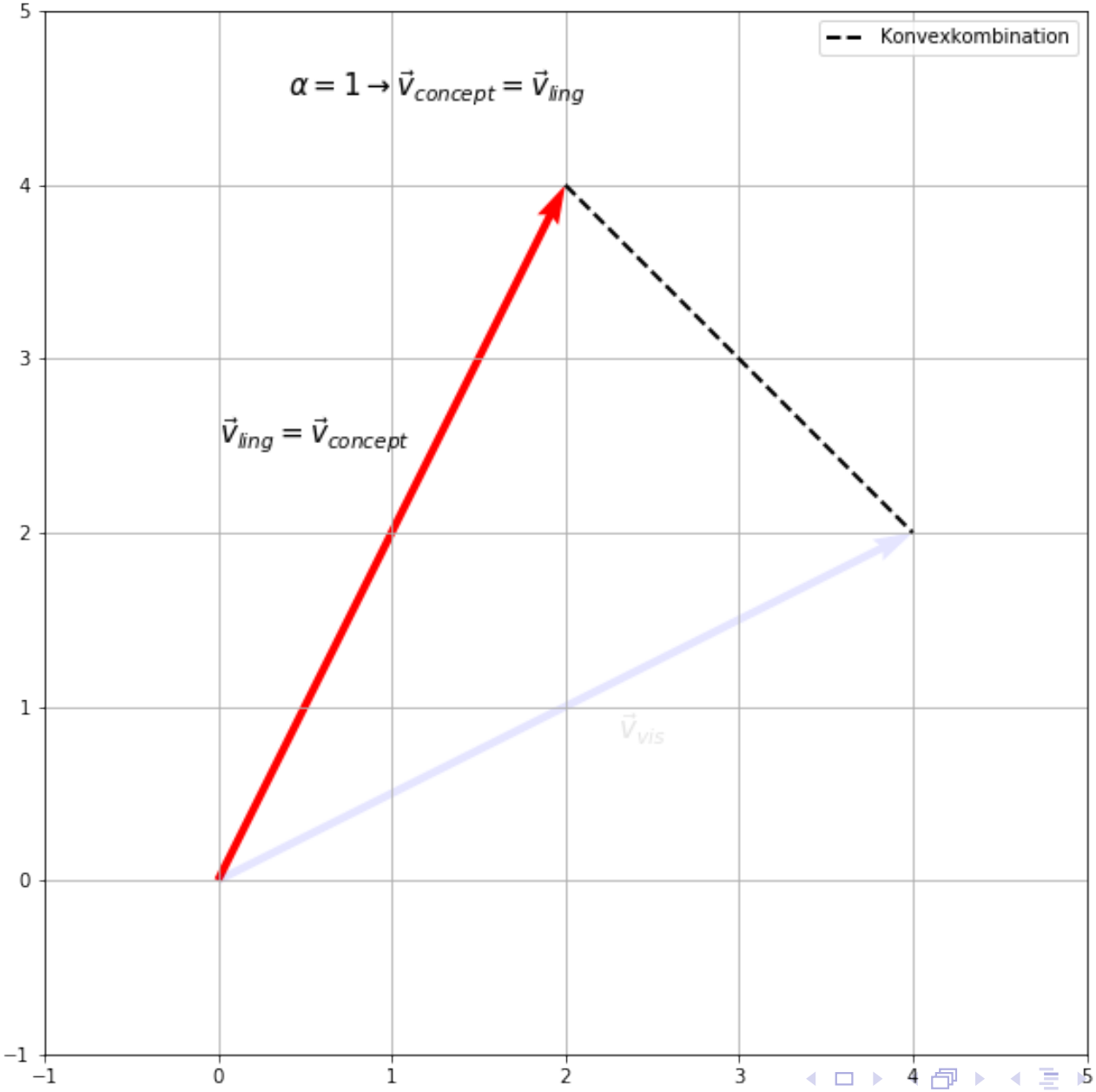
Konvexkombination



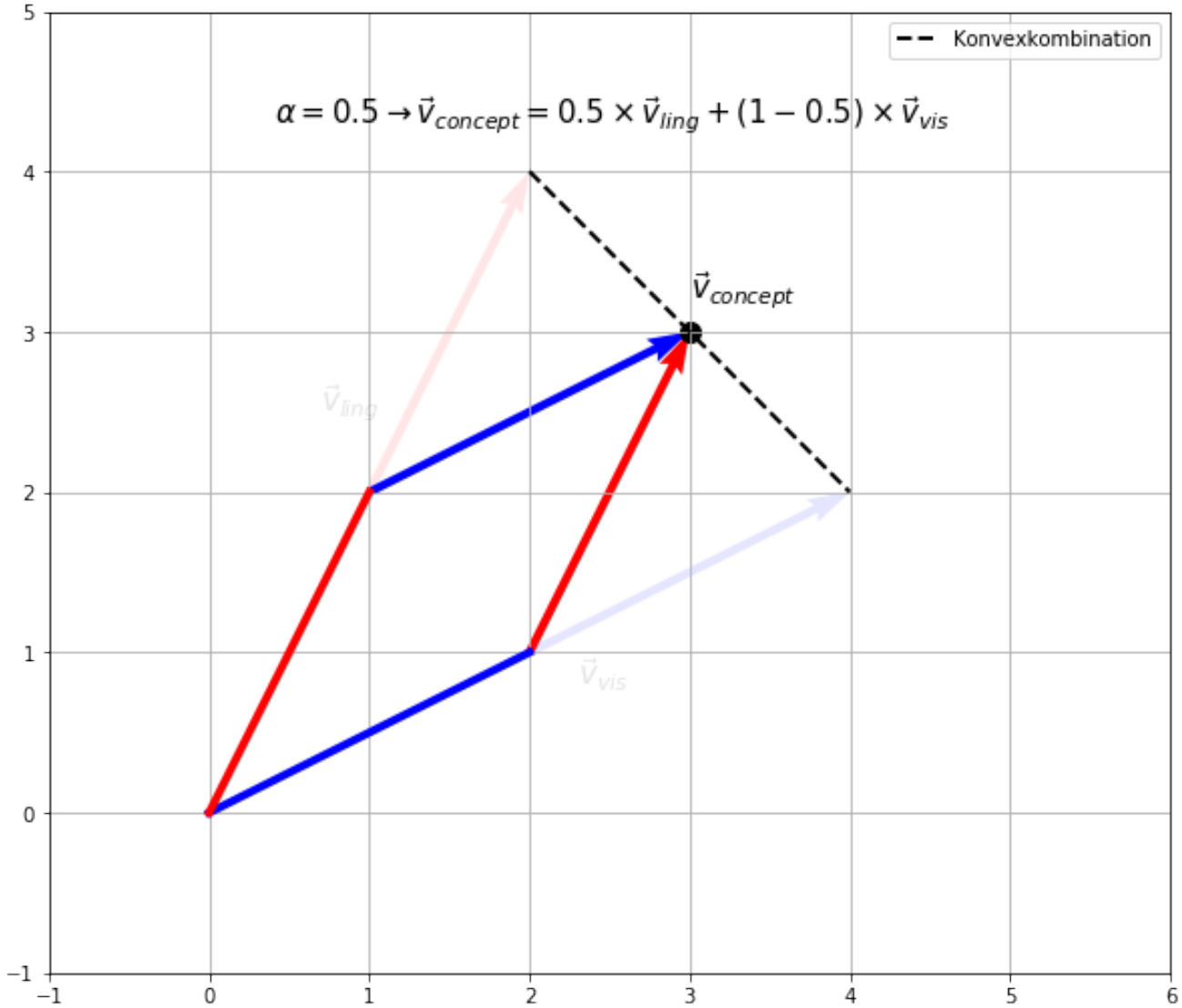
Konvexkombination



Konvexkombination



Konvexkombination



2. Frage - multimodale Repräsentation

$$\vec{v}_{concept} = \alpha \times \vec{v}_{ling} \parallel (1 - \alpha) \times \vec{v}_{vis} \quad (1)$$

2. Frage - multimodale Repräsentation

$$\vec{v}_{concept} = \alpha \times \vec{v}_{ling} \parallel (1 - \alpha) \times \vec{v}_{vis} \tag{1}$$

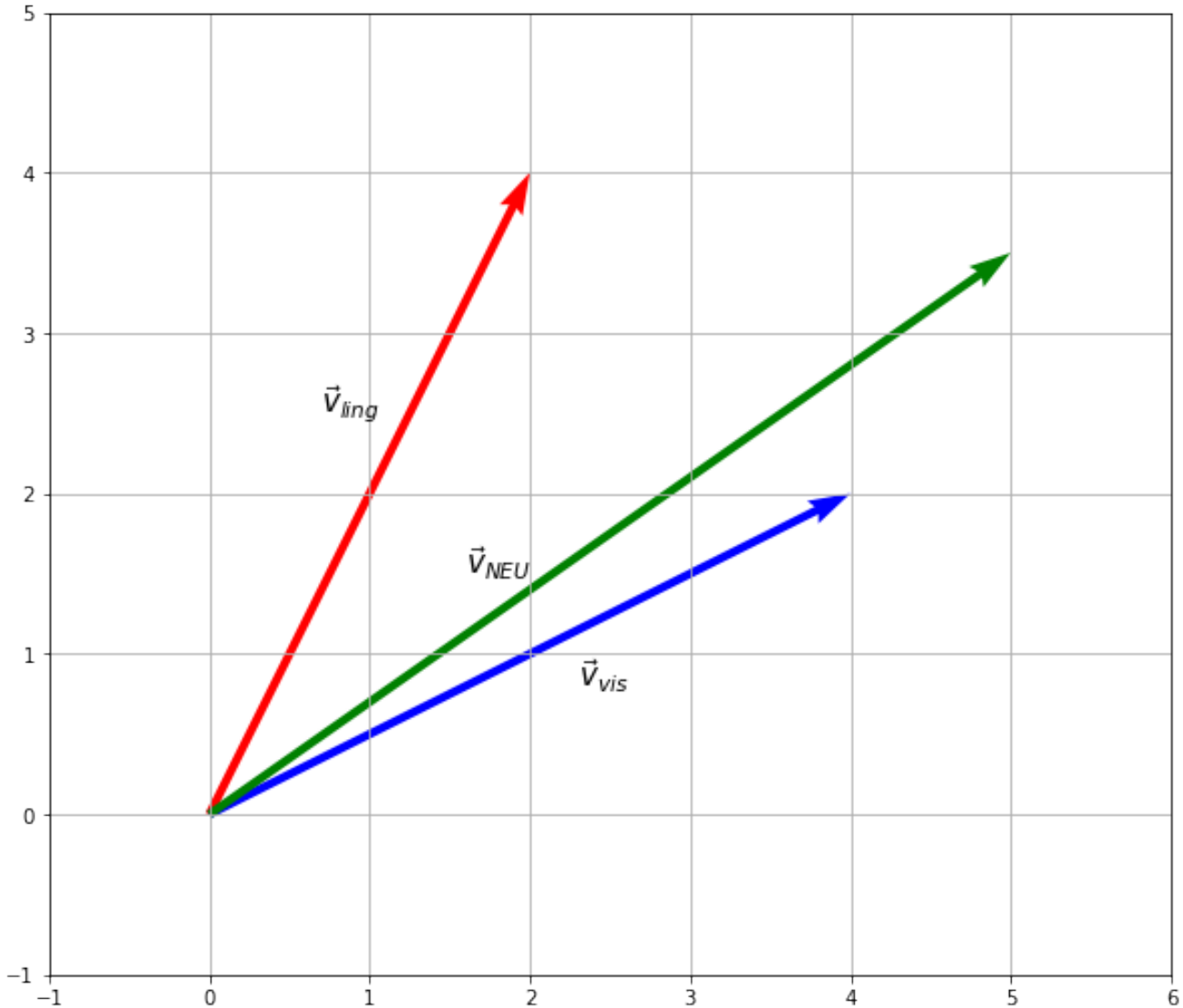
- Wie sähe Formel 1 aus, wenn man noch eine dritte Modalität wie Audio hinzufügen würde?

$$\alpha \times \vec{v}_{ling} \parallel \beta \times \vec{v}_{NEU} \parallel \gamma \times \vec{v}_{vis} = \vec{v}_{concept} \tag{2}$$

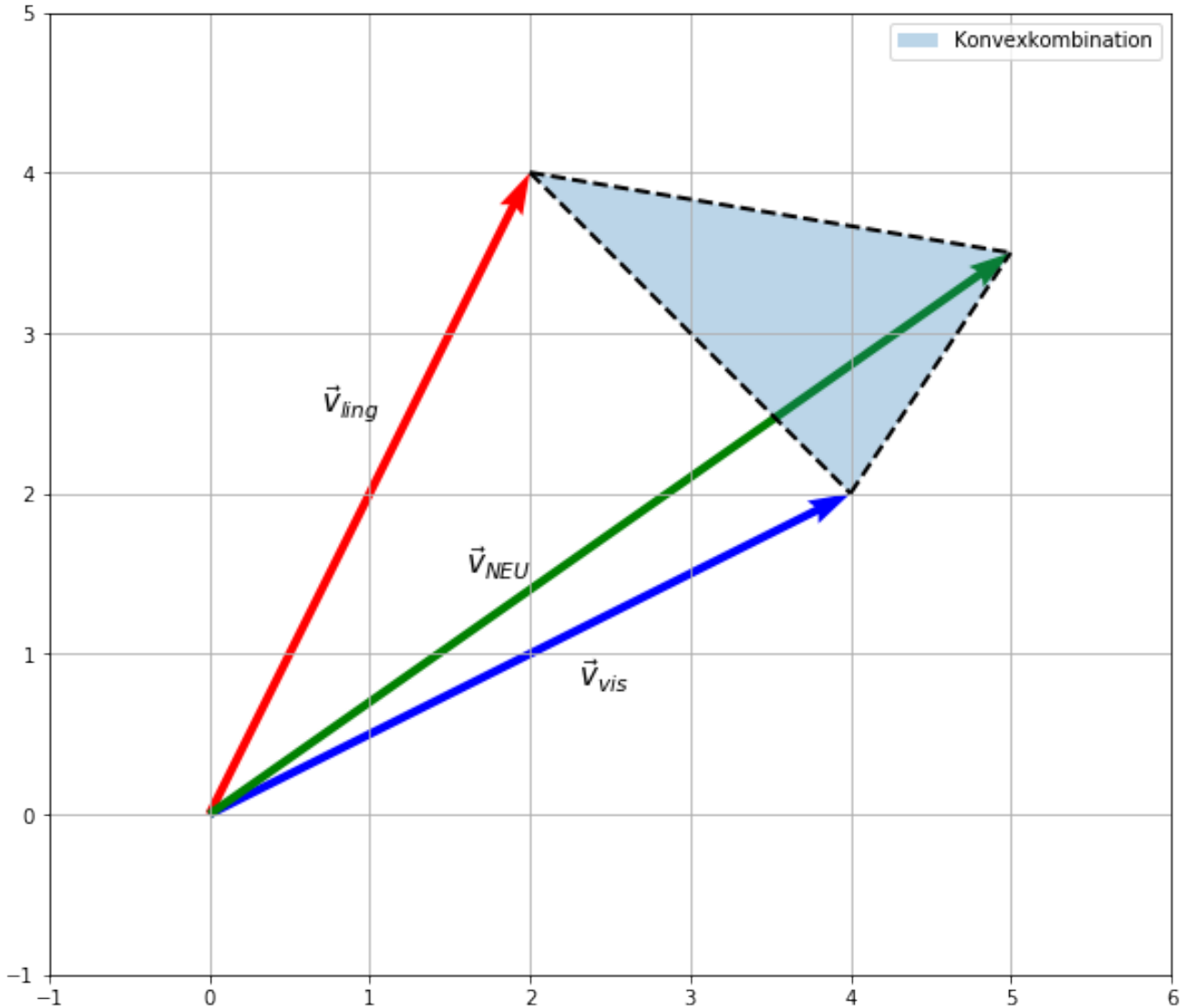
$$\alpha + \beta + \gamma = 1 \tag{3}$$

$$\alpha, \beta, \gamma \geq 0 \tag{4}$$

Konvexkombination für 3 Vektoren



Konvexkombination für 3 Vektoren



Modell - Zusammenfassung

- linguistische Repräsentation: 100 dim. Vektor(word2vec, skipgram), trainiert auf wikipedia-Artikeln/dem BNC
- perzeptuelle Repräsentation: 6144 dim. Vektor (CNN, trainiert auf ImageNet- und ESP Game-Bildern)
- multimodale Repräsentation: Konvexkombination der beiden Vektoren

Inhaltsverzeichnis

Evaluation I

WordSim353 (Finkelstein et al., 2001)

- Auswahl aus 353 concept pairs mit menschlicher Bewertung der Ähnlichkeit
- Problem: enthält Worte, deren Darstellung schwierig ist:
  - Named Entities (*OPEC*)
  - abstrakte Begriffe (*credibility*)

Evaluation II

MEN (Bruni et al., 2012)

- Ziel: Probleme aus WordSim353 zu lindern
- Nur häufige Wörter mit mindestens 50 Wortpaaren im ESP Game
- Much Larger: 3000 Word pairs consisting of 751 individual words

Evaluation III

- Subsets: WordSim Relevant und MEN-Relevant: Bilder in beiden Datensets enthalten!
- Bewertung der Modelle anhand der Spearman-Korrelation
- Ähnlichkeit zwischen der Repräsentation über Kosinus-Ähnlichkeit:

$$\cos(v_1, v_2) = \frac{v_1 \cdot v_2}{\|v_1\| \|v_2\|}$$

Inhaltsverzeichnis

Kernfragen

- Liefern CNNs bessere Ergebnisse für perzeptuelle Repräsentationen?
- Liefern CNNs bessere Ergebnisse für multimodale Repräsentationen?
- Was für eine Auswirkung auf das Ergebnis hat die Beschaffenheit des verwendeten Datensets?

Ergebnisse

Dataset	Linguistic	Visual			Multi-modal		
		BOVW	CNN-Mean	CNN-Max	BOVW	CNN-Mean	CNN-Max
ImageNet visual features							
MEN	0.64	-	-	-	0.64	**0.70**	0.67
MEN-Relevant	0.62	0.40	0.64	0.63	0.64	**0.72**	0.71
W353	0.57	-	-	-	0.58	0.59	**0.60**
W353-Relevant	0.51	0.30	0.32	0.30	0.55	0.56	**0.57**
ESP game visual features							
MEN	0.64	0.17	0.51	0.20	0.64	**0.71**	0.65
MEN-Relevant	0.62	0.35	0.58	0.57	0.63	0.69	**0.70**
W353	0.57	-	-	-	0.58	0.59	**0.60**
W353-Relevant	0.51	0.38	0.44	0.56	0.52	0.55	**0.61**

Table 1: Results (see sections 4 and 5).

Einfluss des α-Parameter

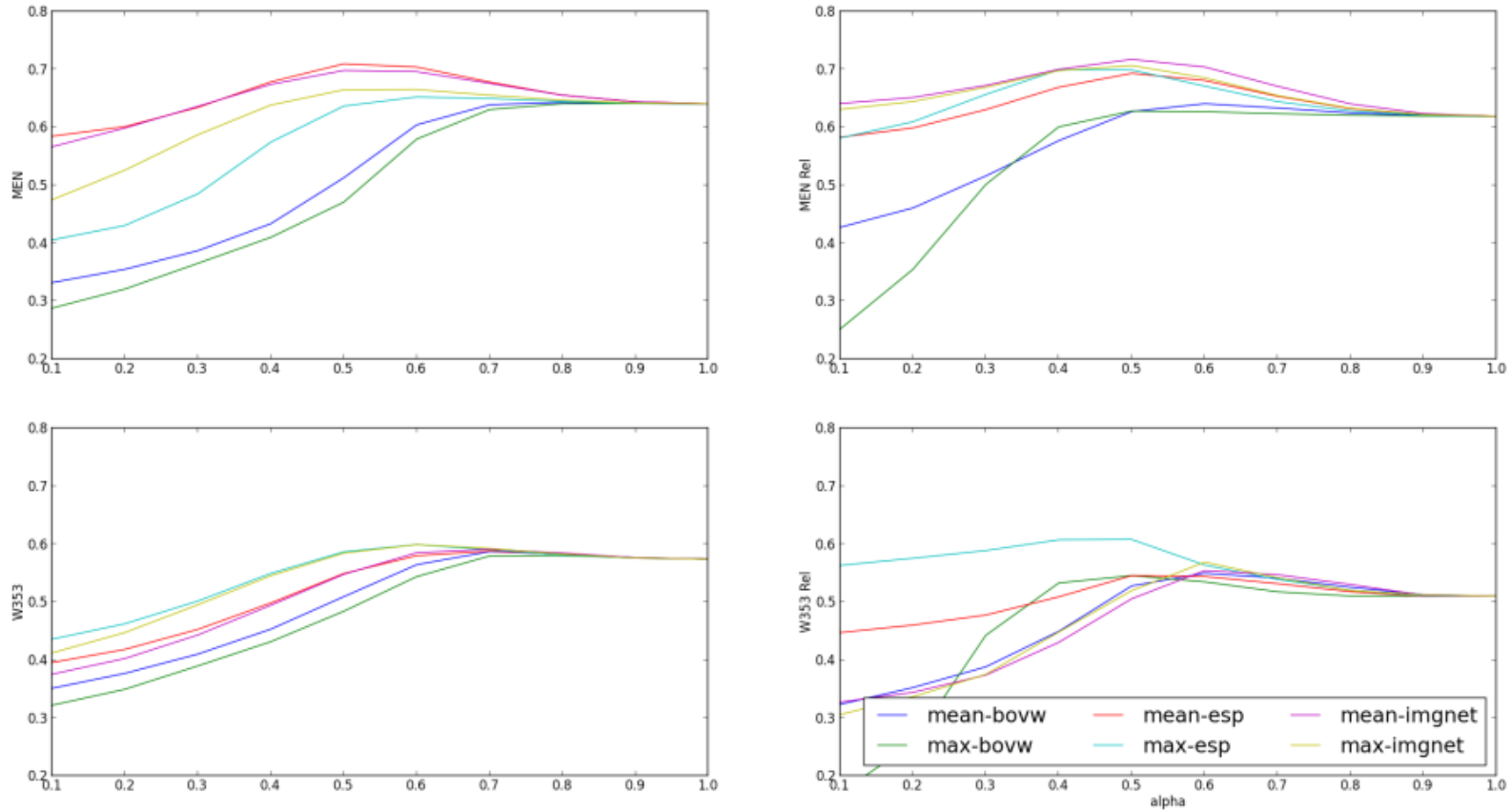


Figure 4: Varying the α parameter for MEN, MEN-Relevant, WordSim353 and WordSim353-Relevant, respectively.

3. Frage

- Für welche Arten von Konzepten/Wörtern wird wohl die Extrainformation aus Bildern das Ergebnis verbessern, für welche eher nicht oder weniger?

Ergebnisse II

W353-Relevant

ImageNet				ESP Game			
word1	word2	system score	gold standard	word1	word2	system score	gold standard
tiger	tiger	1.00	1.00	tiger	tiger	1.00	1.00
man	governor	0.53	0.53	man	governor	0.53	0.53
stock	phone	0.15	0.16	stock	phone	0.15	0.16
football	tennis	0.68	0.66	football	tennis	0.68	0.66
man	woman	0.85	0.83	man	woman	0.85	0.83
cell	phone	0.27	0.78	law	lawyer	0.33	0.84
discovery	space	0.10	0.63	monk	slave	0.58	0.09
closet	clothes	0.22	0.80	gem	jewel	0.41	0.90
king	queen	0.26	0.86	stock	market	0.33	0.81
wood	forest	0.13	0.77	planet	space	0.32	0.79

MEN-Relevant

ImageNet				ESP Game			
word1	word2	system score	gold standard	word1	word2	system score	gold standard
beef	potatoes	0.35	0.35	beef	potatoes	0.35	0.35
art	work	0.35	0.35	art	work	0.35	0.35
grass	stop	0.06	0.06	grass	stop	0.06	0.06
shade	tree	0.45	0.45	shade	tree	0.45	0.45
blonde	rock	0.07	0.07	blonde	rock	0.07	0.07
bread	potatoes	0.88	0.34	bread	dessert	0.78	0.24
fruit	potatoes	0.80	0.26	jacket	shirt	0.89	0.34
dessert	sandwich	0.76	0.23	fruit	nuts	0.88	0.33
pepper	tomato	0.79	0.27	dinner	lunch	0.93	0.37
dessert	tomato	0.66	0.14	dessert	soup	0.81	0.23

Table 2: The top 5 best and top 5 worst scoring pairs with respect to the gold standard.

Fehleranalyse

- Mehrere Fehlerquellen möglich:
  - schlechte linguistische Repräsentationen
  - schlechte ImageNet Repräsentationen

Zusammenfassung/Kritik

- Etwas “naiver” Ansatz der Bewertung (Spearman)
- Auf Fehler wird nur kurz eingegangen

Inhaltsverzeichnis

Quellen

- Mikolov et al.:
  Estimation of Word Representations in Vector Space
- http://image-net.org/
- Ahn et al.: Labeling Images with a Computer Game
- http://cs231n.github.io/convolutional-networks/